2020

Intentions and Creative Insights: a Reinforcement Learning Study of Creative Exploration in Problem-Solving

Colin, Thomas R.

http://hdl.handle.net/10026.1/15845

http://dx.doi.org/10.24382/1161
University of Plymouth

UNIVERSITY OF
PLYMOUTH

INTENTIONS AND CREATIVE INSIGHTS

A REINFORCEMENT LEARNING STUDY OF CREATIVE EXPLORATION IN PROBLEM-SOLVING

by

THOMAS RENAUD COLIN

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing and Mathematics

June 2020

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

This work has been carried out by Thomas R. Colin under the supervision of Prof. Dr. Tony Belpaeme and Dr. Nikolas Hemion. The work was funded by the European Union Marie Curie Initial Training Network FP7-PEOPLE-2013-ITN, CogNovo, grant number 604764.

Publications:

Colin, T. R., Belpaeme, T., Cangelosi, A., and Hemion, N. (2016). *Hierarchical reinforcement learning as creative problem solving*. Robotics and Autonomous Systems, 86, 196-206. doi: 10.1016/j.robot.2016.08.021

Colin, T. R. (2017). *Analyzing ambiguity in the standard definition of creativity*. AVANT, 8, 25–34. doi: 10.26913/80s02017.0111.0003

Colin, T.R. and Belpaeme, T. (2019). *Reinforcement Learning and Insight in the Artificial Pigeon*. Proceedings of the 41st annual meeting of the Cognitive Science Society (cogsci 2019).

Presentations at conferences and talks:

Colin, T.R. (2016). *Reinforcement Learning and Insight*. Invited talk, RLAI.

Colin, T.R. and Belpaeme, T. (2018). *A machine learning model of insight*. 2nd UK Creativity Researchers' Conference.

Colin, T.R. and Belpaeme, T. (2019). *Creativity by any other Name*. 3rd UK Creativity Researchers' Conference.

Word count for the main body of this thesis: **58399**

THOMAS RENAUD COLIN
June 2020

Acknowledgements

I am thankful first of all to my director of studies, Tony Belpaeme, for guiding me through the mazes of academia, for being an attentive reader of the many iterations of this thesis, and for trusting me to produce the kind of work that I wanted - even if that meant taking risks.

In Canada, Richard Sutton made a decisive contribution to this thesis by letting me invite myself (so to say) for an incredibly productive couple of months in Edmonton, where I learned a lot from the entire RLAI lab and its constant stream of visitors. In France, Nikolas Hemion, Alban Laflaquière, and Michael Garcia-Ortiz created at Aldeberan Robotics (now Softbank Robotics) an atmosphere of deep and genuine interest into the nature of intelligence, and offered an ear to my speculations. If it wasn't for my secondments in France and Canada, I would not have dared to go beyond the option architecture... and I might have finished this PhD over a year ago! It was worth it.

In all these places – Plymouth, Paris, Edmonton – I was surrounded by academic colleagues who soon became good friends. Tara Zaksaite put up with me for most of this PhD, and taught me more than Sutton and Barto (2018). Many others played an essential part in keeping me socially busy and somewhat sane, in addition to often offering fascinating and stimulating ideas about reinforcement learning, insight, or creativity. The following list is far from exhaustive: Emmanuel Senft, Frank Loesche, Oksana Hagen, Tian Tian, Francois Lemarchand, Katie Francis, Roshan Shariff, Pontus Loviken, Asya Grechka, the FBUM, Christos Melidis, RJ, Michael Straeubig, Diego Maranan, Pinar Oztop...

Finally, I am forever grateful to my parents, Olivier and Anne-Marie, for their love, and for giving me the drive to accomplish difficult things; and to my sisters, Julie and Célia, for being amazing friends and for keeping me grounded.

Abstract

INTENTIONS AND CREATIVE INSIGHTS
Thomas R. Colin

Insight is perhaps the cognitive phenomenon most closely associated with creativity. People engaged in problem-solving sometimes experience a sudden transformation: they see the problem in a radically different manner, and simultaneously feel with great certainty that they have found the right solution. The change of problem representation is called "restructuring", and the affective changes associated with sudden progress are called the "Aha!" experience. Together, restructuring and the "Aha!" experience characterize insight.

Reinforcement Learning is both a theory of biological learning and a subfield of machine learning. In its psychological and neuroscientific guise, it is used to model habit formation, and, increasingly, executive function. In its artificial intelligence guise, it is currently the favored paradigm for modeling agents interacting with an environment. Reinforcement learning, I argue, can serve as a model of insight: its foundation in learning coincides with the role of experience in insight problem-solving; its use of an explicit "value" provides the basis for the "Aha!" experience; and finally, in a hierarchical form, it can achieve a sudden change of representation resembling restructuring.

An experiment helps confirm some parallels between reinforcement learning and insight. It shows how transfer from prior tasks results in considerably accelerated learning, and how the value function increase resembles the sense of progress corresponding to the "Aha!"-moment. However, a model of insight on the basis of hierarchical reinforcement learning did not display the expected "insightful" behavior.

A second model of insight is presented, in which temporal abstraction is based on self-prediction: by predicting its own future decisions, an agent adjusts its course of action on the basis of unexpected events. This kind of temporal abstraction, I argue, corresponds to what we call "intentions", and offers a promising model for biological insight. It explains the 'Aha!'-experience as resulting from a temporal difference error, whereas restructuring results from an adjustment of the agent's internal state on the basis of either new information

or a stochastic interpretation of stimuli. The model is called the actor-critic-intention (ACI) architecture.

Finally, the relationship between intentions, insight, and creativity is extensively discussed in light of these models: other works in the philosophical and scientific literature are related to, and sometimes illuminated by the ACI architecture.

Table of contents

List of figures

List of tables

Introduction

Intelligent animal and human behavior consists in part of predictable routine. But the repetition of the same is sometimes interrupted by something surprising and valuable, original but appropriate, non-obvious, yet compelling. In one word: *creative*. This thesis investigates such creative anomalies in intelligent behavior, whether animal, human, or artificial.

Creativity research is fragmented across disciplines; to encompass all of the relevant research, one must therefore adopt an interdisciplinary approach (Sawyer, 2011a). Accordingly, this thesis draws and integrates views from multiple disciplines: primarily psychology, neuroscience, and artificial intelligence. As I work to elucidate an essential aspect of creativity, namely *insight*, I will be constructing a two-way bridge: psychology will inspire ideas in artificial intelligence, and artificial intelligence implementations will offer speculative explanations for psychological phenomena.

Creative insight – a phenomenon characterized by its suddenness, by its restructuring of existing representations, and by its frequent discovery of complete, successful solutions to difficult problems – seems to produce something out of nothing; but this is not due to divine inspiration or purely to chance. I propose that most insights have their origins in the extraction of general patterns from successful behavior, such that they can be applied to new problems.

This idea will be investigated in the domain of reinforcement learning (presented in chapter 3). In many reinforcement learning techniques, at any moment, a new action is tried based on how good it is (e.g. SoftMax action-selection), how likely it is to be optimal (Bayesian approaches), and/or based on how much can be learned from it (curiosity). These various techniques make use of experience to select exploratory decisions. However, simple exploration techniques such as ε-greedy action-selection (Sutton and Barto, 2018, pp. 27-28) continue to perform competitively and to be used in state-of-the-art systems (e.g. Mnih et al., 2015): this suggests something is missing.

Whereas advances in deep neural nets have allowed reinforcement learning algorithms to generalize from one action-decision to another, based on the matching structure of their respective inputs, there is not yet a corresponding technique for discovering temporal structure.

This thesis presents such a technique, called the Actor-Critic-Intention (ACI) architecture. In ACI, the purpose of intentions is to remove redundancy in successive decisions, thus allowing for individual decisions to affect an agent's decision-making process over time - including the representation of the situation. Furthermore, the technique allows agents to undergo positive surprise (in the form of positive temporal difference errors) immediately after a promising decision is made - corresponding to the "aha!"-moments seen in human and animal subjects. This technique is presented both as a tentative model for insight problem-solving in humans and animals, and as a promising technique for improving the problem-solving abilities of artificial agents.

Thus an investigation of creative anomalies leads to a theory covering both temporal abstraction for reinforcement learning agents, and insight in animal and humans. Although focused on problem-solving, these results are also meaningful in the wider context of creativity research - including for instance scientific and artistic creativity.

Plan

Part I: Theoretical foundations

The nebulous concept of creativity is the first difficulty encountered in this thesis. In chapter 1, I propose an analysis of the ambiguities that persist within the most widely accepted definition of creativity. Having spelled out the ambiguities, I select a restricted but less-ambiguous definition: I consider creativity *performed by* and *serving the purposes of* a lone creative agent (though in chapter 7 I will return to a more general discussion of creativity). Within this smaller research area, one phenomenon stands out: insight. From a psychological perspective, insight requires explanation; whereas from an artificial intelligence perspective, the ability to have insights seems reliant on one of the unknown ingredients in the recipe for intelligence.

Having narrowed the field of view in chapter 1, in chapter 2 I focus on insight. I review the psychological literature, including animal studies, human studies, and neuro-imaging studies. This review suggests that there is room and promise for a reinforcement learning theory of insight, establishing a connection between the properties of insight and those of hierarchical reinforcement learning algorithms.

Chapter 3 reviews the theory of Reinforcement Learning in a tutorial format, which (without sacrificing mathematical precision) seeks to be accessible for researchers specialized in the psychology of insight, creativity, and in computational creativity.

These three chapters make up the first part of this thesis. Chapter 1, although inspired by other reflections on this topic, is an original contribution to the philosophy of creativity and to creativity research (Colin, 2017, 2019). Chapters 2 and 3 are background chapters surveying what is relevant to subsequent developments in this thesis. However, they also make an original claim about the use of reinforcement learning to model insight, describing a relationship between these two fields which had not been previously identified (Colin et al., 2016).

Part II: Experiments in Deep Exploration

The second part of the thesis is an investigation of insight using models and simulation studies. Experiments are adapted from research on animal insight, particularly the work of Epstein et al. (1984) and Köhler (1921).

In chapter 4, I show how some insight-like phenomena emerge from statistical learning, and investigate whether techniques developed within the options framework (Sutton et al., 1999) can produce insight-like behavior. That is: trying out a new option can lead to a sudden change in the expectation of success, while simultaneously corresponding to a change in the agent's internal representation of the problem. However, learning options that generalize well turns out to be difficult. I briefly discuss why existing techniques, in particular the option-critic architecture (Bacon et al., 2017), may not be appropriate for the discovery of general options for exploration.

Chapter 5 introduces a novel approach for the discovery of temporally extended exploratory behavior. This is the ACI architecture, which exhibits intentions - an internal state of the agent which affects its short-term future decision making and behavior. Intentions are conceptually distinct from memories (which seek to encode relevant information from the past) and from options (which are best understood as discrete entities and have a termination condition). Intentions are learned by detecting temporal patterns in successive decisions made by an agent. I argue that intentions are best suited for adaptive exploration, and are a plausible source of insights.

The two chapters in part II investigate RL perspectives on exploration, starting from a basic paradigm and moving on to sophisticated examples. Along the way, analogies between reinforcement learning and insightful problem solving are uncovered (with some results published in the article by Colin and Belpaeme (2019)), and a model of insight based on intentional exploration is proposed.

Part III: Insight, Intentions, and Creativity

If part I served to restrict our attention and sharpen our focus to a limited domain, and part II sought to capture experimental evidence within that limited view, then part III "zooms out" again to a wider field of view, so that the findings can be seen as part of the broader landscape of creativity research.

In chapter 6, I discuss the ACI architecture in light of the literature on intentions in philosophy, psychology, and artificial intelligence; and in chapter 7, I discuss the meaning of the new model of insight in the context of the wider literature on artificial intelligence, psychology, and creativity.

Finally, in the conclusion I summarize the thesis and its contributions, and review possible extensions made possible by this work.

Main contributions

In order of appearance in this thesis:

1. A conceptual analysis of creativity as a scientific concept.

2. Evidence for the role of learning in insight via simulation experiments.

3. A new Reinforcement Learning technique for temporal abstraction called the ACI architecture, with applications to the related sub-fields of transfer, deep exploration, large action spaces, and planning.

4. A theory of insight based on the relevant interdisciplinary literature and on intentions.

Part I

Theoretical foundations

Chapter 1

Creativity

In 1950, Turing asked the question: "can machines think?" It is a difficult question, at the intersection of mathematics, engineering, and philosophy. This thesis must tackle similar questions ("how can machines be creative?"), so it is sensible to seek inspiration from the manner in which Turing tackled the problem[1]. To ask whether machines can think, Turing argues, one should know what "thinking" means. Likewise, to judge the validity of a model of creativity, one should define creativity. But it is not clear what either thinking or creativity are. Some philosophers (Boden, 1996; Fodor and Crawford, 2005) have argued that creativity is not a natural kind[2]; if they are right, seeking a scientific theory of creativity is absurd.

Let us turn, then, to Turing's solution in his classic article, "Computing machinery and intelligence". Instead of directly addressing the question at hand, Turing discusses its linguistic origin so that the original question is dissolved, and may then be replaced by another, more meaningful question:

[1] Others have looked at Turing's approach with respect to computational creativity (Pease and Colton, 2011b; Boden, 2010). However, the focus of these discussions was usually the adequacy of a Turing-like Test for the evaluation of computer art; this chapter takes a related but different direction.

[2] Fodor says: "It's not clear that [emotion, creativity, and imagination] are, as some philosophers like to say, "natural kinds", that they're the sorts of states over which law-like, reliable, counterfactual supporting generalizations can be stated and in terms of which theories can be elaborated. I like to refer to the sad career of a guy who thought he was going to develop a science of Tuesdays. And it sort of worked for a while, he discovered some generalizations about Tuesdays. That is, they come after Mondays, they come before Wednesdays, they last about 24 hours and so forth and so on, and then the subject seemed to dry up. And the reason it dried up is that the property of being a Tuesday doesn't provide a scientific domain nor unfortunately do most of the properties of most of the things that we're humanly interested in. What I would guess is that emotion and creativity and imagination (...) just aren't going to prove appropriate domains for the kind of theory construction that scientists do, though they may be a perfectly appropriate domain for, as it were, writing novels (...)". Margaret Boden has argued that "Creativity is not a natural kind, such that a single scientific theory could explain every case" (Boden, 1996, pp. 267-268), where it is understood that "to say that a kind is natural is to say that it corresponds to a grouping that reflects the structure of the natural world rather than the interests and actions of human beings" (Bird and Tobin, 2017).

> *I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think". The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words (...) We now ask the question, "What will happen when a machine takes the part of [a participant] in [the imitation game]?".* — A. M. Turing (1950)

Like the word "think", the word "creativity" is shrouded in a cloud of confusion. But no Gallup poll will be necessary to collect opinions regarding the meaning of creativity: numerous academics have devoted considerable attention to the matter of defining it.

My goal in this chapter is not to offer yet another definition, but to identify the possible meanings of the word creativity in the context of specialized academic research. Having done this, I shall replace the question "what is creativity?" by the related and relatively unambiguous "what is insight?". Unlike Turing, however, I will not dismiss entirely the original question. Rather, I will carefully discuss when and why the substitution is valid, showing the manner in which the new question, "what is insight?", is related to the old question, "what is creativity?". This will make it possible to identify the manner in which the results obtained in this thesis, with respect to insight, relate to the wider body of literature on creativity in all its manifestations.

1.1 Defining creativity

It is striking that creativity researchers, compared to other fields, have devoted such considerable energy to the question of its definition: "no topic is more central to research on creativity", and "nearly every article in the [Creativity Research Journal] at least briefly defines creativity" (Runco and Jaeger, 2012, p. 92).

Why should we worry about the definition of creativity? An incomplete or vague definition is not necessarily damaging, even when it affects the most central concept of a field or discipline. For instance, Trifonov (2011) collected 124 definitions of life (the central concept in biology), and Legg and Hutter (2007) brought together 72 definitions of intelligence (the central concept in artificial intelligence and an important one in psychology).

Despite this inability to settle on a definition of their object of study, and indeed a relative disinterest towards the matter, these fields appear to be thriving.

The difference of treatment of definitional questions may be due to the disparity of the roles taken by these concepts (life, intelligence, creativity) in the corresponding disciplines. Indeed, both "life" in biology and "intelligence" in artificial intelligence are usually relegated to a philosophical backstage; the leading roles are instead given for instance to evolution or medicine (biology), or to planning or learning (artificial intelligence). In contrast, the concept of creativity is at the forefront of creativity research. Many creativity researchers seek to make general claims about the nature of creativity: statements of the form *"creativity correlated with x"* or *"creativity requires y"* (e.g. Runco and Jaeger (2012); Sternberg (1996); Takeuchi et al. (2010)). But if we cannot confidently separate the creative from the non-creative, such statements themselves become vague and ambiguous.

To prevent this, we must know what we are talking about when we talk about creativity. The "standard definition of creativity" (Runco and Jaeger, 2012) can be seen as the summary of decades of discussions and debates aimed at reaching a point of (approximative) agreement between creativity researchers. The result is a bi-partite definition: creativity requires (1) originality and (2) effectiveness. This definition is best understood by also considering the family of related definitions to which it belongs: "novelty and value" or "novelty and appropriateness," etc.

I propose investigating this definition via conceptual analysis, a philosophical method consisting of clarifying a concept by exploring possible interpretations and testing their internal consistency, sometimes by using thought experiments. The analysis reveals two sources of ambiguity: the relativity of the criteria of originality and effectiveness to a context and norm, and the potential subjectivity of a judge of creativity.

In the next section I briefly present the history of the field. I then study the standard definition in more detail: what is meant by *originality* and by *effectiveness*? Finally, I consider some of the implications for creativity research.

1.2 Models and taxonomies of creativity

Before discussing the ambiguities of the definition, it is useful to give an overview of past conceptualizations of creativity.[3]

[3] For a more detailed account of the story of creativity research, interested readers can consult the articles by Batey and Furnham (2006) and Hennessey and Amabile (2010). Some of the key characters are cognitive processes, personality traits, social interactions, creative achievements, and practical applications.

The word itself is a surprisingly recent addition to the western vocabulary[4]; its appearance is almost concomitant with that of creativity research as a field, which is generally accepted to be the 1950 address to the American Psychological Association (Guilford, 1950). Prior to the 1940s, the word "creativity" was almost unheard of in English, German (Kreativität), or French (créativité). Its popularity rose sharply and steadily until the 2000s (Google Ngram Viewer, 2017); the European Union branded 2009 the "Year of Creativity and Innovation" (European Commission, 2009).

As the popularity of the word grew, academics were devising models of creativity. Important such models include:

Stage-based models (Wallas, 1926): the creative process is divided into separate stages, such as preparation, incubation, illumination, and verification. Later models (Lubart, 2001) have included stages for problem-finding or for the communication of results.

Convergent and Divergent thinking (Guilford, 1956b): creativity is considered within a theory of the intellect, in which creativity results from the interplay of convergent and divergent productive processes.

Blind Variation, Selective Retention (BVSR) (Campbell, 1960b): this recently revitalized (Simonton, 2011) model focuses on a trial-and-error explanatory framework; it is amenable to cognitive or computational as well as social interpretations.

The Systems Model (Csikszentmihalyi, 2009, p. 6): *"creativity results from the interaction of a system composed of three elements: a culture that contains symbolic rules, a person who brings novelty into the symbolic domain, and a field of experts who recognize and validate the innovation."*

To compare different explanatory models and to map out the scope of their applicability, various taxonomies have been proposed. Such work includes the "four Cs" (Kaufman and Beghetto, 2009), which distinguish types of creativity ranging from "mini-c" (creativity in learning and development) to "big-C" (eminent creativity), with intermediary levels "little-c" (everyday creativity) and "pro-c" (professional creative output). Another example is the four Ps framework (Rhodes, 1961), which refers to the elements of creativity research as the Person, Product (object, idea, behavior. . .), Process, and Press (interactions with the social environment of the creator)[5]. Boden (2004) distinguishes "explorative", "combinatorial" and

[4] However, related terms and concepts predate it; for instance, there was considerable discussion of *genius* in the 19th century.

[5] The four Ps have later been extended and revised by other researchers to form the five As (Glăveanu, 2013) and the seven Cs (Lubart, 2017).

"transformational" creativity, the latter being the more radical, mysterious, and interesting of the three.

These variegated views and frameworks illustrate the fragmentation of the field (recognized with concern in, e.g., Hennessey and Amabile (2010)). This fragmentation is also evident in comprehensive introductions to the field, such as Sawyer (2011a) or Runco (2014b), in which different sections or chapters discuss cognitive, developmental, neurobiological, social, educational, and cultural perspectives... among others.

Indeed, creativity research has produced an awe-inspiring variety of "perspectives" and "approaches", but it has failed to converge towards a single big picture which might be called the theory of creativity. I believe this is in part because the researchers share the same lexicon, but often do not speak the same language.

The word creativity is fought over by academic communities which attach different meanings to it. But a word with multiple competing meanings is impractical for communication. How can misunderstandings be avoided? A radical solution would be to stop using the term altogether in favor of a set of more precise alternatives; but this is unrealistic due to the cachet that "creativity" already has. A more pragmatic approach is called for: let us identify and list the different meanings attached to creativity and use this list to disambiguate the uses of the term in creativity research. In this chapter, the first step of this program takes the form of a conceptual analysis.

1.3 Conceptual analysis of the standard definition

Margolis and Laurence (2014) define conceptual analysis as "a distinctively a priori activity that many take to be the essence of philosophy. [...] Paradigmatic conceptual analyses offer definitions of concepts that are to be tested against potential counter-examples that are identified via thought experiments". Conceptual analysis has been criticized on the grounds that the intuitions with which we navigate thought experiments can be individual- or culture-dependent (ibid.). It would be dangerous, perhaps even "absurd" to rely on such intuitions. Hence I will not consider the concepts of creativity used by lay people or even by creative professionals: these are likely to be sometimes incoherent, and usually inconsistent across individuals, communities, or cultures. Using these would be the equivalent of the "Gallup poll" ridiculed by Turing. Instead, I focus on the definitions and intuitions put forward by creativity researchers whose views are informed by experimental evidence, correspond to an extensive knowledge of the field, and aim to achieve consensus.

1.3.1 The standard definition

The "standard definition", summarized by Runco and Jaeger (2012) as "originality and effectiveness" (p. 92), is related to a cluster of bi-partite definitions which has accumulated since the origin of the field. Each component is better understood by considering the cluster of related terms[6]:

- **Originality**, Novelty/Novel/New, Non-obvious, Uncommon, Unique.
- **Effectiveness**, Adaptive, Appropriateness, Correct, Fit, Good, Realistic and acceptable, Relevant, Valuable, Usefulness, Worthwhile and compelling.

Unfortunately, these definitions make use of concepts that are themselves under-specified. This is no coincidence: they harness ambiguous terms to describe an ambiguous concept. Instead of decomposing creativity into more precise and fundamental concepts (as an explanatory definition would: "water is H2O"), some vagueness is preserved. This could be by design: to be consensual, the definition must preserve the existing inconsistencies and disagreement between researchers. There is only one way to preserve these disagreements in a definition while avoiding incoherence and contradiction: the definition must be ambiguous, under-specified itself.

Indeed, this is precisely what makes this concise definition an interesting starting point for elucidating ambiguity in the word creativity. The definition is likely to compress the ambiguity of "creativity" within just a few words. What ambiguity is there? I find two different sorts.

1.3.2 First ambiguity: context and norm

The first sort is the relativity of the two components. Absolute interpretations are not taken seriously in the field: that is, few believe that creativity requires either absolute novelty or originality, or objective good. Thus, originality and effectiveness must be relative to something. But relative to what?

- **Originality is relative to a context:** originality can only be measured relative to a group, and novelty relative to a history. I will refer to this group or history as the *context*.

[6] The terms in each cluster appear in various bi-partite definitions. Runco and Jaeger (2012) provide references for the corresponding definitions. The earliest formulation of a bipartite definition might be Kant's remark that "since there can also be original nonsense, [the products of genius] must at the same time be models, i.e., exemplary" (Kant, 1790/2000), which could perhaps be rephrased as "genius requires originality and exemplarity". At the time, the word "creativity" was not used.

- **Effectiveness is relative to a norm:** effectiveness, goodness, relevance, can only be measured with respect to goals, criteria, or values; in the most general sense these are called *norms*[7].

In practice, creativity research often lets context and norm vary together by specifying them based on a singular point of view. For instance, one may implicitly choose the point of view of the creator: the creative product must be new to the creator (context), and effective for the creator's purposes (norm). It is also possible to prefer some group external to the creator, such as the field or community in which the creator is integrated: the product must then be original within the group (context), and effective for the group or according to criteria defined by the group (norm).

Thus, the concise definition, "creativity requires both originality and effectiveness", can be rephrased as "creativity with respect to a context and a norm requires both effectiveness relative to the given norm and originality relative to the given context". This more cumbersome definition makes it explicit that the components (originality and effectiveness) are ambiguous whenever context and norm are not specified.[8]

To illustrate this point, consider the following two examples of creative individuals drawn from distant branches of the literature: the New Caledonian crow Betty, and the 20th century post-impressionist painter Vincent van Gogh. Betty was dubbed creative (Weir and Kacelnik, 2006) for building a tool to solve a problem. The crow had never made such a tool before, and had not seen another crow make one, hence the tool-building process was original within Betty's context and was effective at reaching the food reward. Contrast with Van Gogh, whose widely recognized creativity is dependent not merely on the novelty of his paintings with respect to his personal history, nor to their fit with his own norm of artistic value. Instead, the creativity of the painter is measured against the history of art, the artists of his time, and the norms of today's art connoisseurs and art historians. Based on the definition, we are justified in saying that both Betty and Van Gogh are creative in some sense; but they are creative relative to different contexts and norms. The difference between them is not (just) quantitative, but also qualitative, because different norms and contexts and therefore different concepts of creativity are involved.

Figure 1.1 shows how ambiguities with respect to context and norm allow for a large "space of definitions" of creativity. The graph shows a few possible variations, and is not

[7] In this chapter I use the term "norm" in its philosophical sense (a context in which "normative" is opposed to "descriptive"), rather than in its mathematical sense.

[8] Note however that this definition explicitly endorses the validity of multiple possible choices for the norm and context, instead of merely failing to choose between them, making it less consensual to those researchers who reject some of these contexts.

meant to show the entire space: there are many more ways to define both norms and contexts than those shown on the graph's axes.

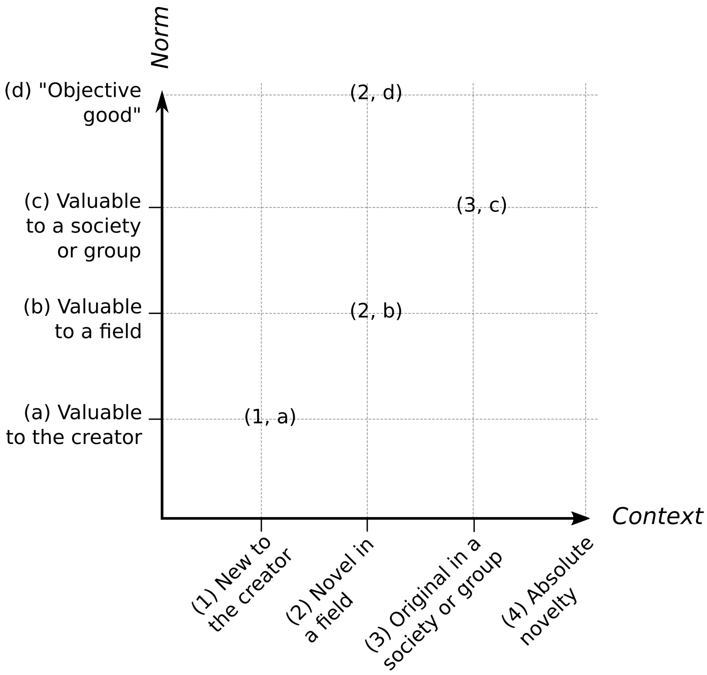


Fig. 1.1 Some possible interpretations of "creativity", depending on context and norm. Below are examples illustrating how different forms of creativity match, or do not match, different definitions:

- Betty the crow (Weir and Kacelnik, 2006) is creative in the sense of $(1,a)$, and arguably also in the sense of $(3,c)$ when considering the group of captive crows to which she belongs.
- Van Gogh is creative in the sense of $(2,b)$, $(3,c)$.
- An artist who sticks to the same style may feel that their own work has become routine, thus failing $(1,\cdot)$ even though art critics may continue to find originality and value in their work $(3,c)$.
- A scientist using ideas from a different field to further their own field might be creative in the sense of $(2,b)$, which is close to that of Csikszentmihalyi (2009), but may fail $(3,c)$ if they have no genuinely new ideas of their own.
- "Malevolent" creators (creative terrorism, for instance; see Cropley et al. (2008)) fit $(2,b)$, but not $(2,d)$ or any other definition in $(\cdot,d)$.
- The Torrance test (Torrance, 1988), or other tests that evaluate originality based on the frequency of responses among a tested group, assume a definition of the form $(3,\cdot)$.

1.3.3 Second ambiguity: the interplay between creator and judge

The second source of ambiguity is the person or group making the creativity judgement, presumably based on the two criteria of originality and effectiveness. This is surprising: while in the previous section I described creativity as relative (to norms and context), it remained nevertheless possible to directly measure it against these elements if they were supplied. For instance, one might say that x is creative given a norm N and a context C. But a subjective judge is now introduced, typically a community or even a "field" (e.g. aviation, biology, dancing . . .). This negates the possibility of assessing the creativity of an individual in isolation; this undermines branches of creativity research that focus on cognitive characteristics at the individual level.

But major figures of creativity research have claimed that creativity requires a judge. For instance, Runco and Jaeger (2012) lament that "The standard definition only pinpoints which criteria must be used; it does not say anything about who is to judge each". Csikszentmihalyi (2009), pp. 23-25, explicitly asks his readers to pick such a judge: either the creators themselves, or relevant members of society, the latter having his vote. According to him, "[Van Gogh's creativity] came into being when a sufficient number of art experts felt that his paintings had something important to contribute" (ibid., p. 31). According to Csikszentmihalyi, Van Gogh therefore became creative after his death, by virtue of the changing opinions of art experts.

Why would the terms of my first distinction not suffice? Distinguishing different contexts and norms, one could write that Van Gogh was both (1) not creative based on the criteria of a first group (his contemporaries), but (2) was creative relative to the criteria of another group (later critics). However, these authors (Runco, Jaeger, Csikszentmihaliy) ignored this approach. I surmise that they view creativity as involving a dynamic interaction between the creator and the judges, such that the creator is able to alter the norms by which the product is evaluated. This is most clear in the case of creativity in the arts, and perhaps best expressed by Proust (1921):

> People of taste and refinement tell us nowadays that Renoir is one of the great painters of the last century. But in so saying they forget. . . that it took a great deal of time, well into the present century, before Renoir was hailed as a great artist. To succeed thus in gaining recognition, the original painter, the original writer proceeds like an oculist. The course of treatment they give us by their painting or by their prose is not always agreeable to us. When it is at an end the operator says to us: "Now look!" And, lo and behold, the world around us (which was not created once and for all, but is created afresh as often as an original artist

> is born) appears to us entirely different from the old world, but perfectly clear. Women pass in the street, different from what they used to be, because they are Renoirs, those Renoir types which we persistently refused to see as women. The carriages, too, are Renoirs, and the water, and the sky. (p. 1131)

The creator, via his creation, proceeds "like an oculist": transforming the vision of the judges, and thereby, their judgement. Is this specific to the arts? Kuhn (1970) sees a similar phenomenon in the sciences, in which an "incommensurable" innovation must lead to the "conversion" of a scientific community to new criteria of good science. In this view, creativity is an emergent social phenomenon rather than a mere cognitive ability. However, this interpretation seems rather more applicable to the eminent, "big-C" creativity of historical importance, than to the "mini-C" or "little-C" creativity of children and daily life. Indeed, Csikszentmihalyi's Van Gogh thought experiment could "seem insane" (Csikszentmihalyi, 2009, p. 31) to some readers, such as, perhaps, experimental psychologists attempting to measure creativity in the lab. It is more useful for those focused on the process of creative thought (for instance the insight phenomenon) to adopt a more prosaic definition, according to which Betty the crow is creative.

1.3.4 On the analysis of creativity

I have presented two types of ambiguity in the definition of creativity. The first concerns the context and norm against which originality and effectiveness are to be measured. The second concerns the existence and identity of a judge of creativity. The combination of these ambiguities makes room for many different interpretations of the term "creativity".

If the meanings of creativity are not identical, they can at least be related. For instance, the first ambiguity can be set aside when the norms and context of the individual and those of society are close enough to produce similar evaluations of creativity. The second ambiguity, despite considering creativity as an emergent social phenomenon, does not preclude the involvement of specialized cognitive processes on the part of the creator, some of which may be similar for multiple interpretations of creativity. In particular, there is a striking parallel between the "paradigm shifts" (Kuhn, 1970) seen in the most eminent of scientific or artistic works and the representational change observed in insight problem-solving (Ohlsson, 1992). The consensus view is that, in spite of the differences between these interpretations of creativity, the areas of agreement justify the unity of the field (Tardif and Sternberg, 1988).

This relatedness notwithstanding, it is not difficult to provide examples which satisfy one interpretation of the definition, but fail to satisfy another; I have done so in the previous section by considering Betty the crow and Van Gogh. There are other such instances. Consider, for example, the much-debated relationship between mental illness and creativity (see e.g. Kyaga et al. (2011)). Mental illness causes ineffective behavior relative to oneself, but exploring the less-traveled path may increase the probability of discovering something valuable relative to the community. Therefore, differing interpretations of creativity with respect to the first ambiguity may help explain the heated disagreement on this issue. Consider also the debate on malevolent creativity (Cropley et al., 2008) and other such discussions[9].

Sawyer thinks creativity research is divided, and claims that as long as the different communities proceed "on separate tracks, we will fail to explain creativity" (Sawyer, 2011a, p. 14). By studying the definition of creativity, I hope to have shed light on the hidden causes of these divisions: creativity researchers study different concepts (creativity relative to oneself, to a group; involving a subjective judge or not), which they refer to under a single label. Acknowledging this diversity of interpretations, rather than denying it in the name of unity, would allow creativity researchers to communicate with each other while avoiding fruitless controversy over the "true meaning" of creativity.

1.4 Focusing on insight

In the remainder of this thesis, I will focus on originality and value relative to the creative agent, rather than relative to a third party (experts, society, etc.); in figure 1.1, this corresponds to coordinates $(1,a)$. I will look into the kind of creativity displayed by Betty the Crow, rather than the kind displayed by Van Gogh. Arguably, the individual level of creativity is the more fundamental concept. Without the individual creativity of animals, children, without the daily small feats of ingenuity that all humans are capable of, can there be eminent creativity? Often, social creativity seems to have individual creativity at its origin and foundation. The different kinds of creativity are related; in the discussion and conclusion chapters I will discuss the implications of my results with respect to other conceptions of creativity.

Unfortunately, a great variety of behaviors, by humans, animals, and algorithmic agents, match this definition of individual creativity in some minimal way. A basic Q-learning algorithm (Watkins and Dayan, 1992), for instance, is creative in the sense that it keeps making improvements that are both original (previously unknown) and useful (increasing the expected return) relative to the agent's previous behavior and efficiency. Indeed if any efficient adaptation to novel circumstances is creative, and the world is constantly changing ("you cannot step into the same river twice" (Plato and Reeve, 1998)); then it is non-creative

[9] The now famous 1971 debate between Noam Chomsky and Michel Foucault (Chomsky et al., 1971) was the theater of an extended discussion of creativity, in which the linguist and the philosopher adopted different concepts of creativity (a cognitive ability for Chomsky, a socio-cultural phenomenon for Foucault). This debate constitutes a fine illustration of the difficulties associated with discussions of creativity across paradigms.

behavior that is exceedingly rare! By this standard, artificial creativity is already solved, albeit in a trivial, unsatisfying manner.

However, a look at the cognitive psychology of creativity suggests that there might be more to individual creativity: there is the phenomenon of insight. Humans and certain animals, such as great apes (Köhler, 1921; Mendes et al., 2007) and certain species of crows (Weir and Kacelnik, 2006), appear to sometimes solve a radically novel problem all at once, rather than by small incremental changes, as most learning algorithms do. Insight is typically accompanied by a seemingly instantaneous and radical change in the *representation* of the problem, which psychologists call "restructuring". Note how this psychological phenomenon is strikingly similar to the radical transformations occasioned in eminent creativity, discussed by Proust and Kuhn, in which "the world around [an agent] appears to [it] entirely different from the old world, but perfectly clear".

According to Tardif and Sternberg (1988), the majority view is that flashes of insight are "a small but necessary component of the creative process" (p. 430); Schooler and Melcher (1995) further note that insight constitutes "one of the major sources of ineffability with which discussions of creativity have grappled". For Ohlsson (2011) "what distinguishes creative from analytical processes is that the former are punctuated by *insights*" (p. 87). Insight is often recognized, especially by psychologists, as an essential aspect of creative thinking. This thesis focuses on the insight phenomenon, which I believe is key to understanding the creativity of biological and artificial agents.

1.5 Conclusion

1.5.1 Summary

1. There is no consensual and unambiguous scientific concept of creativity.

2. Various conceptions of creativity cover a range of phenomena including individuals acquiring new skills, and great artists or scientists transforming entire cultures.

3. Within the scope of individual creativity, the phenomenon of insight is poorly understood, and its analysis has potentially wide implications for understanding other forms of creativity.

The bulk of this thesis focuses on understanding the insight phenomenon, both for its own sake, as an inspiration for more intelligent algorithms, and as a step towards understanding other aspects and conceptions of creativity. I will review the literature on insight in the next chapter.

Contributions

Much of this chapter (apart from the brief review of creativity research) consists in a novel analysis of the concept of creativity. Both the results and the methods with which they are achieved are original to creativity research (although prior works such as that of Wiggins et al. (2015) have presented related ideas).

- The general method employed here is to focus on the standard definition of the term "creativity" to discover potential for inconsistent or contradictory usage.
- The specification of a "space of interpretations" of creativity in terms of two variables (a context and a norm) is a novel view of creativity.

An early version of some of the main ideas presented in this chapter was published as "Analyzing ambiguity in the standard definition of creativity" (Colin, 2017).

1.5.2 Bibliographical remarks

The psychology of creativity is very briefly reviewed in this chapter. Much of this is adequately summarized in the two main academic textbooks on creativity research, "Explaining Creativity" (Sawyer, 2011a) and "Creativity" (Runco, 2014b) (both in their second edition).

The philosophy of creativity has received surprisingly little attention despite the need for conceptual clarification which made this chapter necessary in a thesis otherwise primarily concerned with psychology and artificial intelligence. The most influential work, especially for computational approaches, is undoubtedly "The Creative Mind" by Boden (1996); but see also the agency theory of creativity by Gaut (2010). Because this chapter seeks to make progress in a philosophical domain, it is also indebted to philosophers not directly associated with creativity. Turing's views, mentioned in the introduction, are perhaps best understood as derivative of those of the later Wittgenstein (Wittgenstein et al., 1953/2010)[10] (for the view that the disagreements surrounding the concept of creativity must come from some misuse of the concept). Carnap (1950) is also of particular relevance for the idea that scientific progress can be achieved by redefining a concept.[11]

The geometrical interpretation of the creativity concept presented in figure 1.1 is reminiscent of the notion of conceptual space put forward by the philosopher Gärdenfors (2004).

[10] Turing and Wittgenstein knew each other in Cambridge, and had extensive discussions on the philosophy of mathematics (Wittgenstein and Diamond, 2015). This might explain Turing's use of a seemingly Wittgensteinian approach to deal with the question of whether machines can think.

[11] Carnap famously takes the concept of "Fish" as an example. The concept of Fish used to extend to most sea-based animal life; narrowing it down, for instance by excluding dolphins or octopuses, arguably renders it more useful scientifically.

Creativity as defined by context and norm to some extent generalizes previous ideas, usually expressed separately for the norm/value and the novelty/originality/surprise aspects of creativity. In particular, Boden distinguished between "historical" and "psychological" creativity (H- and P-creativity) Boden (1996), to mark the distinction between absolute novelty and novelty with respect to an individual (but this disjunction does not leave room for novelty with respect to a group or a field; see e.g. Csikszentmihalyi (2009)). With respect to value, there has been some discussion about the "dark side" of creativity, e.g. creative crimes. Cropley et al. (2008) discuss this in terms of "subjective benevolence", i.e. the idea that usefulness or value is dependent on an agent or group. Wiggins et al. (2015) offer views closely related to the present work: they analyse creativity as based on value judgments requiring a context. However, in contrast with the present work, Wiggins and colleagues do not treat separately the "value" and "novelty" components of the definition, and do not distinguish context from norm. Finally, a book chapter by Gruner and Csikszentmihalyi (2019) presents a distinction between individual and social creativity reminiscent of the one presented in this chapter.

The interplay between creator and judge has been most notably discussed by Csikszentmihalyi (2009). In the context of computational creativity, also see Sosa and Gero (2008) and Saunders (2019); the latter contains a literature review addressing this question.

Chapter 2

Insight

In chapter 2, an investigation into the meaning of creativity led to focusing this inquiry on a more specific target: insight. In this chapter I review the interdisciplinary literature on insight from experimental psychology and neuroscience.

This chapter is complemented by three appendices (A, B, and C). Appendix A presents experimental problems, whereas appendix B gathers several first-hand accounts of important creative discoveries. They can be read as informal (and perhaps entertaining) introductions to the study of insight, before or together with this chapter. Appendix C deals with contemporary theories of insight, and is best read after this chapter.

This chapter has three purposes. First, the characteristics of psychological insight described here serve in later chapters as inspiration for a creative reinforcement learning algorithm. Second, the "insightful" algorithm serves as a model of biological insight; the models are evaluated in light of psychological and biological evidence. Finally, in the last section of this chapter I spell out the reasons for choosing reinforcement learning as a tentative model of insight.

Insight is also called the "Aha!" experience, the "Eureka!" moment, and more rarely "illumination", "epiphany"... It is the transition undergone by a problem-solver who suddenly and unexpectedly comes to know how to solve a problem (Mayer, 1995)[1]. Furthermore, insight is associated with the transformation, or *restructuring*, of the problem representation.

[1] There are many other definitions. The earliest technical definition seems to be in an untranslated work by Gestalt psychologist Selz (Selz, 1922, p. 591): "Als einsichtig bezeichnen wir die Anwendung einer Lösungsmethode, soweit die Sachverhältnisse, auf denen ihre Anwendbarkeit beruht, erkannt, d.h. abstrahiert sind, und diese Erkenntnis die Anwendung bedingt" ("The use of a solution method is said to be "insightful" in so far as relationships, on whose applicability they rely, are recognized, i.e. abstracted from it, and this realization conditions the solution" – translation by Frank Loesche and myself). For an extended discussion on the history of the concept of insight, and the connotations associated with the various terms for it, see Loesche (2018), chapter 2.

Research on insight was the theater of regular controversies, making a sober review a difficult exercise. To organize these conflicting opinions into a structured review, I will tell a chronological as well as thematic story, following the movements of the main impetus of insight research over the decades. In doing so I cover most of the diversity of insight research[2].

From the beginning of the century until the 1980s, animal studies yielded the most interesting results; then, until the early 2000s, it was human experiments; currently it is human brain imaging that pulls the field forward[3]. This literature review, after the introductory section 2.1, follows insight research through these three phases, and is concluded by a discussion of the analogies between reinforcement learning and insight.

2.1 Introduction

In most of the psychological literature, insight is considered a form of problem-solving, where *"solving a problem is transforming a given situation into a desired situation, or goal"* (Simon, 2001)[4]. Below I present two competing approaches to human problem-solving: the Gestalt and cognitive views.

2.1.1 Gestalt views

The Gestalt tradition (e.g. Koffka, 1935) analyzes human and animal problem-solving by focusing on representation. For gestaltists, solving a problem, or part thereof, consists in transforming the problematic situation by changing its representation. This is illustrated by figure 2.1a, in which the problem is to find the sum of the areas of the square and the parallelogram strip. In that figure, one may of course proceed analytically by reproducing learned rules for calculating the areas of squares and parallelograms[5]; in which case a Gestaltist might say that there was not much of a problem to begin with. However, a child who has not learned the formula for the area of a parallelogram might instead *restructure* the problem: in place of the square and the parallelogram strip, one might, for instance, see two overlapping triangles of base a and height b, making the solution immediate.

[2] The historical approaches focusing on eminent creators are among those not covered in detail; for instance, the historiometric approach promoted by Simonton (2016). Although it is of great interest, this literature generally focuses on creativity or genius, and only discusses insight in passing; besides, these approaches open too broad a field of inquiry to allow for a careful review in this chapter. Instead, a selection of historical accounts of insights is presented in appendix B.

[3] Although the study of animal and human behavior is continuing to yield useful new results.

[4] This can be done in the mind or in interaction with an environment.

[5] The area of the square is b^2, and the area of the parallelogram is $b \times (a-b)$; therefore the total area is the sum of these terms: $b^2 + b \times (a-b) = ab$

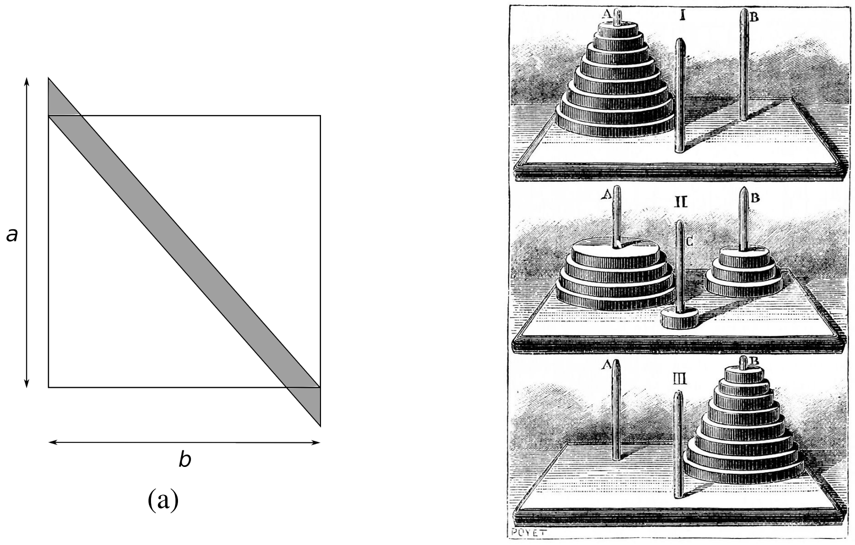
Humans excel at such tasks, and some other animals appear to be capable of similar feats in the type of problems that interest them. Köhler (1921) had chimpanzees attempt to retrieve inaccessible (out of reach) bananas; in order to solve the problem, they had to realize that tree branches in their enclosure could be broken off and used as sticks, after which the solution was trivial. Insightful problem-solving has since been observed in orangutans (Mendes et al., 2007) and corvids (Taylor and Gray, 2009), posing a challenge for explanations of problem-solving in terms of Thornidikian trial-and-error (Thorndike, 1898a) and for the cognitive school, which is described below.

2.1.2 Cognitive views

The cognitive school is best exemplified by the classic work of Newell and Simon (1972): in their book *Human Problem Solving*, problem-solving is interpreted as a search process. A problem-solver "navigates" a problem space, seeking to transform an initial state into a goal state. Formally (Newell and Simon, 1972, p. 810), $\mathcal{U}$ is the set of knowledge states containing the initial state $u_0 \in \mathcal{U}$, which must be transformed into a state u_n, such that $u_n \in \mathcal{G}$, where $\mathcal{G} \subseteq \mathcal{U}$ is the set of goal states. This is done using operators from a set $\mathcal{Q}$ which govern the transitions between states. Within this framework, problem-solvers can adopt a divide-and-conquer approach, or other more elaborate strategies; indeed, a problem can often be understood as consisting of smaller problems, themselves composite, and so on until an atomic granularity is reached. In contrast to the top-down approach of Gestalt psychology, then, cognitive views seem to solve problems from the bottom up.

A classical example is the tower of Hanoi problem (cf. figure 2.1b). This problem can be represented as having an initial state (the initial position of the disks on the rods), a goal state (the desired final configuration). Operators allow for moving disks from one rod to another, according to the rules of the game, arriving in intermediary states. Solving the problem then consists in finding a sequence of operators leading from the initial state to the goal state.

This approach (unlike that of Gestalt psychology) lends itself straightforwardly to software implementation. One may progressively construct a graph of the state transitions enabled by the operators, such that a variety of search algorithms can solve the problem (e.g. A^*, (Hart et al., 1968)). But it suffers from a crucial limitation: the problem-space (states and operators) must be defined by the programmer, and no provision is given for transforming it in the course of problem-solving. If a problem-space makes use of inappropriate representations for the problem at hand, perhaps grouping features in an improper way, or excluding certain operators, finding a solution can become difficult or impossible.



(a)

(b)

Fig. 2.1 **(a)** Problem-solving for Gestalt psychology. In this square with a parallelogram strip across it, the lengths a and b are given. Find the sum of the areas of the square and the parallelogram strip; the (insight) solution is in the main text. *Illustration adapted from Wertheimer (1938).*
(b) Problem-solving for the cognitive tradition: the tower of Hanoi. In the initial state (I), all disks are on rod A. The solver can move the disks, for instance to intermediary state (II), but only one disk can be moved at a time, and no disk may be placed atop a smaller disk. The goal state is (III), in which all disks are on rod C. The problem can be solved by iteratively constructing the transition tree until a suitable "solved" state is found. *Illustration reproduced from Lucas (1885).*

2.1.3 Towards a unified theory

Both methods claim to be general theories of problem-solving; however they seem to have distinct scopes of application, being efficient on different kinds of problems.

Current research on insight, overwhelmingly, adopts the cognitive view, but recognizes the difficult challenge posed by insight[6]; a sign, perhaps, that something is missing in the cognitive picture. A key question in contemporary discussions (e.g. Weisberg, 2015) is whether a "special process" is needed to account for restructuring (special process theory); or whether, instead, a "standard" analytic search procedure or associative learning principles (Shettleworth, 2012) could account for it (business-as-usual theory). I believe this question is ill-formed and should be altogether abandoned: whatever mechanisms account for restruc-

[6] In a seminal article by Ohlsson (1984) that marks the renewal of psychological research on insight, he writes: "A unified theory of thinking should interpret restructuring in information processing terms, and explain the relation between restructuring and search."

turing, they must be so closely integrated with analytic or associative learning procedures that the question of whether they constitute a "special process" or "business-as-usual" is not likely to have a clear answer.

In this thesis, I will propose a model of insight that interleaves Thorndikian trial-and-error and Gestalt restructuring. But first, let us take a closer look at what is known about insight.

2.2 Animal studies

Animal psychology can often unveil fundamental properties of thinking by focusing the research effort onto "simpler" thinkers on the spectrum of animal intelligence, from insects to apes. Indeed, if one wishes to reverse-engineer biological thinking, it seems sensible to begin not with human cognition, but with life forms making use of a smaller number of the evolutionary building blocks of intelligence[7].

But the study of animal problem-solving has certain drawbacks with respect to investigating insight. Some of the most surprising properties of the insight phenomenon are its phenomenology (what insight feels like to the subject), which includes surprise, delight, and the sudden modification of internal representations. Because animals lack the ability to unambiguously report these feelings, much less the representations they are using, animal studies of insight are vulnerable to anthropomorphic misinterpretations. Indeed, much of insight research on animals has focused on figuring out whether seemingly "insightful" behaviors could not be better explained by associative learning (see Heyes (2012) for a relevant discussion of associative learning in animals).

In addition to these limitations owing to the difficult measurement of surprise and restructuring, the interpretation of the classical studies of animal insight has been colored by theoretical assumptions with regards to animal reasoning, behaviorism, and recently cognition as planning. In this section I focus on a handful of classic experimental results presented in chronological order, since they were intended as responses to one another in the wider context of these theoretical debates. By considering the results in their historical context, one can hope to better distinguish experimental outcomes from biases, and to build a clearer picture of the evidence accumulated so far.

[7] Moore (2004) offers a useful review of the evolution of learning mechanisms; but recent reviews on the evolution of "cognition" (e.g. van Horik and Emery, 2011) unfortunately limit their attention to *planning*-like views. See Wiggins et al. (2015) for a discussion of the evolution of creativity.

2.2.1 Early conditioning studies

In order to find out what "sort of thinking" animals were capable of, Thorndike designed the following protocol:

> Dogs and cats were shut up, when hungry, in inclosures from which they could escape by performing some simple act, such as pulling a wire loop, stepping on a platform or lever, clawing down a string stretched across the inclosure, turning a wooden button, etc. In each case the act set in play some simple mechanism which opened the door. A piece of fish or meat outside the inclosure furnished the motive for their attempts to escape.

Thorndike (1898b, 1899) concludes that animals are not capable of reason, but instead are creatures of habit: "little by little, the one act becomes more and more likely to be done in that situation, while the others slowly vanish." Their behavior corresponds to "the wearing smooth of a path in the brain, not the decisions of a rational consciousness"; even though appearances can be deceptive: "people who witnessed the performances of my animals after they had fully learned a lot of these acts, but had not seen the method of acquisition, all unanimously wondered at their wonderful intellectual powers."

Thorndike's experiments on operant conditioning, together with the work of Pavlov (1927/2010) on classical conditioning, paved the way for an experimental psychology seeking to find mathematically expressed laws of thought and behavior, such as, for instance, the Rescorla-Wagner model (Rescorla et al., 1972).

2.2.2 The first insight studies: Köhler

In the opinion of Köhler, "there has arisen among animal psychologists a distinct negativist tendency, according to which it is considered particularly exact to establish non-performance, non-human behaviour, mechanically-limited actions, and stupidity to animals."

Köhler sets up to experiment on the most human-like of animals, chimpanzees. The experiments are inspired by Gestalt ideas: "All of the experiments [...] are of one and the same kind: the experimenter sets up a situation in which the direct path to the objective is blocked, but a roundabout way left open." A great number of experiments is conduced, during which Chimpanzees repeatedly display an ability to solve problems quickly, with limited trial-and-error, usually learning from a single trial. In perhaps the most famous example, a chimpanzee named Sultan uses a box to reach a banana (Köhler, 1921, chapter 2):

> The objective [a banana] was nailed to the roof in a corner, about two and a half metres distant from the box. All six apes vainly endeavoured to reach the fruit

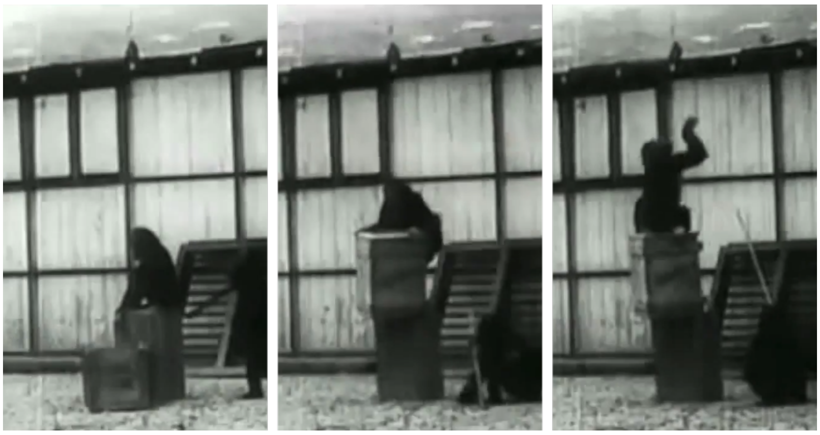


Fig. 2.2 From left to right, one of Köhler's chimpanzees performs the final motions to solve an especially difficult task: the chimpanzee brings the second box on top of the first, and stands on top of it to retrieve the banana which had been fixed to the ceiling. The frames are extracted from Köhler's footage of his experiments.

> by leaping up from the ground. Sultan soon relinquished this attempt, paced restlessly up and down, suddenly stood still in front of the box, seized it, tipped it hastily straight towards the objective, but began to climb upon it at a (horizontal) distance of half a metre, and springing upwards with all his force, tore down the banana. About five minutes had elapsed since the fastening of the fruit; from the momentary pause before the box to the first bite into the banana, only a few seconds elapsed, a perfectly continuous action after the first hesitation. Up to that instant none of the animals had taken any notice of the box; they were all far too intent on the objective; (...).

The difference with Thorndike's box experiments is striking: the chimpanzee appears to be attempting a full-blown strategy rather than doing a large number of random, instinctive movements in a trial-and-error fashion. The solution is immediately learned (the next day, Sultan solves the problem readily). Köhler calls the behavior "insightful".[8]

2.2.3 "Replications" by Epstein and Birch

The feats of Sultan, though perhaps not particularly impressive to contemporary readers accustomed to the idea of clever animals or of animal cultures, were considered "extraordinary"

[8] Towards the end of his life, Köhler (1959) defined insight as follows: "In its strict sense, the term refers to the fact that, when we are aware of a relation, of any relation, this relation is not experienced as a fact by itself, but rather as something that follows from the characteristics of the objects under consideration," where the relation is not necessarily causal – it can be, for instance, the comparative "oddity" between objects. See Loesche (2018) for a discussion.

and remained controversial for decades (see the excellent article by Ruiz and Sánchez (2014) for a contextualization of Köhler's work). This attracted work that sought to "demystify" Köhler's results. We consider two such lines of work, one by Birch (1945) and one by Epstein (1984)[9]. I begin with Epstein's work.

Epstein et al. (1984) endeavored to show that he could "reproduce" Köhler's experiment cited above, using mere pigeons. Epstein trained the pigeons for up to two months. Among other things, this involved rewarding the pigeons for pushing the box in random directions, then only in the direction of a "green spot" of diameter 4cm, with the banana not present. In separate sessions, the pigeons were also rewarded for climbing onto the box and pecking the banana with the box positioned appropriately. Finally, the birds having successfully graduated from this preparatory course were presented with a situation in which they had to push the box towards the banana in order to climb on it and peck the banana. Epstein notes that:

> Pigeons that had acquired relevant skills solved the problem in a remarkably chimpanzee-like (and, perforce, human-like) fashion.

Epstein emphasizes that "we did not train the birds to push the box towards the banana", and therefore the pigeon's behavior was "genuinely novel". Echoing Thorndike's comments 85 years earlier, Epstein notes that people observing only the result of this training attributed "a wide range of human thoughts to the pigeons".

There are substantial differences between the experiments of Epstein and Köhler[10]. Epstein nevertheless concludes that their "insightful" problem solving is simply the chaining of learned skills. Thus intelligent behavior, in pigeons, apes, and probably also in humans, is the result of the rule-controlled application of the skills present in the agent's repertoire.

The other major work following up on Köhler's chimpanzee studies is by Birch (1945). Birch also re-enacted some of Köhler's experiments, but on chimpanzees born in captivity and having lived under controlled conditions. Birch found that chimpanzees who never had played with sticks struggled to solve basic puzzles of the sort used by Köhler involving the use of a stick (although two out of six did succeed, their success did not seem insightful, but resembled trial and error behavior). Birch then let the chimpanzees play with sticks for several days; after which they were given another opportunity to solve the problem. This time, all chimpanzees readily solved the problem.

[9] In Epstein's case, the declared objective was also to "further Skinner's longstanding campaign against cognitive psychology" and in favor of accounts in terms of "contingencies and reinforcement" (Epstein, 1991, p. 366).

[10] Besides the obvious difference that pigeons received training, one may remark that the birds had been trained to push the box towards the only remarkable object in sight; that the green dot may have looked, to a pigeon, similar in color to a yellow banana; and so on.

Birch's conclusions, which can be seen as an early expression of many of the views expressed in this thesis[11], were:

1. That the perception of functional relations in a situation is dependent in large part upon the previous experiences of the animal.
2. That insightful problem solution represents the integration into new patterns of activity of previously existent part-processes developed in the course of the animals' earlier activities.
3. That any interpretation of insight in situational terms alone[12], or even predominately in situational terms, is invalid.
4. That the functional relationships which are perceived in a given problem by the animal, and which serve as the basis for an insightful response, are the product of the dynamic interaction of a) the available repertoire of experiences (superimposed upon the basic species characteristics), with b) the objective features of the situation.
5. That in insightful problem-solving, in contrast with trial-and-error solution, previous experience provides the materials out of which an adequate pattern of response may be fabricated, rather than the stereotyped problem-solving response itself.

2.2.4 Animal insight then and now

In recent years, the debate on "insight" has moved to the case of corvids (Bird and Stokes, 2006; Bird and Emery, 2009; Taylor and Gray, 2009; Taylor et al., 2012), after surprisingly intelligent displays of problem-solving behavior were seen in certain New Caledonian crows and other members of the corvid family (for instance the crow named Betty, which I mentioned in chapter 2). New experimental setups, for instance one inspired by Aesop's fable "The Crow and the Pitcher", are being applied on various animals including corvids (Jelbert et al., 2014), apes (Hanus et al., 2011), and most recently raccoons (Stanton et al., 2017); other setups have been used for rats (Maier, 1929, 1931; Neves Filho et al., 2015) and even elephants (Foerder et al., 2011).

The eventful history of animal insight is profoundly affected by changing theoretical backgrounds and research goals (see Shettleworth (2012) for a recent overview of this research). Thorndike showed that animals lack reasoning powers; Köhler sought to show

[11] Also see Hebb's take on insight (Hebb, 1949, p165) for a more neurological presentation of similar views.

[12] This is the Gestalt view of restructuring "from scratch".

that animals can display intelligence resembling that of humans; Epstein found, in his results, evidence that animals and humans alike obey behavioral laws. Other relevant work[13], not reviewed above, received limited attention, perhaps because it had less to say about the scientific controversies that occupied the center of the scene at the time.

This continues today; in the most recent work, the underlying question has become: are animals capable of a *causal* understanding of their environment (Taylor et al., 2012)? Seed and Boogert (2013) argue that "perhaps comparative psychologists need to re-visit what a test of 'insight' is trying to capture". The concept of insight, in animal experiments, continues to be used to attack or defend paradigms concerning animal intelligence, human intelligence, and intelligence in general. This leads to interpretations of results that are sometimes only tangentially relevant to the insight phenomenon itself.

Looking at the accumulated evidence obtained via animal studies, and ignoring the controversies, a clearer picture of animal insight emerges[14]. With varying levels of ability, different animals demonstrate the capacity to rapidly solve novel problems. In the case of pigeons, very precise component behaviors must be drilled for weeks immediately preceding the test; whereas an ape solves the "same" problem by quickly and successfully adapting a skill that perhaps was previously used a long time before, and/or in different circumstances. In such cases, the animals display some characteristic features of human insightful behavior: particularly, a sudden and successful switch to a new behavioral pattern, often including paying attention to environmental features that had until then been ignored (e.g. the shape of the tool for Betty, the box for Sultan).

2.3 Human studies

In the 1980s, insight research experienced a revival, coinciding with the decreasing influence of the behaviorist paradigm. This revival consisted in a switch to experiments on humans, with considerable reliance on self-report.

One advantage of working with humans is their ability to report on their own mental processes, either during or after problem-solving. Even though self-report does not suppress all the difficulties related to the study of insight (Ash et al., 2009), and is not without its own methodological complications (Ericsson and Simon, 1980; Schooler et al., 1993), it is nevertheless easier to interpret than the attitudes of an animal. With the increased attention given to cognition over behavior (cognitivism), it was therefore quite natural that the study of

[13] Of particular interest is the work of Tolman and Honzik (1930) on rats. Rats displayed good performance at a rewarded task, compared to control rats, after being allowed to explore a maze without receiving any sort reinforcement: this suggest that they could learn without receiving rewards.

[14] Also see the summary and conclusions of Shettleworth (2012).

insight should turn to human subjects. These subjects could report their progress in real time, could be asked about their impressions of progress or pleasure, or about what representations they experienced, and so on; types of data that behaviorists often did not collect. (New behavioral data was also collected, notably gaze (Knoblich et al., 2001).) This resulted in the discovery of new features of insight, discussed below.

2.3.1 The insight sequence

The first such feature is the temporal pattern of insight. Insight appears to roughly obey the "insight sequence" (Ohlsson, 2011; Weisberg, 2015):

1. Search in a problem space
2. Consistent failure or "impasse" (but the necessity of this stage is disputed; see Fleck and Weisberg (2013), Danek et al. (2014), and Webb et al. (2016)).
3. Restructuring, and solution or significant progress
4. Test of perceived solution

One of the important characteristics of insight problem-solving is the unpredictability and suddenness of the third step, the "Aha!"-moment. The problem-solver becomes aware that significant progress has been suddenly achieved, in contrast to a slowly increasing "feeling of warmth" for non-insight solutions (Metcalfe and Wiebe, 1987). This is sometimes a mistaken impression (Danek and Wiley, 2017), hence the necessity to test it in the fourth step.

Search, failure, or testing are easy enough to understand, even when it is not clear precisely how humans might perform them. Even sudden progress can be explained as a chance discovery as part of a planning process. The mystery of insight lies in *restructuring*.

2.3.2 Types of restructuring

What is restructuring? In a cognitive approach, one may distinguish aspects of the representations that could be subject to modification. Several such modifications of the problem space have been empirically observed:

1. The heuristics used to select operators (Kaplan and Simon, 1990);
2. The chunking of perceptual elements into objects (Knoblich et al., 2001) (cf. figure 2.3);

3. “Hard” constraints on available operators (MacGregor et al., 2001).

Whatever its form, restructuring can be triggered either from new information obtained in the course of problem-solving[15], or from failure to succeed in the initial problem-space.

The cognitive view, then, seems to liken restructuring to changing some parameter of the search-space in which the problem-solver is working. In the context of computational creativity[16], related ideas were formalized by Wiggins (2006b), for whom search, if it includes search at the meta-level, can be a valid simulation of human creativity. However, if insightful restructuring consisted of meta-search, then one might expect the change of search space to be followed by additional search, rather than being seemingly simultaneous with the discovery of the solution; especially since some insights alter the problem representation radically, rather than just one aspect of it. How could meta-search change so much about the problem perception, yet arrive immediately at a correct solution?

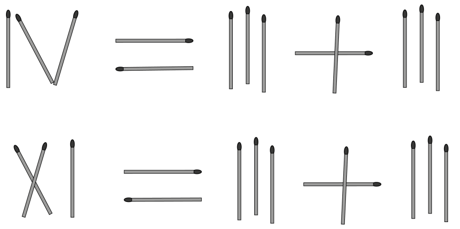


Fig. 2.3 Two of the problems used in Knoblich et al. (2001). The objective is to transform an incorrect equation into a correct one by moving a single matchstick, where forming the sign “$\neq$” is prohibited. The solution to the first problem consists in changing “IV” into “VI”, and in the second problem, “XI” into “VI”. The second problem proves to be more difficult for participants. The authors argue that this is due to “chunking”: subjects make two chunks for “IV”, whereas “X” is perceived as a single element, thus inhibiting search on the latter.

In Gestalt theory restructuring was thought to require viewing the problem naively – rejecting the contribution of experience (Koffka, 1935). But this view seems even less tenable, considering the computational complexity of solving such problems. Moreover, it fails to account for experimental evidence showing that insight is more likely to occur with prior experience (Birch, 1945; Harlow, 1949; Stickgold and Walker, 2004; Wiley, 1998) and may even be impossible to subjects lacking domain experience. For example, Weisberg and Alba (1981) report a success rate below 5% on the 9-dot problem (cf. figure 2.4), even when preventing fixation on an incorrect solution strategy; whereas Kershaw and Ohlsson

[15] In which case problem-solving has still been described as “analytical” as opposed to corresponding to a “special process” Weisberg (2015); although it is widespread, I do not think the distinction is meaningful, since the two categories do not seem to be mutually exclusive.

[16] Rather than insight specifically.

(2001) demonstrated 89% success on variations of the 9-dot problem following acquisition of relevant experience, compared to 18% for controls.

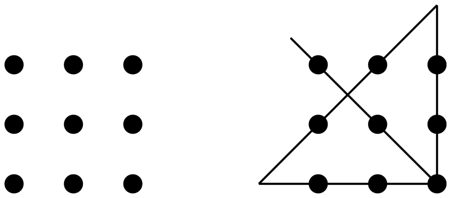


Fig. 2.4 Left, the 9-dots problem: subjects are instructed to draw four straight lines connecting all dots, without lifting their pen. Right, the solution, which involves drawing lines outside the square formed by the dots and turning on non-dot points.

2.3.3 Other findings

Above, I have reviewed results about insight acquired within the gestalt and cognitive paradigms. Below I present pell-mell results that do not fall neatly into either paradigm, but nonetheless have contributed to the present understanding of insight.

Relationship between insight and mood

The effect of mood on insight seems to be important, but remains poorly understood. The findings are inconsistent, with e.g. Suzanne K. Vosburg (1997) finding that negative moods improve performance on insight problems, whereas e.g. Isen et al. (1987) observed an opposite effect. Both findings were replicated. The findings are, however, not necessarily contradicting, as the methodology and the problems used differed; they may simply reflect a complex relationship between mood and insight, potentially related to attentional mechanisms (Subramaniam et al., 2009).

Sleep is conducive to insight

Another branch of research has investigated the role of sleep into insightful problem-solving. It found a strong effect on the proportion of insightful solutions for participants allowed to “sleep on a problem” (Stickgold and Walker, 2004; Verleger et al., 2013; Wagner et al., 2004). This can be speculatively related to machine learning concepts (such as for instance the wake-sleep algorithm (Hinton et al., 1995)).

Historic insights and first-hand accounts

Finally, some researchers have focused on historic and first-hand accounts of insight; see for example Simonton (2012) for a study of Galileo's discoveries. Other famous insights include Archimedes' original (but apocryphal) "Eureka!" moment (Pollio et al., 1914), Newton's apple (Stukeley, 1752), Kekulé's discovery of the structure of benzene (Rothenberg, 1995), etc. The mathematician Henri Poincaré (1909) in particular provides an exceptionally detailed and thoughtful account of his insights . Reviewing this is difficult, in part due to the frequent controversies surrounding these accounts and their interpretation. In appendix B, I have compiled a summary of several of these stories and, when available, the first-hand accounts.

2.4 Insight and Reinforcement Learning

The study of insight, from its inception, interacts with and challenges approaches in terms of associative learning, of which reinforcement learning is a descendant: Köhler challenges Thorndike, and is in turn challenged by Epstein. With the advent of cognitivism, insight becomes likewise a challenge for cognitive explanations. Considering the origins of reinforcement learning within the cognitive sciences, it seems natural to confront theories of RL with the insight phenomenon. Insight is a recalcitrant counter-example to grand theories of cognition, demanding explanation.

Despite that, there is very little work at the intersection of insight research and reinforcement learning research. Recent reviews of insight neuroscience do not mention reinforcement learning at all (Dietrich and Kanso, 2010; Kounios and Beeman, 2014, 2015; Shen et al., 2017; Sprugnoli et al., 2017), nor do other works on psychological insight (Ohlsson, 2011). Indeed, even connections between reinforcement learning and creativity are rare (the *Blind Variation, Selective Retention* model (Campbell, 1960b; Simonton, 2010) for instance is usually compared to biological evolution, rather than to reinforcement learning, despite some obvious analogies), with few exceptions (e.g. Smith and Garnett (2012b); Vigorito and Barto (2008)). Even studies that focus on the link between dopamine and creativity (e.g. Boot et al., 2017; Lhommée et al., 2014; Zabelina et al., 2016) rarely mention reinforcement learning.[17]

This section makes the case for seeking a reinforcement learning theory of insight.

[17] With the occasional exception such as Stahlman et al. (2013), but with little engagement with the AI literature on RL.

2.4.1 Machine learning and creativity

Reinforcement Learning is a promising field with respect to creativity generally speaking, because some of its most important concepts are analogous to the *novelty* and *usefulness* criteria of creativity: exploration for novelty, reward for usefulness. Indeed, whereas supervised and unsupervised learning (see Goodfellow et al. (2016) for an introduction) can produce novel and useful output, they fundamentally seek to generalize from a training set, rather than to try new things whose value cannot be inferred from this training set. Hence products of supervised and unsupervised learning are limited by this training set: with rare exceptions (such as active learning (e.g. Wang et al., 2017)) there is no exploration mechanism that could yield new training data and, thereby, interestingly novel outputs.

Optimization techniques are perhaps better equipped to be creative. Such techniques include Reinforcement Learning, which seeks to maximize an agent's return as it interacts with its environment, or Evolutionary Algorithms (EAs), which rely on the fitness of agents to guide the search for a good solution. At the core of these algorithms is a utility measure which should be maximized (return for RL, fitness for EAs), as opposed to an error measure which should be made close to zero – allowing for a more open-ended learning process. In addition, they contain explicit mechanisms to discover novelty: mutation and crossover for EAs, exploration in RL. This contrasts with paradigms which explore parameter space exclusively by slowly following a gradient[18].

Moreover, Reinforcement Learning techniques are characterized by their ability to learn based on estimates of future utility. Because of this, adjustments can be made on-line in the course of interaction. This corresponds to the improvisational, interactive nature of human creativity (Glăveanu, 2013; Ingold, 2014): humans do not wait for the finished result or product to evaluate it (contrary to most optimization techniques, for instance genetic programming (Koza et al., 2004)). Most RL algorithms make explicit use of progress estimates, learning "online" without waiting for the final result. This is important to achieve good performance in large problem spaces, where experience is costly and episodes take considerable time to complete; many creative domains have these characteristics.

Reinforcement Learning, then, seems well-equipped to model creative behavior. However, I am interested more specifically in one manifestation of creativity, insight.

[18] Although techniques such as dropout (Srivastava et al., 2014) can arguably be seen as active forms of exploration in parameter space in a supervised context.

2.4.2 RL at the crossroads of representation and search

In the psychology of insight, a major hurdle for cognitive approaches (such as those defended by Weisberg (1986) or Newell and Simon (1972)) is the *representation* of the problem. Insight is challenging for the cognitive paradigm because it seems to operate on the (very large) representation space, rather than on the typically smaller state-space. (That is, whereas Newell and Simon's operators change the state, insights seem to transform the representation which defines all at once states and operators.) Discrete, tabular versions of reinforcement learning algorithms have much in common with the search paradigm exemplified by Newell and Simon and presented in section 2.1: the operators of Newell and Simon correspond to the actions of RL; their knowledge states, to states; and goal states can be emulated using rewarding terminal states.

But unlike Newell and Simon's approach, the reinforcement learning paradigm is designed to work with function approximation (Sutton and Barto, 2018). Recently, spectacular results were achieved with non-linear function approximation in Atari (Mnih et al., 2013, 2015), Go (Silver et al., 2016), and Starcraft II (Vinyals et al., 2019). This is largely due to advances in artificial neural networks, which have made progress towards finding good representations (Bengio et al., 2013a). Contemporary RL algorithms (e.g. Mnih et al. (2013)) make use of artificial neural networks to generalize between similar states, while still viewing problem solving as investigating trajectories in a state space.

Perhaps the challenge of the insight phenomenon, from a machine learning perspective, is to bridge the gap between search in a state-space (corresponding to the cognitive paradigm) and discovery of good representations (corresponding to the Gestalt paradigm). Reinforcement learning is ideally positioned to answer this challenge, having at its disposal the tools necessary to search along trajectories within a state space, as well as those needed to discover good representations via function approximation.

2.4.3 Experimental evidence

Besides this promising potential, there is experimental evidence suggesting a link between reinforcement learning and insight. For this section I will assume a broad familiarity with psychological and neuroscientific theories inspired by RL; appendix C reviews the reinforcement learning theories used in psychology and neuroscience, which are closely tied to the RL paradigm in AI.

Fundamental to RL as a psychological paradigm is the notion of temporal difference errors: an agent should learn based on unexpected changes in the evaluation of a situation. After an action, if the new situation is more promising (alt. less promising) than expected, the tendency to repeat that action in similar contexts should be reinforced (alt. weakened). The subjective feeling of insight is characterized by an unexpected, sudden rise in the perception of progress (Metcalfe and Wiebe, 1987): this corresponds to positive temporal difference error in a reinforcement learning model, and suggests a possible explanatory role for RL[19].

Much of the fMRI and EEG evidence suggests that the executive control network, involving primarily the frontal lobe, plays a crucial role in insight (Sprugnoli et al., 2017). These are the brain regions and the functions which the neuroscientific RL paradigm seeks to model. Although insight correlates with activation in several frontal areas, the most replicated result is ACC activation, which might be "initiating processes that lead to the breaking of the mental mindset that keeps one stuck in the wrong solution space" (Dietrich and Kanso, 2010). Influential reinforcement learning theories of the ACC are consistent with this view (Holroyd and Yeung, 2012; Shenhav et al., 2013). In particular, Holroyd's view of the role of the ACC in implementing biological Hierarchical Reinforcement Learning is highly suggestive of the theory of insight of Ohlsson (2011)[20] (summarized in appendix C.1). For an introductory account of the neuroscience of reinforcement learning and insight, see appendix D.

A recent study Tik et al. (2018) was the first, and the only at the time of writing, to use a 7 Tesla scanner to investigate insight. Perhaps thanks to this increased precision, this study found activations in the substantia nigra, ventral tegmental area, the thalamus, and the striatum (including caudate and especially nucleus accumbens), all consistent with the reinforcement learning subcortical network (see appendix C) that would be expected to activate in the event of a positive temporal difference error (Liu et al., 2011). Although these results will need to be confirmed, they contribute to the body of evidence already discussed so far.

2.4.4 Hierarchical reinforcement learning

We have presented several indicators suggesting a connection between insight and reinforcement learning. Perhaps this connection can be further narrowed down, relating insight with *deep, hierarchical* reinforcement learning. In hierarchical reinforcement learning, an agent possesses multiple policies[21], between which it switches according to a higher-level policy. Such a switch between policies would radically alter the search process and the chances of success of the agent. Furthermore, if different policies are implemented as distinct

[19] Unfortunately, there have been no studies investigating the role of dopamine in insight problem-solving. A reinforcement learning theory of insight would surely predict a spike in dopaminergic activity during the "Aha!" moment; but this has never been specifically tested.

[20] Ohlsson's proposes that restructuring constitutes redistribution of activation following an "option" change, unknowingly using the same word as Sutton's in his "option" framework.

[21] A policy can be understood as a function controlling the agent's behavior.

neural networks, this would result in a simultaneous transformation in the perception of the environment; and because this new perception is based on relevant experience, it might be immediately, or at least rapidly, successful. There already exists a reinforcement learning architecture capable of much of this: the option-critic architecture of Bacon et al. (2017), a fully differentiable hierarchical actor-critic system. This suggests a path forwards for modeling insight.

I have discussed this approach in greater depth in Colin (2017); in this thesis, I leave extended discussion of it to chapters 4 and 5, in which experiments are conduced to verify this theory.

2.5 Conclusion

2.5.1 Summary

- Insight is the transition undergone by a problem-solver who suddenly and unexpectedly comes to know how to solve a problem.

- There is about a century's worth of research on the insight phenomenon, beginning with Köhler (1921) work with chimpanzees. Major strands of research have included Gestalt, behaviorist, and cognitive psychology, experiments on animals and humans, and neuroscience.

- The essential features of insight are as follows:

    1. The insight sequence: search – (impasse) – restructuring – verification (Ohlsson, 2011; Weisberg, 2015).
    2. Insights are sudden and pleasantly surprising to the problem-solver (as measured with "feeling-of-warmth" ratings (Metcalfe and Wiebe, 1987)).
    3. Restructuring involves changes in "chunking" (Knoblich et al., 2001), in the heuristics used (Kaplan and Simon, 1990), and in the constraints on operators (MacGregor et al., 2001).
    4. Insight depends on previous experience Wiley (1998) and is facilitated by sleep (Wagner et al., 2004).
    5. Insight involves most importantly the superior and medial temporal gyri and the anterior cingulate cortex (Sprugnoli et al., 2017), the latter suggesting an important role for executive control and attention networks.

- Parallels between insight and RL suggest that insight could be explained by RL mechanisms, but until now this connection has not been noticed or explored.

Contributions

Sections 2.1 through to 2.3 are a literature review, containing no novel contribution beyond the summarizing of state-of-the-art research. Section 2.4 is, to our knowledge, the first to discuss a link between reinforcement learning and insight (save for our own article Colin (2017)).

2.5.2 Bibliographical remarks

The Gestalt view of insight is best discussed in the works of Köhler (1921) and Wertheimer and Wertheimer (1959). Köhler's work constitutes simultaneously the origin of insight research and the first studies of insight on animals. However, there have not been influential new Gestalt studies on insight since the 1950s. Nonetheless recent work by Friston et al. (2017) and less directly by Schmidhuber (2010) (see appendix C.4), in some ways resemble a revival of Gestalt perspectives, albeit harnessing subsequent progress in information theory.

The cognitive research program on insight is probably most explicitly stated by Ohlsson (1984), which also denotes the regain of interest in human studies of insight involving self-reports, accompanying the loss of popularity of behaviorism. This research program has been ongoing since then, see e.g. Danek et al. (2013); Kaplan and Simon (1990); Knoblich et al. (2001); Metcalfe and Wiebe (1987), etc.

The neuroscience of insight is a younger discipline, initially dominated by the research of Kounios and Beeman (2014). As more groups joined the research effort, other views emerged. The review presented here is based for the most part on review articles, including (in addition to Kounios and Beeman (2014)), Sprugnoli et al. (2017), Dietrich and Kanso (2010), and Shen et al. (2017).

The discussion of RL in relation to creativity and insight is original. Very little has been written on the relationship between RL and creativity; among the rare exceptions are the article by Vigorito and Barto (2008) on using hierarchical reinforcement learning for creative search and a chapter by Smith and Garnett (2012b) on using intrinsically motivated RL for music improvisation (but neither mentions insight). For a more psychological perspective (not including the RL framework from AI), see for instance Stahlman et al. (2013) which discusses the importance of conditioning in the creative process. Little has been written on the relationship between insight and reinforcement learning, although there is considerable discussion in terms of operant conditioning (but such work sometimes contests the existence of an “insight” phenomenon by ignoring or denying representational change; see e.g. Epstein (2014)).

Chapter 3

Reinforcement Learning

In this chapter, I introduce the technical backbone of this thesis: the field of Reinforcement Learning, towards which some of the main contributions of this thesis are made. I introduce the formalization of RL problems as Markov Decision Processes (MDPs), and several techniques for solving them. Because this thesis is intended for an audience beyond RL specialists, this chapter is written as a tutorial introduction rather than as a mere statement of the notation. This chapter has two objectives:

1. Establishing the theoretical bases and the notation needed in the experimental chapters.
2. Introducing the challenges in Reinforcement Learning research to which this thesis makes contributions.

This chapter is divided into two sections corresponding respectively to the two objectives above: firstly introducing basic concepts, and secondly presenting advanced ideas and challenges. The first section introduces *MDPs* (the framework for describing reinforcement learning problems), *policy evaluation* (methods for evaluating how "good" the agent's situation is, depending on the agent's policy), and *policy improvement* (methods for improving the agent's policy). The second section discusses function approximation (a method for dealing with MDPs with a large number of states), actor-critic algorithms, exploration, and hierarchical reinforcement learning.

Although this chapter aims to cover all the concepts and techniques necessary to understand the rest of the thesis, there exist alternative introductions to reinforcement learning. The article by Kaelbling et al. (1996), although dated, is an excellent, accessible introduction, as is the video lecture series by Silver (2015)[1]. The up-to-date reference for reinforcement learning is the second edition of the book by Sutton and Barto (2018).[2].

[1] Available online at: https://www.youtube.com/watch?v=2pWv7GOvuf0

[2] Also available online, at: http://incompleteideas.net/book/the-book-2nd.html

At the end of the previous chapter, I briefly discussed parallels between reinforcement learning and insight. Unfortunately, this discussion was limited, as I had not yet introduced the field of RL. Before concluding the present chapter, I will provide further details on this analogy, in light of a closer look at the RL paradigm and RL techniques.

3.1 The Reinforcement Learning paradigm

Reinforcement learning is learning what to do, how to map situations to actions, in order to maximize a reward signal. Reinforcement learning is further distinguished by the necessity to explore by trial-and-error (Sutton and Barto, 2018), and by the credit assignment problem (Minsky, 1961) which arises from delayed rewards.

Reinforcement learning methods have been successful in challenging tasks, including most recently playing Atari games (Mnih et al., 2015) or the game of Go (Silver et al., 2016). These methods are often capable of learning with limited pre-processed external input, compared for instance to supervised learning, making them a promising model for human and animal intelligence. Indeed, there are connections between RL and biological intelligence. The reinforcement learning temporal difference (TD) model in psychology can be viewed (Sutton and Barto, 2018) as an extension to Rescorla-Wagner's model (Brisch et al., 2014; Rescorla et al., 1972). Several reinforcement learning concepts, born in an artificial intelligence context, have since found their way back in neuroscience, as discussed in appendix D.

In the context of AI, Reinforcement learning is not only a class of problems, nor a class of solutions for these problems; it is also a field of research (with interdisciplinary connections to psychology, neuroscience, operations research, as well as other subfields within machine learning), and the intellectual home of a community of researchers. It has therefore developed a shared formal language. This chapter introduces the problems, the known solutions, and the expression of these problems and solutions in the notation of Sutton and Barto (2018), in which all the technical contributions of this thesis are written.

3.1.1 The MDP framework

Reinforcement Learning problems are often formalized as finite Markov Decision Processes (MDPs). MDPs can be understood by contrast with Markov Processes and Markov Reward Processes; all have the *Markov property*. Having the Markov property means that the future does not depend on the past, given the present: "the present state must include information about all aspects of the past agent-environment interaction that make a difference for the future" (Sutton and Barto, 2018). Below, I successively describe MPs, MRPs, and MDPs. Figure 3.1 illustrates these differences in a simple example.

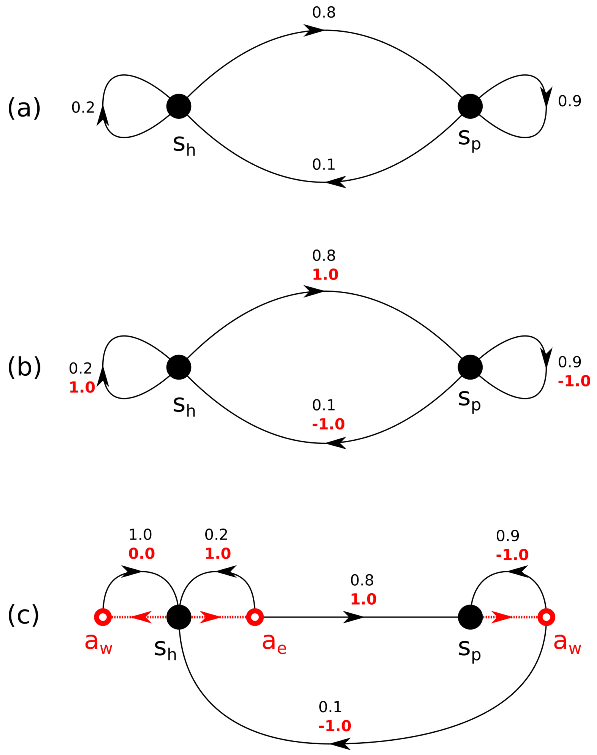


Fig. 3.1 From Markov process to Markov Decision Process:
(a): a Markov process with two nodes s_p and s_h representing the states, and four weighted edges representing transitions and their probabilities. This can be viewed as a description of an ape eating food that has gone bad: the healthy ape (state s_h) eating food will, with probability 0.8, transition to the food poisoning state (s_p). From s_p, at every time step, the ape recovers with probability 0.1, in which case it transitions back to s_h, and so on. This is a predictive model, modeling neither rewards nor decisions: it is not suitable for evaluation or control.
(b): a Markov Reward Process (MRP) with the same states. Rewards (bold, red numbers) are now associated with state transitions. This representation introduces an evaluative signal: the rewards incurred as the process unfolds in time. An MRP can be used for policy evaluation.
(c): a Markov Decision Process (MDP). There now are action nodes (red nodes) in addition to state nodes. The healthy ape at s_h can now decide to eat (a_e) or to wait (a_w). Eating incurs a reward of 1.0 and causes with probability 0.8 a transition to s_p, from which the ape can only wait for recovery, while receiving negative rewards for every time-step spent in the food poisoning state. This formulation distinguishes aspects of the systems that depend on the decision-maker from those that depend on the dynamics of environment.
If the ape adopts the policy of always eating whenever possible, then the evolution of the MDP shown in **(c)** can be predicted with the Markov process at (a); that policy can be evaluated based on the Markov Reward Process at (b).

Markov Processes

A Markov process is a descriptive/predictive model of the evolution of a system through time. A discrete Markov process could be fully described, for instance, by a finite set of states $\mathcal{S}$, and a probability $p(s,s') = \Pr(s'|s)$ of transitioning from a state s to a state s' at the next time step. Hence that process can be specified by the double $\langle \mathcal{S}, p \rangle$, where p is called a *transition function*. An example of a Markov process is shown in part (a) of figure 3.1. This process is useful to model the evolution of the state of the system through time, but is insufficient to evaluate how "good" it is, or to control it towards better outcomes.

Markov Reward Processes

In contrast to Markov processes, the transitions of a Markov Reward Process (MRP) incur rewards r from a reward set $\mathcal{R}$. Hence for MRPs the transition function is $p(s',r,s) = \Pr(s',r|s)$ (in general rewards can be stochastic; however in figure 3.1 rewards are deterministic given the current and next state). Rewards are a rich way of defining what is good for the process - for instance, one may define a "goal state" as having a positive reward, or one may attribute special costs (negative rewards) to particular transitions. As it unfolds from a starting time-step t, gathering rewards at each time-step, an MRP produces a *return*: $\sum_{k=0}^{\infty} \gamma^k r_{t+k}$, where γ is a discount factor between 0 and 1 defining a preference for immediate over future rewards. The discounted return from timestep t onwards is denoted G_t (the capital R denotes a random/not yet observed reward):

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k} \tag{3.1}$$

Because of the Markov property, the expected return is always the same (regardless of history) for a given state; this is called the *value* of that state $v(s)$:

$$v(s) \doteq \mathbb{E}[G_t|S_t = s] \tag{3.2}$$

Markov Decision Processes

An MRP can be used to quantify how good or bad an agent's present situation is, given the dynamics of the system, but says nothing about *controlling* its trajectory. A Markov Decision Process introduces actions: actions specify what can be done to alter that process, so as to maximize expected return. Hence each transition is, as it were, divided in two parts: one that depends on the dynamics of the environment, and one that depends on the decision-maker or agent. In the graph (c) corresponding to the MDP in figure 3.2, this is shown by the introduction of action nodes, which result in two types of edges: from states to actions, and from actions to new states. The transition function now only correspond to edges of the second type, and therefore takes the form[3] $p(s', r, s, a) = \Pr(s', r|s, a)$. Thus an MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, p, \gamma \rangle$. An MDP by itself is insufficient for predicting the behavior of a system, since the actions taken by the agent are not known. Instead, an MDP is best understood as describing a *control problem* – that of finding a policy, that is, of finding out which action to choose, given the current state, in order to maximize return.

The MDP formulation allows for the representation of policies π which map states to actions (either stochastically, in which case $\pi(s, a)$ denotes the probability $\Pr(a|s)$, or deterministically, in which case $\pi(s)$ denotes the action that is to be taken in state s). If transitions are the physics of the system, policies are their controllable counterpart: the choices of the agent. One may view MDPs as representing the "freedom of choice" of the agent with respect to action.

Partially observable Markov Decision Processes

One important limitation is that, in many real-world cases, the agent does not have direct access to enough information to uniquely identify the current state of the process. This means that, from the perspective of a memory-less agent, the Markov Property usually does not hold. In such cases, one says that the MDP is *partially observable* (POMDP). One way to tackle such cases is to attempt to maintain memories of relevant past information to augment the state, such that the augmented environment (perceived state plus memory) "is Markov". This thesis does not engage much with POMDPs.

Summary: the MDP framework

A Markov Process can model the unfolding evolution of a system through time, but the primary concern of an agent is not *what will happen*, but *what to do*. This justifies the use of MDPs[4]. An MDP, compared to a Markov Process, distinguishes what depends on the agent (*actions* $\mathcal{A}$) from what depends on the environment (*transitions* $p(s', r, s, a)$). Furthermore, MDPs introduce *rewards* providing an evaluative signal, defining some complex goals or constraints. In short, a Markov Process describes the evolution of a system, whereas a Markov Decision Process is a decision problem which an agent faces.

[3] In general we give the same name to many action-nodes in the graph, because most agents have a certain number of fixed ways of interacting with the world (e.g. a robot has a certain set of actuators which rarely change depending on the state). Hence it is necessary to specify both the state and the action in order to uniquely specify an action node in the MDP graph.

[4] Alternatively called multistage or sequential decision processes.

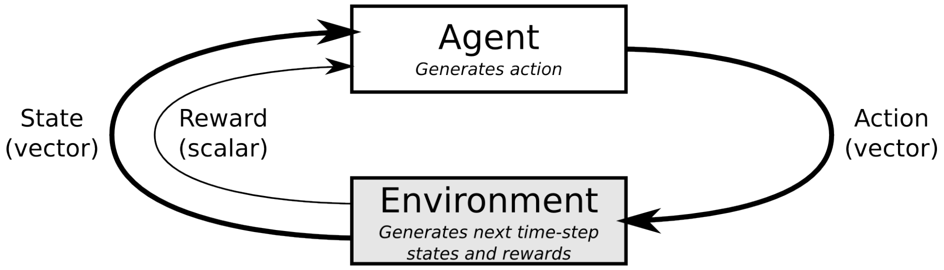


Fig. 3.2 The agent-environment interaction in an MDP, adapted from Sutton and Barto (2018). In some of the cases studied in this thesis, both states and actions will be vector-valued. Note that the border between agent and environment is not the same as the border between organism and environment: for instance, in an organism the reward function is implemented internally, but from a reinforcement learning perspective that function is *outside the control* of the agent, and therefore is considered part of the "environment".

An MDP consists of a quadruple $\langle \mathcal{S}, \mathcal{A}, p, \gamma \rangle$ defining a set of states, a set of actions, a transition function, and a discount factor respectively; this describes the interface between an agent and an environment. The dynamics of this environment include the production of an evaluative signal, the reward. An MDP thus describes the problem of selecting actions to maximize the expected discounted sum of rewards, or return, G.

3.1.2 The value function and the credit assignment problem

In reinforcement learning the central concern is perhaps the *credit assignment problem* Minsky (1961): "How do you distribute credit for success among the many decisions that may have been involved in producing it?"(Sutton and Barto, 2018, p17). For example, in a chess game it may not be clear whether the final victory was due to a move in the early, middle, or late game. To solve the credit assignment problem, part of the answer is to evaluate intermediary positions, or states, using a value function $v : \mathcal{S} \rightarrow \mathbb{R}$.

The contribution of the value function to credit assignment is illustrated, in this subsection, by way of an extended example. Imagine an office worker seeking to optimize her commute. She does not like walking or cycling, especially when carrying a bag along, but likes to read (even mediocre books are better than nothing). However she suffers from motion sickness, so she can read only in the subway - she cannot read in the bus without suffering significant discomfort. We further assume that the agent begins solving the problem with no prior knowledge of the transitions (she never gets to see the actual graph of the MDP[5]).

[5] If the commuter uses her experience to progressively build such a graph, and then uses that graph to improve her behavior, this is model-based reinforcement learning. In this thesis I focus on model-free reinforcement learning. Some additional discussion of model-based techniques is found in chapter 7.

What is the least annoying or best commute? This question is fully specified in the MDP graph[6] shown in figure 3.3. Let us assume that the agent has some policy to start with; a default approach to the problem, which does not need to be any good. Two examples of such starting policies are shown in figures 3.4 and 3.5. In figure 3.4, the agent randomly (1/2) picks a book or doesn't, then cycles to the subway, reads in the subway if she had a book, and finally walks to the station. In figure 3.5, the agent makes all decisions uniformly at random (i.e. if there are three possible decisions, the agent picks each choice with probability 1/3). Because some of the transitions are stochastic (in this case, only the quality of the book), it may be safer not to rush to judgment from only a few observations; instead making progressive changes. The general approach of most RL algorithms, then, is to proceed by statistical trial and error, slowly changing one's policy for the better.

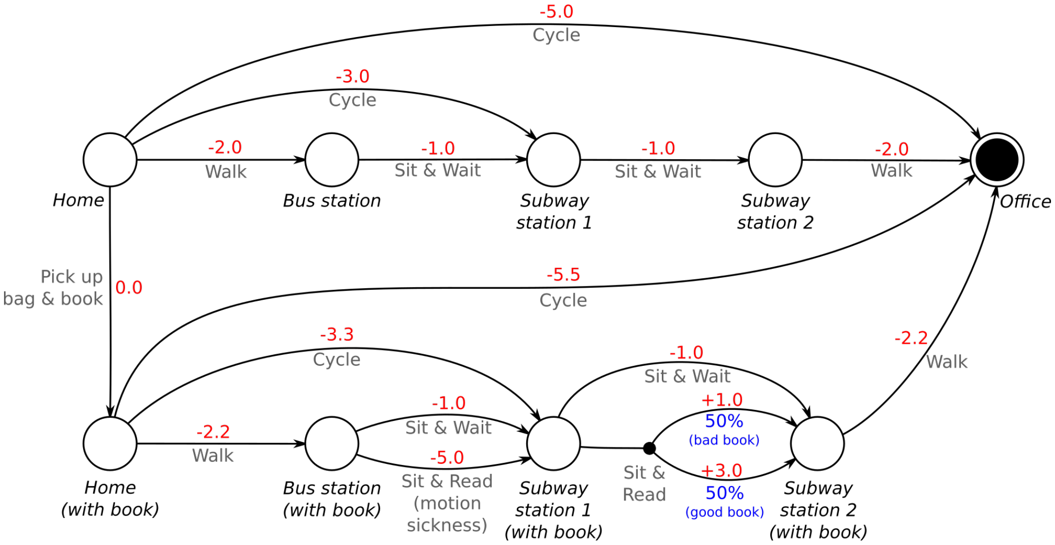


Fig. 3.3 A commute optimization MDP. The state "office" is terminal.

How can the commuter progressively improve its policy? A tempting solution would be to only ever change one action at a time, and test whether the end result (the total reward accumulated) was better or worse. However, this is not feasible when MDPs are large, or when the problem cannot easily be separated into independent episodes. But if one tries out several different actions at once, how is one to tell which action helped and which did not? And how is one to distinguish the effect of the new action from the stochasticity of the environment (e.g. whether this specific book was good)? The solution is to use a value function. In figure 3.4, the value of each state is shown.

[6] In figure 3.1, there were two types of nodes: states and actions. Here (except in one place), no separate nodes are used for actions, because all (but one) actions have a deterministic outcome.

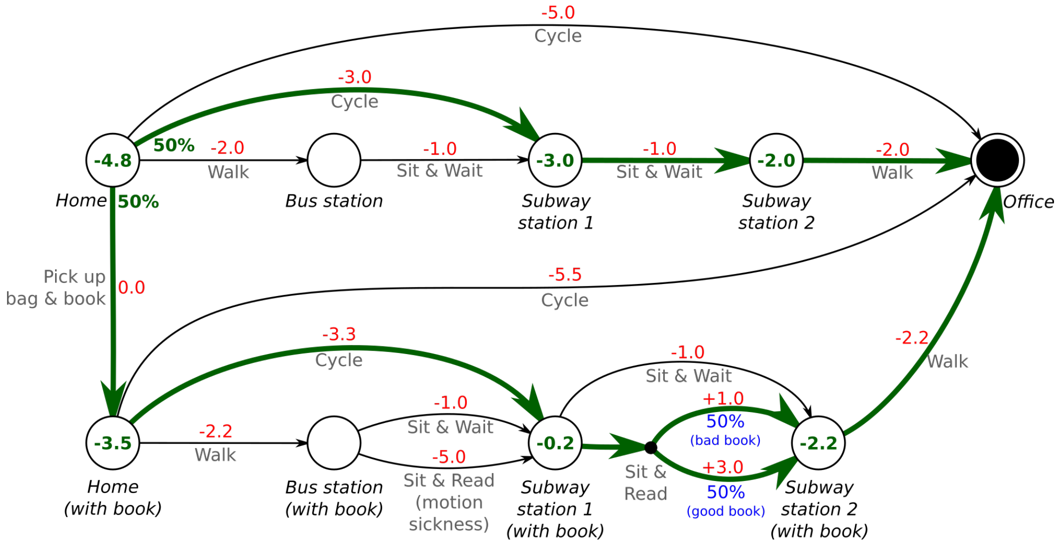


Fig. 3.4 State values for the policy: randomly pick or do not pick a book, cycle to the subway station, then go to work via the subway. In this example values are not given for states not visited by the policy.

Once the value for states are known (using one of the techniques presented in section 3.1.3), it becomes easy to assign credit to actions. For instance, in figure 3.4, when the agent goes from "Home" to "Home (with book)", she receives a reward of 0, while switching from a state with value -4.8 to one with value -3.5, thus making a total gain of 1.3 compared to expectations[7]. The surprise is positive; the agent did better than usual; better than expectations. This means that (assuming the policy is otherwise fixed) increasing the probability of picking up a bag and a book will probably improve the average commute.

The value function thus enables an agent to know, immediately after an action, whether it has improved the situation of the agent: the agent does not need to wait until it actually experiences the reward. Furthermore, the value function "smoothes" over stochastic effects. In the commute MDP, the agent does not know in advance whether the particular book selected is good or not; so at the state "Home (with book)", it is not possible to predict the *actual* return. Using the *expected* return instead of the actual return reduces the variance of the updates, and thereby allows for faster learning (there is no need to test this change over a large number of different books if we already know the average value of having a book).

[7] This "gain" $\delta = v(s') + r - v(s) = -3.5 + 0 - (-4.8) = 1.3$ (which can also be a loss when taking a worse action) is called the time-difference error, and will be discussed later in this chapter.

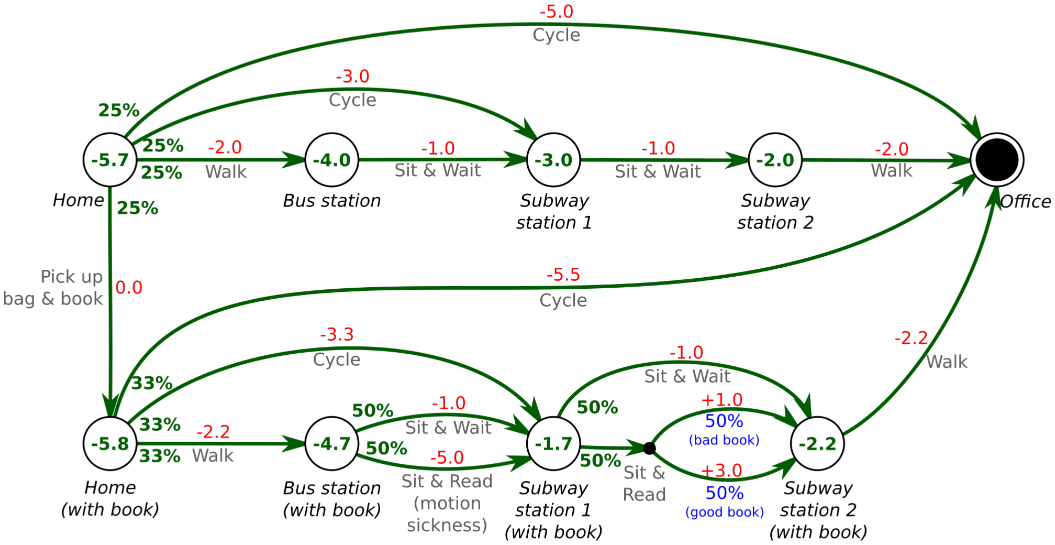


Fig. 3.5 State values for the policy: pick every action uniformly at random.

So long as an agent explores enough (i.e. uses a policy that continues to visit all state transitions), the agent eventually finds the best policy for a finite MDP. It does so by estimating the value according to policy evaluation techniques (presented in subsection 3.1.3) and then increasing the probability of using "good" actions according to policy improvement techniques (presented in subsection 3.1.4). In the commute case, the optimal policy is shown in figure 3.6 along with the optimal value function.

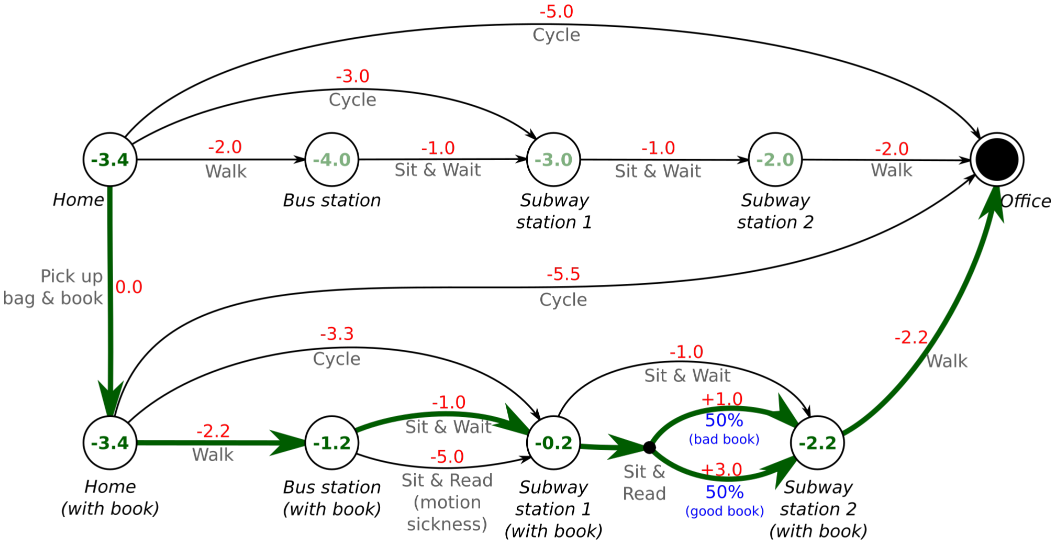


Fig. 3.6 Optimal policy and value function.

The rest of this section presents the basic techniques for estimating the value function (subsection 3.1.3) and improving the policy (subsection 3.1.4). These techniques constitute the basis for all the algorithms used in this thesis. The following section (3.2) introduces more advanced techniques which, relying on the same basic principles, are capable for example of tackling large state spaces.

3.1.3 Policy evaluation

When considering an MDP and a given policy π, one may construct a Markov Reward Process corresponding to the evolution of the system as the agent follows π. This makes it possible to estimate, for every state, the expected return obtained by an agent following the policy, which is the value of that state.

Note that policy evaluation is not yet solving the reinforcement learning problem - indeed, we are ultimately interested not in evaluating a policy, but in finding one that is good. Nonetheless, most reinforcement algorithms contain a policy evaluation component, and indeed much of the difficulty and diversity in RL is in efficiently evaluating a policy. How to find good policies, via policy *improvement*, is discussed in the next subsection.

Monte Carlo methods for policy evaluation

Recall that the value of a state is defined as the expected return, and the return is defined as:

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k}$$

The most obvious approach to evaluate a state is to record many independent samples of the return observed following a state visit, and to average them. This is possible if the MDP is episodic, and the result is an unbiased estimate of the return[8] (in the equation below, $\mathbb{1}_{S_t=s}$ equals 1 if $S_t = s$ and 0 otherwise):

$$v_\pi(s) = \frac{\sum_{t=0}^{\infty} \mathbb{1}_{S_t=s} G_t}{\sum_{t=0}^{\infty} \mathbb{1}_{S_t=s}}$$

This is the spirit of Monte Carlo (MC) methods for Reinforcement Learning. These methods can be used, for instance, directly from observed experience, or by sampling trajectories ("rollouts") from a model. A problem for Monte Carlo methods is that, if the environment or the policy are stochastic, a large number of trajectories must be seen in order to evaluate each state with a sufficient degree of confidence. But this is not an issue when it is

[8] It is not obvious that this should be the case when the returns are not independent, for instance when a state is visited several times during the same episode. It can be shown (Sutton and Barto, 2018) that *every-visit MC*, whose sampled returns are not independent, also converge to the expected value of the return, but in what follows we will only consider *first-visit MC*.

cheap to generate trajectories (e.g. if there is a cheap and precise model and a fast-to-compute policy)[9]. Pseudocode for MC prediction is shown in algorithm 1 (algorithm adapted from Sutton and Barto (2018)).

Algorithm 1: First Visit Monte Carlo prediction

```
Input    : the policy to be evaluated, $\pi$
Output   : an estimate V of the value function $v_\pi$
Require  : An episodic MDP, or a model thereof, which can be interacted with using
           the functions observe() (which returns the state), perform(action)
           (which accomplishes an action, thereby modifying the world), and
           newEpisode() (which initializes an episode)
// Initialize returns and value:
foreach s in States do
    $V[s] \leftarrow$ an arbitrary value $\in \mathbb{R}$
    $returns[s] \leftarrow$ an empty list
// Learning loop:
repeat
    // (1) Generate an episode following $\pi$:
    newEpisode()                                   // Initialize episode
    $S, R \leftarrow$ empty dictionaries
    $t \leftarrow 0$
    $S[t] \leftarrow$ observe()
    while $S[t]$ is not terminal do
        $a \leftarrow \pi(S[t])$
        $t \leftarrow t + 1$
        $R[t] \leftarrow$ perform(a)
        $S[t] \leftarrow$ observe()
    // (2) Parse collected episode data, in reverse chronological order:
    $G \leftarrow 0$
    while $t \geq 0$ do
        $t \leftarrow t - 1$
        $G \leftarrow \gamma G + R[t+1]$                // Accumulate rewards
        if $S[t]$ is not in $S[0:t-1]$ then        // First visit of the episode to S[t]
            Append G to returns[S[t]]
            $V[S[t]] \leftarrow$ average(returns[S[t]])   // Update estimated value
until user stoppage, time limit, or some other stopping condition...
return V
```

[9] Monte-Carlo methods were used for instance in AlphaGo Silver et al. (2017), where the environment model is simple (the rules of Go), and a special “rollout network” enabled fast computation of trajectories.

Dynamic programming for policy evaluation

Dynamic Programming (DP) methods assume a perfect model of the transitions of the MDP (Sutton and Barto, 2018, p. 73). They seek to exploit the Markov Property by treating each transition separately, in a divide-and-conquer approach. This is concisely captured by the Bellman equation (Bellman, 1957) for v_π, which relates the value of a state to the value of successor states under a policy π. To derive the Bellman equation, remark that the value of a state can be expressed as the next reward, plus the return from the subsequent state:

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \end{aligned}$$

We can then express the value of a state in terms of the transition probabilities and the return from the next state (weighted by its probability of being next):

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']]$$

But the last term ($\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']$) is by definition just the value of the next state, s'. Therefore the Bellman equation for v_π is:

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')] \tag{3.3}$$

The Bellman equation defines a system of linear equations (one for each non-terminal state) which can in principle be solved directly; in practice, iterative solutions are often preferred (Sutton and Barto, 2018, p. 74). These solutions use the Bellman equations as an update rule; they can be shown to converge to the correct value as the number of sweeps of the states goes to infinity. Pseudocode for a basic dynamic programming algorithm for policy evaluation is provided in algorithm 2.

Note that by using the Bellman equation, dynamic programming methods “bootstrap”, meaning they update an estimation (the value of s) based on another estimation (the value of s') (Sutton and Barto, 2018, p. 89). One of the advantages of bootstrapping is that one may update estimates before having observed the actual return; this opens the way for *online* methods, which learn even as they are interacting with an MDP.

Algorithm 2: Iterative Policy Evaluation

```
Input      : the policy to be evaluated, $\pi$
Require    : a model of the MDP, from which one can obtain transition probabilities
             of the form $\mathsf{p}(\textit{next state, reward} \mid \textit{state, action}) \to [0,1]$
Parameters : Small $\theta > 0$, corresponding to the maximum error tolerated
Output     : an estimate $\mathsf{V}$ of the value function $v_\pi$
// Initialize values:
foreach $\mathsf{s} \in \mathsf{States}$ do
    if $\mathsf{s}$ is terminal then $\mathsf{V[s]} \leftarrow 0$ else $\mathsf{V[s]} \leftarrow$ an arbitrary value $\in \mathbb{R}$
// Evaluation loop:
repeat
    foreach $\mathsf{s} \in \mathsf{States}$ do                    // One sweep of state space
        $previousValue \leftarrow \mathsf{V(s)}$
        // Propagate value ‘‘backwards’’, using the Bellman equation as an
           update rule:
        $\mathsf{V(s)} \leftarrow \sum_a \pi(a|\mathsf{s}) \sum_{\mathsf{s}',r} \mathsf{p}(\mathsf{s}', r|\mathsf{s}, a)\,[r + \gamma \mathsf{V}(\mathsf{s}')]$
        $\Delta \leftarrow \max(\Delta, |previousValue - \mathsf{V(s)}|)$
until $\Delta < \theta$
return $\mathsf{V}$
```

Time difference methods for policy evaluation

Monte Carlo methods do not bootstrap, and Dynamic Programming methods require a pre-existing model. In that sense, neither is a paradigmatic RL method, in contrast with temporal difference (TD) learning. Temporal-difference methods use the divide-and-conquer idea of DP, but, like Monte Carlo methods, they can learn directly from experience, absent any model. Furthermore, they can learn online, making them applicable to life-long learning scenarios such as those faced by animals and humans.

The basic idea is to bootstrap, as in DP (cf. algorithm 2), but sampling from direct experience as in Monte Carlo methods. First, note that the value can be rewritten in terms of the expectation of the value at the next state, as follows:

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}|S_t = s)] \end{aligned} \tag{3.4}$$

Informally, a situation (state) is exactly as promising (valuable) as whatever pleasure/satisfaction (positive reward) or pain/dissatisfaction (negative reward) the agent is currently experiencing, plus however promising the next situation is likely to be.

When we replace v_π in 3.4 by an approximation V obtained from experience, and the random variables R_{t+1}, S_t and S_{t+1} with sampled values r_{t+1}, s_t, and s_{t+1}, the equality no longer holds. The difference between the two sides of the equation is the expected temporal difference error. When the temporal difference error is 0 in expectation, we have the correct value function. The expected temporal difference error is denoted $\mathbb{E}\delta_t$, where the temporal difference error δ_t is:

$$\delta_t \doteq r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{3.5}$$

However, in contrast to DP, we are sampling from experience, and thus we may be affected by stochasticity in the policy and/or in the transitions (the equality in 3.4 is only true in expectation). To deal with this, a learning rate α is introduced. Therefore, when we turn 3.4 into an update rule, it becomes:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{3.6}$$

$$V(S_t) \leftarrow V(S_t) + \alpha\delta_t \tag{3.7}$$

Using this update rule, TD(0) converges to values that are correct for the maximum-likelihood model of the MRP. Pseudocode for TD(0) is given in algorithm 3.

Algorithm 3: Tabular TD(0) for estimating v_π

```
Input      : the policy to be evaluated, $\pi$
Require    : Interaction with a (potentially continuing) MDP using functions
             observe(), perform(action)
Parameters : A learning rate $\alpha$
Output     : an estimate V of the value function $v_\pi$
// Initialization:
foreach s ∈ States do
    if s is terminal then V[s] ← 0 else V[s] ← an arbitrary value ∈ $\mathbb{R}$
s ← observe()
// Learning:
repeat
    a ← $\pi$(s)
    r ← perform(a)
    s' ← observe()
    V(s) ← V(s) + $\alpha$[r + $\gamma$V(s') − V(s)]
    s ← s'
until user stoppage, time limit, or some other stopping condition...
return V
```

TD(0) is perhaps the most important reinforcement learning algorithm: it can learn directly from experience, without requiring a model of its environment, even in the continuing (non-episodic) case. (It is therefore essential for the rest of this thesis to understand this algorithm.) This makes it suitable as a model for animal learning, for instance. However, like the other algorithms presented so far, it is merely an evaluation algorithm. In the next subsection I discuss the extension from evaluation to control; but first some potential improvements to policy evaluation are presented.

Beyond TD(0)

There are many ways to improve the performance of temporal difference algorithms. Often these methods can be used jointly to combine their benefits, though depending on the case they may also interact in counter-productive manners (for instance, function approximation and off-policy learning can be difficult to combine; see e.g. Sutton and Barto (2018), chap. 11). I list here only some of the techniques that are considered important:

- **N-step methods:** for instance, the TD(λ) algorithm (of which TD(0) is a degenerate special case) uses *eligibility traces* to compute temporal differences across multiple time-steps, thus forming a bridge between Monte Carlo and single time-step temporal difference methods.

- **Efficient use of models:** many methods can make use of models in creative ways, sometimes in combination with direct experience, thus avoiding complete sweeps of state-space. (This can include approximate models built online from experience, e.g. in Sutton (1990).) For instance, *prioritized sweeping* methods update preferentially along transitions for which there is a large temporal difference error.

- **Off-policy learning:** the two basic learning algorithms presented here (First visit Monte Carlo and TD(0)) are *on-policy* algorithms: they evaluate the policy that they are following. It is also possible to learn *off-policy*, that is, to evaluate one policy π while gathering experience using a different policy π', called the behavior policy. This is done by correcting for the difference between the distribution of transitions observed by π' compared to π; the technique is called *importance sampling*.[10]

- **Function approximation:** instead of using point-like states and actions and representing the value function and policy with tables, various techniques can be used

[10] The famous Q-learning algorithm is an example of this, though it can be seen as a degenerate case of importance sampling due to the use of a deterministic estimation of the best policy $\hat{\pi}$.

to generalize across states and state-action pairs. I discuss this briefly in a separate subsection of this chapter.

3.1.4 Policy improvement and control

In the previous subsection, I have discussed three basic methods (Monte Carlo, Dynamic Programming, and Temporal Difference) for evaluating the expected discounted return at every state, or value $v(s)$, under a policy π. But in Reinforcement Learning the objective is not merely to evaluate an MRP, but to discover a good (or even optimal) policy for the MDP. The purpose of the policy *evaluation* techniques described above is to serve in policy *improvement*.

Policy improvement theorem

Action-values are given by the action-value function denoted $q_\pi(s,a)$:

$$q_\pi(s,a) \doteq \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s, A_t = a] \tag{3.8}$$

$$= \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')] \tag{3.9}$$

Action-values are conceptually similar to state values, except they evaluate not just states, but states accompanied by the next action that will be performed.

To improve the policy, it is enough to increase the probability of selecting actions for which $q_\pi(s,a) > v_\pi(s)$; this can be done simultaneously for many states. This general result is called the policy improvement theorem (Bellman, 1957); that is, if $q_\pi(s,\pi'(s)) \geq v_\pi(s)$ for all $s \in \mathcal{S}$, then $v_{\pi'}(s) \geq v_\pi(s)$ for all $s \in \mathcal{S}$.

Armed with this theorem and a model, it is straightforward to improve a policy in a Monte Carlo, Dynamic Programming, or Temporal Difference setting. This procedure is called Policy Iteration, and is presented below. Policy Iteration converges to the optimal policy for a finite MDP (Sutton and Barto, 2018).

Generalized Policy Iteration

In principle, then, it would be sensible to start with a policy, evaluate it with one of the algorithms discussed in the previous section, improve it based on the policy improvement theorem, and repeat these steps until convergence. However, policy evaluation is in many cases an expensive operation; including a full policy evaluation in the policy iteration loop

Algorithm 4: Policy Iteration

```
Require    : a model of the MDP, from which one can obtain transition probabilities
             of the form $p(\textit{next state, reward} \mid \textit{state, action}) \rightarrow [0,1]$
Parameters : Small $\theta > 0$, corresponding to the maximum error tolerated
Output     : an estimate V of the value function $v_\pi$
// Initialization:
$\pi \leftarrow$ an arbitrary deterministic policy
// Policy iteration loop:
repeat
    $V \leftarrow$ iterative policy evaluation of $\pi$, initializing state values to V
    policyStable $\leftarrow$ true
    foreach $s \in \mathcal{S}$ do
        oldAction $\leftarrow \pi s$
        // Policy improvement step, using a model:
        $\pi(s) \leftarrow \arg\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S},r\in\mathcal{R}} p(s',r|s,a)[r+\gamma V(s')]$
        if oldAction $\neq \pi(s)$ then policyStable $\leftarrow$ false
until policyStable
return $\pi$
```

is unwieldy (Sutton and Barto, 2018, p. 82). Alternating policy evaluation steps and policy improvement steps would result in slow learning.

But as it turns out, it is usually possible to interrupt the policy evaluation procedure, and to improve the algorithm based on a partially evaluated policy. The two processes, policy evaluation and policy improvement, can then be interleaved, with the value function tracking the moving target of the improving policy. This idea is called Generalized Policy Iteration. Figure 3.7 illustrates the idea of policy iteration and generalized policy iteration. Most Reinforcement Learning algorithms are some flavor of Generalized Policy Iteration.

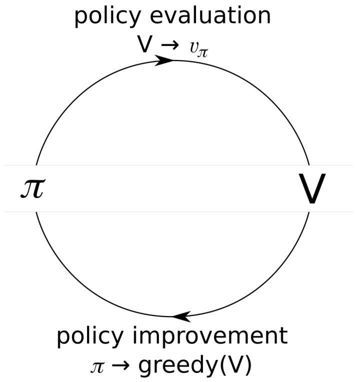


Fig. 3.7 In the special case of *policy iteration*, the policy is evaluated until convergence; then the value function is used to improve the policy until it is greedy with respect to the value function. In *generalized policy iteration*, the two processes may be interleaved, such that a policy evaluation process tracks the value of a constantly changing policy, whereas the policy moves towards the greedy policy (the policy maximizing expected return) according to the current value estimate. Figure adapted from Sutton and Barto (2018).

Critic-only methods

Several generalized policy iteration algorithms can be shown to converge to the optimal policy, π^*, at least in the tabular and linear function approximation cases. In particular Q-Learning (Watkins and Dayan, 1992) and Sarsa (Rummery and Niranjan, 1994) both converge (Singh et al., 2000; Watkins and Dayan, 1992). Both of these popular algorithms are "critic-only" algorithms: meaning the same structure is used as a basis for the policy (or actor) and the value function (or critic).

Sarsa can be viewed as updating both its action-value function and its policy with every learning step, thus forming a particular instance of GPI. Sarsa converges with probability 1 to the optimal action-state value function Q^*, if all state-action pairs are visited an infinite number of times (in the continuing example presented above, this depends on each state being accessible from every other state, so that the agent never gets indefinitely stuck), and the policy itself converges to the greedy policy with respect to the value function, i.e. the policy which chooses the actions with highest expected return according to this value function. One way to achieve this is to reduce the value ε progressively with time; see Singh et al. (2000) for details.

Another popular critic-only algorithm is Q-learning, which resembles Sarsa, except for the update rule:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a \in \mathcal{A}} Q(s',a') - Q(s,a)] \tag{3.10}$$

```
Algorithm 5: Sarsa
Input      : the policy to be evaluated, $\pi$
Require    : Interaction with a (potentially continuing) MDP using functions
             observe(), perform(action)
Parameters : Step size $\alpha \in (0,1]$, small $\varepsilon \in (0,1)$
Output     : in the limit, the optimal policy $\pi*$
// Action-selection function based on action values (variants are possible).
   See sub-subsection “exploration” for a discussion of the role and
   importance of this part of the algorithm.
Function $\varepsilon$-greedy(s):
    with probability $1-\varepsilon$ do
        return $\arg\max_{a\in\mathcal{A}}$ Q(s, $a$)
    otherwise
        return a random action
// Initialization:
foreach s $\in$ States do
    if s is terminal then Q(s, a) $\leftarrow$ 0 else Q(s, a) $\leftarrow$ arbitrarily $\in \mathbb{R}$
s $\leftarrow$ observe()
a $\leftarrow$ $\varepsilon$-greedy(s)
// Learning:
repeat
    r $\leftarrow$ perform(a)
    s$'$ $\leftarrow$ observe()
    a$'$ $\leftarrow$ $\varepsilon$-greedy(s’)
    Q(s, a) $\leftarrow$ Q(s, a) $+ \alpha[$r $+ \gamma$Q(s’, a’) $-$ Q(s, a)$]$
    s $\leftarrow$ s$'$
    a $\leftarrow$ a$'$
until user stoppage, time limit, or some other stopping condition...
return Q
```

Plugging this update rule in algorithm 5, in place of the Sarsa update rule, gives a version of the Q-learning algorithm. Despite this superficial similarity, Q-learning is quite different conceptually: it is an *off-policy* algorithm, meaning it learns directly about the optimal policy, and can do so using *any* exploratory policy that visits all state-action pairs. In practice, an ε-soft version of the current approximation to the optimal policy works well, as in algorithm 5. Q-learning also converges with probability 1 to the action-state value function Q^* (Watkins and Dayan, 1992).

3.2 Advanced Reinforcement Learning techniques

This chapter has focused so far on the basic insights, abstractions, and techniques of reinforcement learning: the MDP framework, policy evaluation (MC, DP, TD), policy improvement; and elementary applications of these techniques (Sarsa).

The methods do not form separate paradigms. On the contrary, they can often be combined. For instance, one may combine model-free methods (Monte Carlo or Time Difference) with model-based methods (corresponding to Dynamic Programming) via the Dyna-Q architecture (Sutton, 1990), which learns from both the model and direct experience, even as the model is being built.

The techniques described in the previous section constitute the basic theory and intuition of reinforcement learning. However they are still limited in ways that make them inapplicable to many real problems. These limitations are:

- Scaling with the state-space: tabular reinforcement learning methods do not scale to large problems; and problem size grows surprisingly fast due to the curse of dimensionality (Bellman, 1957). One solution to this is function approximation.
- Scaling with the action-space: the critic-only methods presented so far assume that a small number of actions are available from each state, such that maximization is easy. In many problems this is not the case (for instance the human body has hundreds of muscles, resulting in a very large action space). One solution to this is actor-critic algorithms (Sutton and Barto, 2018, chap. 13).
- Efficient, deep exploration in large state and action spaces. This thesis can be viewed as an attempt at tackling this problem; this section reviews some existing methods.

In this section I focus on introducing state-of-the-art techniques for solving the above-mentioned limitations. Namely: function approximation, in which reinforcement learning is combined with other machine learning techniques such as deep neural networks; actor-critic methods, which represent separately the policy and the value function; and various techniques for tackling exploration, including Hierarchical Reinforcement learning (see Sutton et al., 1999).

3.2.1 Function approximation

Recall the commute MDP (figure 3.3). In the absence of the book, the MDP would have had only 5 states and 6 transitions; however, introducing the book just about doubled the size of the MDP, resulting in a total of 9 states and 16 transitions. A more realistic version of the MDP would have to take into account even more factors - e.g. the day of the week, strikes, weather, pollution levels, our commuter's mood, and so on and so forth. Each of these decisions would increase the size of the MDP, and together they result in a combinatorial explosion – the "curse of dimensionality" – such that it would soon become impossible to model this problem as a graph, even using a supercomputer, let alone repeatedly explore and evaluate every state and action. To circumnavigate the issue, one must abandon the idea of independent "atomic" states, instead encoding each state as a combination of factors, and generalizing across states. This is the role of function approximation.

Many of the most celebrated achievements of Reinforcement Learning were obtained in combination with function approximation techniques, often deep neural networks. These achievements include TD-Gammon (Tesauro, 1993), which learned to play backgammon at grandmaster level, Deepmind's Atari playing algorithm (Mnih et al., 2013), in which a single algorithm, using the same metaparameters, learned to play several Atari games at human levels of performance; and in 2016, still by Deepmind, AlphaGo (Silver et al., 2016), which overcame one of the world's best Go player in a highly publicized match featuring several "creative" moves from AlphaGo's[11].

Function approximation has been a key element in these advances. In the discussion so far, both the value and the policy assumed discrete states, between which no generalization was made. For instance, a value function could be represented by a table with one entry per state, an action-value function by a table with one entry per state-action pair. This facilitates both intuitive understanding and formal proofs. However, there are (at least) three reasons to be dissatisfied with that approach:

1. **Storage:** for most useful problems, state spaces quickly become prohibitively large: for instance, a 10×10 pixels shades of grey image has 255^{100} distinct states. This makes it impossible to store the value function or the policy. In contrast, with function approximation a comparatively small number of parameters can be stored instead, provided some compression is possible.

2. **Data efficiency:** if one must experience every state several times, then considerable experience is required to learn a useful policy. The cost is very large when a model is available; it is prohibitive when the agent must directly experience negative rewards in trial-and-error exploration. With function approximation, each sampled transition serves to learn about all related situations that the agent might experience, thus increasing data efficiency.

[11] Although the most creative move in that encounter was, perhaps, the one move that secured the lone human win in the 4-1 result - move 78 of game 4.

3. **Behavior in novel situations:** finally, discrete states prevent generalization to novel states. With function approximation, the agent may be able to behave well even in situations that were never previously encountered, if some generalization is possible from past to prevent.

Function approximation is usually synonymous with *generalizing from examples* or *supervised learning* in the reinforcement learning context. Reinforcement learning algorithms make use of the existing methods from supervised learning, integrated into RL algorithms. These include interpolation techniques, regression techniques; Bayesian methods; etc. The most dramatic successes (Mnih et al., 2013; Silver et al., 2016; Tesauro, 1993) have been accomplished using non-linear function approximation such as deep neural networks.

However, the combination of reinforcement learning with supervised learning techniques is not trivial (Sutton and Barto, 2018, chap. 9 and 10). RL with (nonlinear) function approximation is often used despite a lack of theoretical guarantees of convergence, because of its proven practical efficiency on a number of problems. Among the issues that arise, the following two are particularly noteworthy:

1. RL methods such as temporal difference learning, and approximation methods such as regression can interfere with one another; sometimes this leads to RL algorithms converging to erroneous values, or even diverging, especially in the off-policy case (Sutton and Barto, 2018, chap. 11).

2. The data sampled in RL often doesn't meet the requirements of supervised learning algorithms. Many supervised learning algorithms require *independent identically distributed* (iid) data - but because RL agents (1) change their behavior as they learn and (2) the order of experienced data is imposed by the MDP's dynamics, the data is not iid.

To mitigate these issues, techniques have been developed which improve the compatibility between RL methods and deep neural networks. Because these techniques are used in the experimental section of this thesis, they are briefly introduced here:

- **Experience replay:** this technique (Lin, 1992; Mnih et al., 2015) consists in recording observed transitions or "sarses" (for State-Action-Reward-next State), such that they can be re-used for learning. One of the advantages of experience replay is improved sample efficiency (since each experienced transition is re-used multiple times); another is a decorrelation between successive updates (since many are sampled from memory rather than from the flow of experience).

- **Target networks:** since successive states or action-states often have similar features (they are correlated), function approximation can lead to a “moving target effect” on temporal difference algorithms, where changing the value of $v(s)$ also changes the value of $v(s')$. Such moving targets can lead to divergence. It helps to use a target network (especially in the off-policy case, due to other contributing factors (Hasselt, 2010; Wang et al., 2015)). In double-DQN, two value networks are maintained, one serving as a target for the other. The target network is regularly updated.
- **Parallel agents:** one way of decorrelating data is to save it and shuffle it as in experience replay. Another approach is to simply run multiple agents in parallel and to update using a corresponding batch (Mnih et al., 2016). Such agents can also better exploit parallel processing capabilities.

In summary: even though the RL paradigm is best understood by considering the simple cases involving discrete states and actions, the field’s main focus, and its greatest successes, occur when RL is combined with function approximation. The most popular and successful function approximation technique to date has been deep neural networks. Function approximation introduces new challenges, which the field has tackled using a range of techniques such as those presented above: experience replay, target networks, parallelism, etc.

3.2.2 Actor-Critic methods

In place of approximating the value function, or in addition to it, one may approximate the policy. One may learn directly in policy space, using a parameterized policy function. This is for example the principle of the REINFORCE algorithm (Williams, 1988, 1992), which updates policy parameters following the gradient of performance.

However, REINFORCE learns slowly due to the high variance of its reinforcement signal, owing to the inefficient manner in which it solves the credit assignment problem (Sutton et al., 2000). To improve the efficiency of REINFORCE, it is therefore useful to combine its parameterized policy or *actor* (which selects the actions or “acts”) with a parameterized estimate of the value function, or *critic* (which evaluates the behavior or “criticizes”). This family of methods is called “actor-critic” (Sutton and Barto, 2018, chap. 13).

The temporal difference methods presented so far are “critic-only”: they directly represent a value function (more specifically the action-value function), from which they deduce in real time the policy: for instance by picking the action with the highest value for the current state. Actor-critic methods form another class of temporal difference reinforcement learning algorithms. In contrast with critic-only methods, they can tackle problems with large or continuous action spaces, for which it can be difficult or computationally expensive to find the best-valued action from a given state. Furthermore, a critic and an actor may be able to discover and exploit different kinds of regularities (Sutton and Barto, 2018, p. 323). Below, I present the method in more detail.

Actor-critic architectures are an obvious application of the Generalized Policy Iteration interpretation of Reinforcement Learning presented in this chapter. Whereas GPI presents policy evaluation and policy improvement as two distinct (though interleaved) processes, the critic-only methods we have seen so far (Q-learning, Sarsa...) "collapse" both processes into the state-action value function. In contrast, in actor-critic, two separate structures are maintained that correspond to the policy and the value. Thus the critic learns by truncated policy evaluation, and the actor learns by truncated policy improvement. Under certain conditions (e.g. that the critic be updated at a faster rate than the actor), actor-critic algorithms can be shown to converge (Konda and Tsitsiklis, 2000) (however, unlike tabular Q-learning, not necessarily to the global optimum).

Typically in actor-critic algorithms, the learning of the actor (policy improvement) follows the gradient of performance (Sutton and Barto, 2018, p. 337). To explain what this is and how it is done, we need to introduce a few additional concepts, beginning with the policy gradient theorem. Firstly, consider that the policy as implemented by the actor relies on a vector of parameters θ. Secondly, we define the performance of the policy as $v_{\pi_\theta}(s_0)$: that is, the performance is equal to the value of the starting state in the episodic case[12]. Thus to modify the policy towards improved performance, one should use the following update rule:

$$\theta \leftarrow \theta + \alpha \nabla_\theta v_\pi(s_0) \tag{3.11}$$

Thereafter we refer to the performance as $J(\theta) \doteq v_\pi(s_0)$. If we knew the gradient of performance in the policy parameters $\nabla_\theta J(\theta)$, we would know in which direction to modify the parameters to ascend the gradient of performance. The policy gradient theorem (Konda and Tsitsiklis, 2000; Sutton et al., 2000) indicates that:

$$\nabla J(\theta) = \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s) \tag{3.12}$$

$$\nabla J(\theta) = \sum_s \mu(s) \sum_a \pi(a|s) q_\pi(s,a) \nabla \log \pi(a|s) \tag{3.13}$$

[12] In the continuing case, the policy gradient theorem remains true but some modifications are required. For instance, the performance of the policy cannot be defined based on a start state, but must instead be defined based on the rate of reward that is achieved over the policy's stationary distribution over the state-space. To keep this chapter simple I only discuss the episodic case here; but the theorem remains true in the continuing case; see Sutton et al. (2000) for a proof.

Where $\mu(s)$ denotes the probability of encountering s (μ is the distribution over states). There are several things to note about this result:

- The distribution $\mu(s)$ is outside the gradient equation. This means one can get an unbiased estimate of the gradient by sampling states on-policy, applying updates of the form $\sum_a q_\pi(s,a)\nabla\pi(a|s)$. Furthermore, note that by using the log trick in equation 3.13, we have extracted the probability of selecting an action from the gradient. When learning on-policy, we can then use updates of the form: $\theta \leftarrow \theta + \alpha q_\pi(s,a)\nabla\log\pi(a|s)$.

- Perhaps surprisingly, $q_\pi(s,a)$ can be replaced by $q_\pi(s,a)+b(s)$, for any b (Sutton and Barto, 2018, p. 329). This is called using a baseline; the technique can allow for drastically reducing the variance of the updates. (Intuitively: this works because all the actions for a given state are affected equally by this baseline, and thus the preference for each action comparatively to the others remains the same.) A common baseline is the value $v_\pi(s)$. Note that the advantage function[13] $adv(a|s) = q_\pi(s,a) - v_\pi(s)$ is equal in expectation to the temporal difference error observed when performing action a in state s. Therefore one may update the policy directly by using the temporal difference error. This reduction in variance is an improvement of credit assignment.

Thus for a one-step actor-critic algorithm, one may update the critic in the ways that we have already seen (e.g. TD(0)), while the following update can be used for the actor:

$$\theta \leftarrow \theta + \alpha\delta_t\nabla\log\pi(A_t|S_t,\theta_t) \tag{3.14}$$

[13] So called because it gives the advantage of selecting each action, relative to the best action for that state.

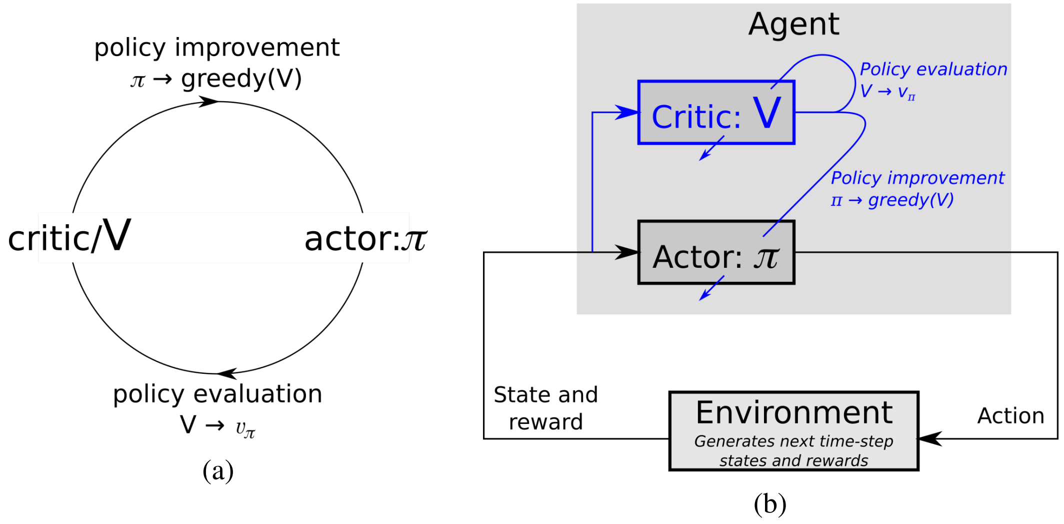


Fig. 3.8 The actor-critic architecture.
(a) Actor-critic architectures use separate dedicated structures to represent each element of generalized policy iteration: the value and the policy.
(b) An actor-critic agent interacting with its environment. The elements dedicated to learning are shown in blue: the critic serves as a basis for the learning signals of both policy evaluation (comparing expected with observed rewards) and policy improvement (increasing the probability of greedy actions with respect to the value function).

Recapitulating: this section has presented actor-critic algorithms, from a rapid overview of theoretical foundations, to parameter update equations. This family of algorithms, although seemingly more complex than critic-only algorithms, is derived from many of the same theoretical principles. In some important cases, actor-critic algorithms are better applicable than their critic-only counterparts; in particular, when the action-space is large, or when a stochastic policy must be learned.

3.2.3 Exploration

In the treatment of reinforcement learning in this chapter so far, very little attention has been paid to the question of exploration. Instead, we have discussed agents which gradually learn to act more efficiently, but which do not go “out of their way” to acquire experience that could potentially lead to policy improvements. Exploration is of course a central aspect of creative endeavors, including problem-solving and, perhaps, insight.

To gather rewards, an RL agent must use the actions that have proven useful in the past; but in order to improve its policy, it must try out new actions and observe their consequences (Barto et al., 2013; Sutton and Barto, 1998). Thus, rather than greedily using the estimated best policy $\hat{\pi}_t^*$, the agent must miss out on predictable rewards in exchange for observing something new. When and how to explore is one of the main open problems in RL. Indeed, convergence to an optimal policy in Sarsa and Q-learning, for instance, depends on continued exploration of the entire problem space. In algorithm 5, exploration was achieved using an ε-greedy action-selection procedure, thus randomizing action-selection. In fact, a range of techniques is available.

In critic-only algorithms, dithering techniques introduce some noise into action-selection at each timestep. Popular approaches are ε-greedy and SoftMax action selection (Sutton and Barto, 2018, pp. 28, 37). An ε-greedy agent ($0 < \varepsilon < 1$) picks actions according to $\hat{\pi}^*$ with probability $1 - \varepsilon$, and otherwise picks another action at random. In actor-critic algorithms, one may rely on the policy's inherent stochasticity; however this can lead to premature convergence. In practice, it has been found that introducing an entropy-based regularizer (punishing low-entropy action-selection given the state) helps maintain sufficient levels of exploration (Mnih et al., 2016).

Since exploration occurs on a fraction of time-steps, these algorithms tend to focus exploration around promising trajectories (Thrun, 1992). This restricts the state space in a manner that can be beneficial (ignoring seemingly irrelevant regions and focusing on promising ones), but in the absence of other mechanisms it can also be detrimental (intensifying exploration around an already well-known policy). Several methods have been proposed to diversify exploration.

One such approach is to encourage exhaustive or near-exhaustive exploration of the state space, typically by keeping count of state visits, as in the popular R-MAX algorithm (Brafman and Tennenholtz, 2003). The sample complexity of these techniques grows linearly with the size of the state space (Thrun, 1992), but by relaxing the exhaustiveness of this approach (ceasing to explore when "good enough" behavior has been discovered), it can produce good results even in robotics domains (Hester et al., 2010). Nonetheless, it seems insufficient to deal with creative domains characterized by very large or even continuous and therefore infinite problem spaces; in such domains, systematic exploration – even in simulation – could last an agent's lifetime without discovering anything "good enough". Further, exhaustiveness seems antinomic with creativity: surely an efficient search would explore mostly the most promising states.

Exploration does not have to be random or exhaustive – it can be guided by the expectation of learning. This has been done in two ways, which we will label intrinsic motivation (Barto, 2013; Singh et al., 2010) and artificial curiosity (Oudeyer and Kaplan, 2008). The latter can be understood as a special case of the former[14].

[14] According to our taxonomy; in the artificial intelligence literature, "intrinsic motivation" and "artificial curiosity" are often used interchangeably.

The first of these two ways, intrinsic motivation (Singh et al., 2010), consists in modifying the reward function to improve the performance of an agent. Whereas the traditional approach to RL is to provide reward exclusively when the goal is achieved, intrinsically motivated agents also receive “shaping” rewards whenever they encounter an interesting state (or “salient state"). These shaping rewards can be provided by the programmer, or, for example, learned via an evolutionary algorithm over several generations of agents. Indeed, these techniques are inspired by the drives of biological agents: from an evolutionary perspective, the “purpose” or fitness function of living agents is to reproduce; but they also receive pleasure from activities such as eating food or playing, which are, ultimately, useful steps towards securing reproductive success. Intrinsic motivation emulates this “design”.

Artificial curiosity (Oudeyer and Kaplan, 2008) is based on rewarding a special kind of activities: those conducing to learning. An artificial agent is “curious” when it uses information-theoretic measures to detect and predict learning or surprise, and receives reward when that occurs. Implementations have used various such measures. Examples include rewarding state prediction error (surprising events) (Huang and Weng, 2002; Schmidhuber, 1991), prediction improvement (Oudeyer et al., 2007), or competence improvement (Santucci et al., 2012). These systems typically include a learning mechanism to compute the amount of surprise or learning that the agent undergoes, whereas intrinsic motivation can often rely on a static reward function. Some of that work investigates optimal exploration from a Bayesian perspective (Sun et al., 2011), although that remains unfeasible for the general RL problem (Jong, 2010, pp. 25-28).

Below, we consider a fourth kind of method to improve exploration, which consists in using the structure of the problem-space to increase the efficiency of exploration. A common form of this is what one may label “state generalization”, or (with a slight departure from common usage) “state abstraction”, which finds commonalities between states and uses these commonalities for action selection, both exploitative and explorative. A deep reinforcement learning agent with SoftMax action-selection can therefore be said to use the similarities between states in order to guide its exploration: state-actions that resemble “good” past state-actions will be selected preferentially. “Function approximation” is therefore one of the most widespread, but least often acknowledged, exploration techniques for reinforcement learning: it is a method that allows an agent to “learn to explore”, that is, to approximately infer promising actions based on experienced ones.

A limitation of these methods, however, is that they continue to treat an action separately from those that occur before and after it. That is, $\pi_\theta(A_t|S_t,S_{t-1},A_{t-1}) = \pi_\theta(A_t|S_t)$. If the system is Markov, then it remains theoretically possible to select actions optimally with this constraint. However, just because a system is Markov does not mean that there is no (exploitable) temporal structure; that is, it might be easier to solve the problem by removing the constraint of having to act independently at each time-step. This point is discussed more extensively in chapter 7.

In order to discover and exploit temporal structure, there exists a range of Hierarchical Reinforcement Learning (HRL) methods. These methods are often fruitfully combined with curiosity or intrinsic motivation; as a result, their contribution to exploration often goes unnoticed (Barto et al., 2013). They are discussed below.

3.3 Hierarchical Reinforcement Learning

3.3.1 Overview of HRL methods

The dominant theoretical paradigm for HRL is the Options framework of Sutton et al. (1999). This work makes use of the notion of semi-MDP or SMDP (Bradtke and Duff, 1995; Parr and Russell, 1998) (an idea also useful for continuous-time RL), in which standard *actions* are a special case of *options*, whose duration can vary. An option consists of three components: a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$, a termination condition $\beta : \mathcal{S}^+ \to [0,1]$, and an initiation set $\mathcal{I} \in \mathcal{S}$. The policy is the manner in which the agent act after the option has been initiated and before it has terminated, the termination condition is what causes the option to end and thereby "return control" to a policy-over-options, and the initiation set consists of the states from which the option may begin (this may be, and often is in practice, the whole state set $\mathcal{S}$). One may then treat the problem of discovering the policy over options as solving an SMDP, which is remarkably similar to solving an MDP.

A range of Hierarchical Reinforcement Learning algorithms have been developed, usually on the basis of the Options framework. The simplest consist largely of providing hand-crafted options, option heuristics, or even just macro-actions (macro-actions are a special case of options which simply execute a predefined sequence of sub-actions, without reacting to intermediary states); see e.g. Lakshminarayanan et al. (2017) for a recent use of repeated actions, Stone and McAllester (2001) for a successful use of hand-crafted options for robot soccer, and McGovern and Sutton (1998) for an investigation of the advantages of macro-actions for exploration and value propagation.

But a more challenging objective is the automated discovery, or learning, of good options. The challenge of discovering options is also the challenge of discovering the relevant actions (atomic actions being a special case of options). In most non-hierarchical applications of reinforcement learning, the correct actions are selected by the programmer on intuitive grounds; a poor choice can increase learning time or reduce performance (McGovern and

Sutton, 1998). However, discovering good actions (that are likely to serve a purpose) is difficult.

More ambitious approaches seek to learn the options and the policy over options simultaneously. The justifications used for these methods vary greatly. One category of approaches ignores rewards in its analysis of options, instead considering only the structure of the environment (e.g. the presence of bottlenecks or other interesting states); see e.g. Machado et al. (2017); Mannor et al. (2004); Şimşek and Barto (2009). These techniques, related to the notion of *empowerment* (Salge et al., 2014), allow for learning faster in environments with sparse reward (though they risk learning purposeless skills). In contrast, other approaches make use of reward and learn options from the trajectories achieved in that manner, for instance the fully differentiable option-critic architecture (Bacon et al., 2017), or the technique of "skill-chaining" (Konidaris and Barto, 2009). Techniques using the second approach focus on discovering useful skills based on the observable reward function, but rarely cite theoretical reasons why the options thus discovered should generalize well to related problems.

Thus despite a rich and varied literature, the problem of finding options based simultaneously on reward and state transitions, and which are likely to generalize well, remains unsolved. One of the main contribution of this thesis (cf. chapter 6) is a proposal for learning general, useful skills, though as we will see it veers away from the "option" framework.

3.3.2 Options, exploration, and insight

This thesis defends a connection between temporal abstraction and insight. In the next chapter, I explore this connection experimentally; but in this subsection, the analogous properties of the insight process and certain characteristics of hierarchical reinforcement learning are discussed. First I discuss some of the properties of options in exploration, then I relate these properties to those of insight.

There are at least three potential advantages of options for exploration. The first is the possibility of using different state abstractions depending on the option (Dietterich, 2000a; Jong and Stone, 2005; Jonsson and Barto, 2001). This allows the policy for an option to select actions using only relevant features (akin to the "chunking" discussed in Knoblich et al. (2001)), independently of which features are used in the policies for other options. Depending on the currently running option, the agent takes into account different perceived features of the environment when selecting an action and when learning from an exploratory move; and a single exploratory decision could lead to a change in the representation in use. Such changing representations can be implemented for instance as separate neural networks (Bacon et al., 2017).

A second advantage is the possibility to discover and re-use general sub-policies between different tasks (Perkins and Precup, 1999) – a property much discussed in the emerging research area of transfer learning in RL (see Taylor and Stone, 2009). Because options are closed-loop, can use approximate value functions and learn on-line (depending on the learning algorithm), they can adapt to being used in a slightly different context - such that a new (but related) problem can be explored using an option learned in earlier problems; furthermore, on-line learning can allow for resolving minor differences in the new problem. Note that this resembles analogy-making as described in work on neural-symbolic integration, for instance by Besold and Kühnberger (2015), where a similar sequence of abstraction, transfer, and repair is found. Furthermore, different policies have different biases in terms of which actions to select, corresponding to what has been dubbed changes in "operator constraints" (MacGregor et al., 2001) or "heuristics" (Kaplan and Simon, 1990) in the insight context.

Insight	**HRL**
Integrated with analytic processes (Weisberg, 2015)	Based on the standard RL framework (Sutton et al., 1999)
Sudden, discontinuous progress (Metcalfe and Wiebe, 1987)	Explorative "jumps" (Vigorito and Barto, 2008)
Operator constraints changes (MacGregor et al., 2001)	Operator constraints changes (Parr and Russell, 1998)
Heuristic change (Kaplan and Simon, 1990)	Option-dependent initial value function (Sutton et al., 1999)
Representational change (Knoblich et al., 2001)	Option-dependent state abstraction (Dietterich, 2000b)

Table 3.1 Characteristics of insight and their HRL counterparts.

Third, options provide temporal abstraction, allowing for exploring the state space at multiple granularity sizes. An obvious advantage is the reduction of processing costs in a potential use of options for planning. But this also makes it possible to reach otherwise unattainable sections of the state space - especially in environments where undirected or unmotivated exploration does not allow for reaching some states (Vigorito and Barto, 2008). Many real life tasks are of this form: for most robot models, the state "broken" is absorbing and accessible from many other states, preventing both random walks and exhaustive search from reaching many interesting regions of state space via dithering. By offering choices between temporally extended strategies, HRL allows for "explorative jumps" reaching far away from its exploitative trajectory.

Stage	Observed behavior	HRL
Problem perception	Read text; manipulate material; etc. Understand what the problem is about.	Pick high-level policy
Problem-solving	Regular progress with occasional trial and error	Transfer of high-level policy, resolve errors
Impasse	Pause in activity	Encounter negative temporal difference errors, re-evaluate high-level policy
Restructuring and insight	Report change in strategy and perception; Exclaim “Aha!”	Switch policies; transfer new policy, encounter positive errors
Verification	Resume problem-solving, sometimes fail	Finish transferring new policy, sometimes fail

Table 3.2 The insight sequence and its HRL counterparts.

These advantages are integrated with the standard RL methods of learning by trial and error (and potentially in simulated trial-and-error if using a model). Thus HRL exploration is integrated and combined with the standard exploration process used in “flat” RL.

The analogies between insight problem-solving and HRL are summarized in tables 3.1 and 3.2 (both based on Colin et al. (2016)).

3.4 Conclusion

3.4.1 Summary

- Reinforcement Learning is a machine learning paradigm for learning how to act, depending on the situation, in order to maximize a scalar reward signal.
- Reinforcement Learning is formalized using Markov Decision Processes of the form $< S,A,p,\gamma >$: a set of states, a set of actions, a transition function including a reward signal, and a discount factor. The goal of the agent is to find a policy π that maximizes the *return* (the expected sum of future discounted rewards).
- Many algorithms make use of “function approximation” techniques to generalize between states or actions, for instance deep neural networks;
- Reinforcement learning techniques can be divided into:

  - Model-free and model-based, depending on whether they build a model of the environment or not;
  - Off-policy or on-policy, depending on whether the policy being optimized and the policy in use are different during learning;
  - Online and offline, depending on whether learning can take place within an episode, or only after an episode has ended;
  - Dynamic Programming, Monte Carlo, and Temporal Difference methods, depending on whether the algorithm learns based on sweeps of the state-space, roll-outs of trajectories, or from experience as it occurs;
  - Critic-only methods, actor-only methods, and actor-critic methods, depending on whether the algorithm learns a value or action-value function, a policy, or both simultaneously;

- Some key challenges (partly addressed in this thesis) in contemporary reinforcement learning research are:
  - Action/skill discovery: what makes an action or skill good for learning, and how to learn such actions/skills?
  - Temporal abstraction: how to capture and use temporal regularities? Existing methods either disregard reward altogether, or lack theoretical justification with regards to generalizability.
  - Deep exploration: how to stray far from the trodden path in order to discover radically different possibilities?

Contributions

This chapter, apart from its last subsection, is a review of existing methods offering no novel contribution. The last subsection, discussing analogies between HRL and insight, is a novel contribution. An extended version was published in Colin et al. (2016); that article also discusses HRL in the more general context of creative problem-solving (rather than just insight).

3.4.2 Bibliographical remarks

Sections 3.1 and 3.2 are heavily indebted to the second edition of Sutton and Barto's "Introduction to Reinforcement Learning" Sutton and Barto (2018). The section on Markov Processes, Markov Reward Processes and Markov Decision Processes is also inspired by the presentation in the UCL course by Silver (2015). This tutorial introduction to RL departs from such manuals mainly in its selection of topics: algorithms and methods used in the experimental sections receive particular attention, others are largely omitted.

Part II

Experiments in Deep Exploration

Chapter 4

Insight in the (artificial) pigeon

In this chapter, a study of insight in the pigeon by Epstein et al. (1984) is reproduced in a simulation. This shows how reinforcement learning policies, leveraging relevant experience, can produce behavior that is novel, useful, and seems insightful. A comparison between flat and hierarchically structured policies helps to identify some of the shortcomings of a naive approach to learning hierarchies for insight.

First, in section 4.1, I discuss the literature which inspires the experiment. I focus in particular on the role of verbal instructions in the insight literature involving humans, and then on the specifics of Epstein's experiment. Second, in sections 4.3 and 4.2 I describe the experimental set-up, highlighting commonalities with Epstein's experiment and departures from it, and the AI models used; in section 4.3.3 I summarize the experimental procedure. Finally, in sections 4.4 and 4.5 I discuss results. These results suggest a role of learning in insight, but contrary to expectations they do not validate a hierarchical reinforcement learning (HRL) model.

4.1 Motivation and background

4.1.1 Hypotheses

This chapter is intended to test the hypothesis that hierarchical reinforcement learning (HRL) can produce insight-like behavior, such that the solution to a problem is arrived at suddenly, by a change of strategy and representation, concomitant with a surprising rise of expected value. To test this hypothesis, the task used by Epstein is implemented as a virtual environment, such that it can be performed (both the training and the final test) by a simulated "pigeon" implemented using HRL techniques. Flat reinforcement learning agents are also implemented in order to provide a comparison with the HRL agents, and to test secondary hypotheses.

This main hypothesis would be verified if a single high-level decision led to solving the problem, while switching to a different neural network for processing its input (thus undergoing representational change), and producing a large temporal difference error. Due to the engineering aspects of this work, however, this hypothesis can be either confirmed (if the model succeeds at reproducing similar behavior) or be left unanswered (the model is one specific instance of a more general idea, HRL: if the model fails to produce similar behavior, the failure of one instance will not suffice to invalidate the general idea).

If this main hypothesis cannot be confirmed, the chapter can nevertheless address less ambitious hypotheses, namely:

- Whether Reinforcement Learning agents are capable of solving the insight problem used by Köhler, Birch, Epstein, and others;
- Whether RL and HRL benefit from experience in solving this insight problem, transferring solutions from old problems to new ones, like the pigeons of Epstein et al. (1984), the chimpanzees of Birch (1945), or the human participants of Wiley (1998);
- Whether expected value behaves like the "warmth ratings" used in insight research, most notably by Metcalfe and Wiebe (1987).

The rest of this section discusses the experimental work in psychology against which the performance of the artificial pigeon is compared.

4.1.2 Animal insight experiments

In chapter 2, I have reviewed the insight literature focusing on animal behavior, human behavior, and the human brain. Here, I will briefly summarize a few particularly relevant studies, with a view to designing experiments for the study of artificial insight.

Most of the contemporary insight literature (whether focusing on behavior or brain imaging) investigates human insight, using a wide array of experimental designs (for instance, consider the nine-dots problem used e.g. by MacGregor et al. (2001) or the mutilated checkerboard problem used by Kaplan and Simon (1990); see appendix A for a review of the problems used). Despite their apparent variety, insight studies in humans nearly always make use of verbal instructions which define the objective for the problem-solver (*"without lifting your pen, draw four straight lines which connect the nine dots"*). These problems thus belong to the subset of problems given by *teachers* to *students*, as opposed to the more general setting of seeking to make the most out of one's present circumstances.

In such problems, *instructions* are used to designate (a set of) target states: for the nine-dot problem, it specifies a number of lines (4), and constraints over which lines are allowed (straight lines without lifting the pen), and what a solution ought to achieve (covering every dot). Note that if the goal state was specified in a more direct non-linguistic manner, e.g. by showing the agent an example solution, discovering the solution would be easy: it would be enough to copy the proposed solution. Using language, this is avoided in a deliberate manner, using instructions which unambiguously specify a unique solution while remaining abstract enough not to reveal what this solution is. Because of this, making sense of insight problems that are set up as "exercises" or "problems" might require technology capable of human-level linguistic abilities[1], which, at the time of writing, remains a distant dream (Cambria and White, 2014).

A more general class of problems consists in overcoming a difficulty or taking advantage of an opportunity, whatever that may be, without the benefit of guiding instructions. Animal problem-solving studies use such settings. Problems consist typically in obtaining food (Birch, 1945; Epstein et al., 1984; Köhler, 1921) or freedom from some constraint (Thorndike, 1899). Even though there might be cues in the non-verbal interaction between the animal and the experimenter, there is no linguistic component: no constraints other than those defined by the animal's body and physical environment, and usually no goal other than the satisfaction of the animal's wants or needs. This makes these problems better suited to an AI approach, and particularly a reinforcement learning approach.

Among animal studies, those of Köhler (1921), Birch (1945) and Epstein et al. (1984), already discussed in chapter 2, are of particular interest. These experiments have identified with increasing clarity the role and nature of learning in solving insight problems:

1. Köhler's insightful chimpanzees had unknown prior experience (Köhler, 1921).

2. When Birch replicated Köhler's work using chimpanzees whose prior experience was fully known and did not include relevant activities (such as using sticks), he found that no insight occurred. However, after naive chimpanzees had *played* with sticks, they did display insight-like behavior (Birch, 1945).

3. Epstein replaced the chimpanzees with pigeons, and play-time with extensive training towards acquiring problem solution components. He found that pigeons could then solve the banana-and-box problem (in which the animal must put a box underneath a rewarding item and climb on top of the box to get the item), even though they had not seen the full problem before (Epstein, 1985, 1987; Epstein et al., 1984), by "automatic-chaining" of learned skills.

[1] At least for the general case; there have been results for toy problem domains, see for instance the work of Bundy et al. (1976).

More recently, Cook and Fowler (2014) replicated Epstein's results (with some modifications in the design) and extended them to demonstrate that the pigeons did not use "means-end processing" based on an understanding of the functional relationships between objects (i.e. an understanding that a box is useful because it is physically possible to stand on it)[2].

These results suggest that insight is heavily dependent on prior experience, and that it might not rely on planning using a physical model of the task. In this chapter, I focus on replicating Epstein's "simple" task, using training similar to that received by his pigeons. Not only does Epstein's work provide a good basis for more in-depth understanding of insight, but replicating it can also serve to highlight the limitations of automatic-chaining approaches to insight and creativity[3]. In the next chapter I will attempt to move beyond some of these limitations.

4.1.3 Epstein's experiment

The problem to solve

The experiment[4] by Epstein et al. (1984) is a reproduction of Köhler's banana-and-box experiment. In Köhler's task, a banana was nailed to a ceiling, out of chimpanzee reach. The chimpanzees then used a box, present in their enclosure among other objects, as an elevated platform to reach for the banana. They moved the box underneath the banana, then jumped from it. The behavior was considered "insightful" because it occurred all at once after a period of apparent hesitation and confusion, thus presenting signs resembling those of insights as observed on and experienced by human beings.

Chimpanzees would naturally want to acquire a banana; pigeons are instead trained to peck a facsimile banana (after which they received food as reward). In the "test" situation the facsimile banana is suspended from the ceiling of the room, such that pigeons cannot reach it by stretching towards it (they do not attempt to fly towards it (Cook and Fowler, 2014)). However, a light carton box is placed in the enclosure, which can be used to reach

[2] Much of the recent literature on animal insight focuses on similar issues of "functional understanding", which were of some importance in Gestalt views. Often this leads to "killjoy" results (Shettleworth, 2010), i.e. results that show the animals do not have an understanding of the objective physical properties of their environment. In the next chapter I will argue that insight need not involve "functional understanding" of that sort, in either humans or animals.

[3] Although there is no prior work connecting reinforcement learning and insight, there is some work on "skill-chaining" (Konidaris and Barto, 2009; Konidaris et al., 2011) in hierarchical reinforcement learning, which is independent and seemingly unaware of Epstein's "automatic chaining".

[4] The reader is encouraged to consult the original description - it is a short letter to Nature (1.5 pages). Subsequent similar experiments are described in more detail Epstein (1985, 1987).

the banana: the problem is solved when the animal pushes/pecks the box underneath the banana and, standing on the box, reaches for/pecks at the (facsimile) banana; see figure 4.1.

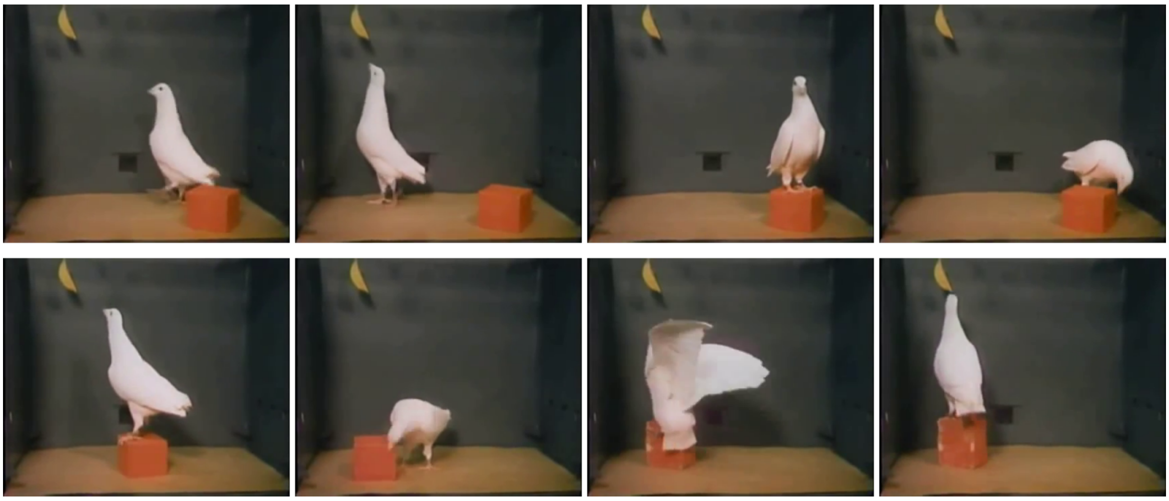


Fig. 4.1 Left-to-right, then top-to-bottom: an "exceptionally human-like" successful test by a pigeon (the snapshots are from a video recording by Epstein et al. (2007)). This figure does not show the whole process, only certain key steps. In this particular case, the whole process takes 60 seconds. (This figure was made using screenshots of a video from Epstein's experiment, with permission.)

Training

Pigeons, like chimpanzees, are not able to solve the task with no prior experience. But unlike the chimpanzees of Köhler or even Birch, the pigeons are given extensive training in the type of behavior required to solve the task. This is done using *shaping*. Skinner (1953) [5] describes shaping as follows:

> We first give the bird food when it turns slightly in the direction of the spot from any part of the cage. This increases the frequency of such behavior. We then withhold reinforcement until a slight movement is made toward the spot. This again alters the general distribution of behavior without producing a new unit. We continue by reinforcing positions successively closer to the spot, then by reinforcing only when the head is moved slightly forward, and finally only when the beak actually makes contact with the spot. (...) The original probability of the response in its final form is very low; in some cases it may even be zero.

[5] Many of Epstein's experiments on pigeons, though not the one that interests us here, were done in collaboration with B.F. Skinner, e.g. Epstein and Skinner (1980).

> In this way we can build complicated operants which would never appear in the repertoire of the organism otherwise. By reinforcing a series of successive approximations, we bring a rare response to a very high probability in a short time. (...) The total act of turning toward the spot from any point in the box, walking toward it, raising the head, and striking the spot may seem to be a functionally coherent unit of behavior; but it is constructed by a continual process of differential reinforcement from undifferentiated behavior, just as the sculptor shapes his figure from a lump of clay.

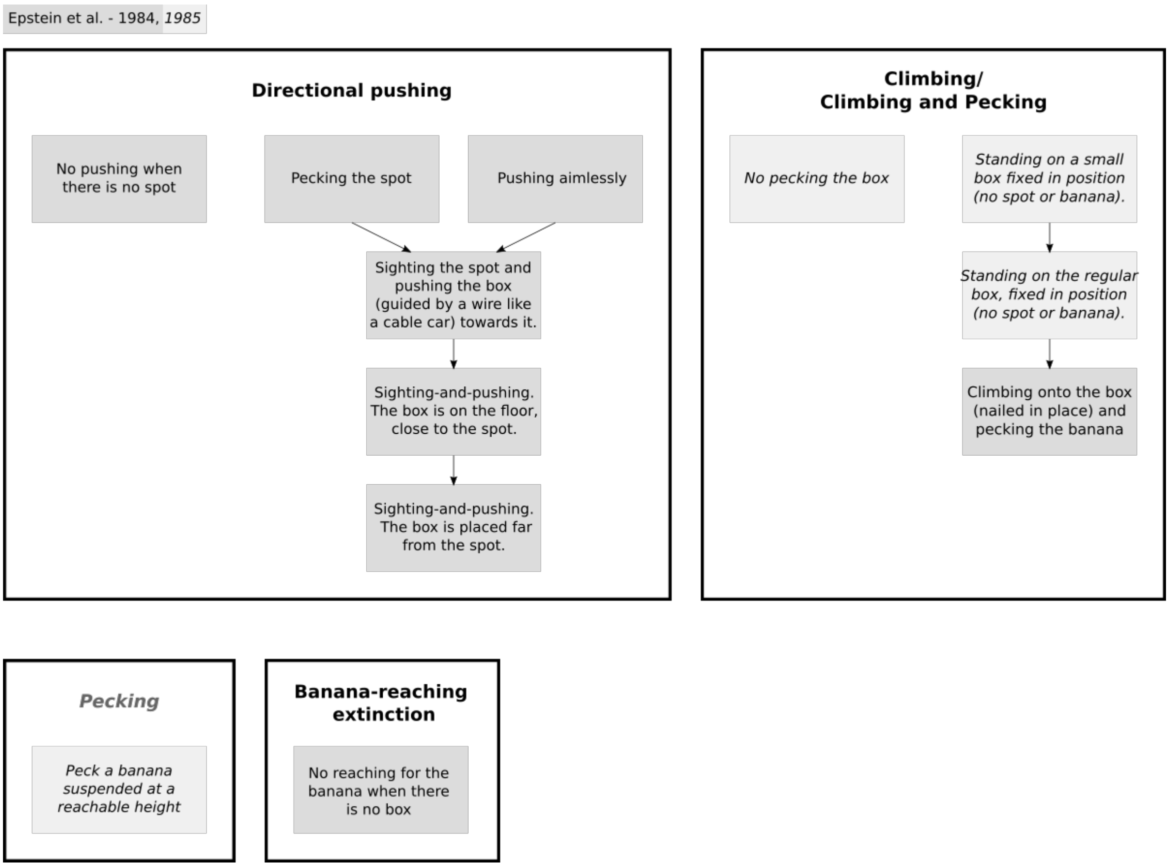

Epstein et al. - 1984, 1985
Directional pushing
No pushing when there is no spot
Pecking the spot
Pushing aimlessly
Sighting the spot and pushing the box (guided by a wire like a cable car) towards it.
Sighting-and-pushing. The box is on the floor, close to the spot.
Sighting-and-pushing. The box is placed far from the spot.
Climbing/ Climbing and Pecking
No pecking the box
Standing on a small box fixed in position (no spot or banana).
Standing on the regular box, fixed in position (no spot or banana).
Climbing onto the box (nailed in place) and pecking the banana
Pecking
Peck a banana suspended at a reachable height
Banana-reaching extinction
No reaching for the banana when there is no box

Fig. 4.2 Shaping as used by Epstein and colleagues to train pigeons in the banana-and-box task. Arrows indicate temporal progression. This figure is based scrupulously on what Epstein et al. report in two separate experiments published in 1984 (dark grey) (Epstein et al., 1984) and 1985 (light grey, italics) (Epstein, 1985); however the description provided by Epstein is itself possibly incomplete: the 1984 article reads “Major training steps included...”, suggesting there was more to the training. Shaping the behavior took up to eight weeks (1984) or 39 sessions/28 hours (1985).

The shaping regimen used by Epstein et al. (1984), and Cook and Fowler (2014), are shown in figures 4.2 and 4.3 respectively. Note that the exact training differs importantly,

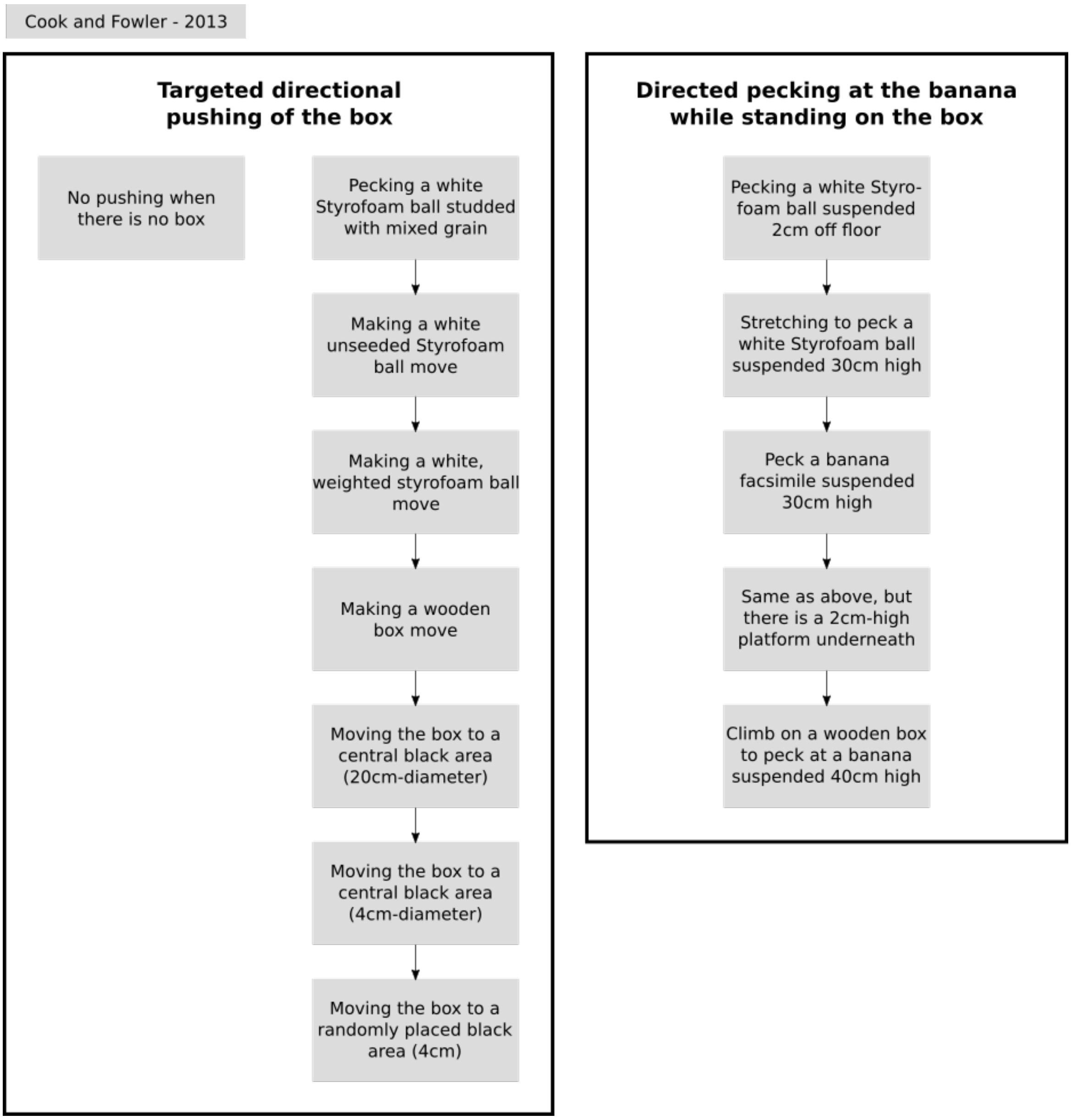

Cook and Fowler - 2013
Targeted directional pushing of the box
No pushing when there is no box
Pecking a white Styrofoam ball studded with mixed grain
Making a white unseeded Styrofoam ball move
Making a white, weighted styrofoam ball move
Making a wooden box move
Moving the box to a central black area (20cm-diameter)
Moving the box to a central black area (4cm-diameter)
Moving the box to a randomly placed black area (4cm)
Directed pecking at the banana while standing on the box
Pecking a white Styrofoam ball suspended 2cm off floor
Stretching to peck a white Styrofoam ball suspended 30cm high
Peck a banana facsimile suspended 30cm high
Same as above, but there is a 2cm-high platform underneath
Climb on a wooden box to peck at a banana suspended 40cm high

Fig. 4.3 Shaping as used by Cook and Fowler (2014) in a replication of Epstein's pigeon and banana experiment. This is likely a closer, more detailed description of the shaping involved in training pigeons. The training took "several weeks to complete".

even though the pigeons' final performance is highly sensitive to the kind of training they have received[6].

[6] Cook and Fowler did not immediately succeed: *"During the first test, the pigeon's extensive and directionless box-directed pushing behavior was the obvious problem. A rereading of Epstein et al. (1984) noted that extinguishing pushing behavior in the absence of the dot was included in their training, something we had not done. Instead, during training, #2B"* (one of the pigeons) *"had received immediate reward upon pushing the box onto the dot. Although this increased the frequency of getting the box to the target dot, it did not require*

4.2 Pigeon models

This section describes the models of the learning and decision-making of the simulated pigeons. I begin by describing the neural network architecture which is common to all models. I then test several popular deep reinforcement learning algorithms on a simplified version of the task. Finally I present three algorithms: an actor-critic algorithm, and two hierarchical systems designed to instantiate the skills implied by Epstein's theory of "automatic-chaining".

4.2.1 Neural network architecture

The artificial pigeons are implemented using deep reinforcement learning architectures, involving one or several neural networks trained via a combination of temporal difference based algorithms and backpropagation. But all algorithms tested share a similar network architecture.

The architecture of an actor network is shown in figure 4.4. The critic network is identical save for the output layer, which has only one unit and has no non-linearity (thus allowing it to output any scalar value). Training in these networks is achieved using gradient descent (backpropagation), according to the update equations given in the next subsections.

4.2.2 Flat model

Preliminary testing

Models that make no use of hierarchical structure are called "flat"; most popular reinforcement learning algorithms are flat. To determine the suitability of existing reinforcement learning algorithms for this environment, and to inform the design of a more complex HRL algorithm, a range of flat reinforcement learning algorithms was implemented and tested on a simplified version of the task.

- Deep Q-Networks (DQN) (Mnih et al., 2015): a critic-only, off-policy algorithm using experience replay (Lin, 1992). It is a variant of Q-learning (see chapter 3, section 10) that makes use of a neural network to approximate the value function. Experience replay serves to increase data-efficiency, while also stabilizing learning by reducing the correlations between successive "sarses" (a SARS or sars stands for the quadruple consisting of a state, action, reward, and next state).

the pigeon to adjust or stop the box onto the dot as a "goal". Thus, the pigeon may have learned to just push the box until reward was delivered." (Cook and Fowler, 2014, p. 210). Subsequently, Cook and Fowler added extinction training.

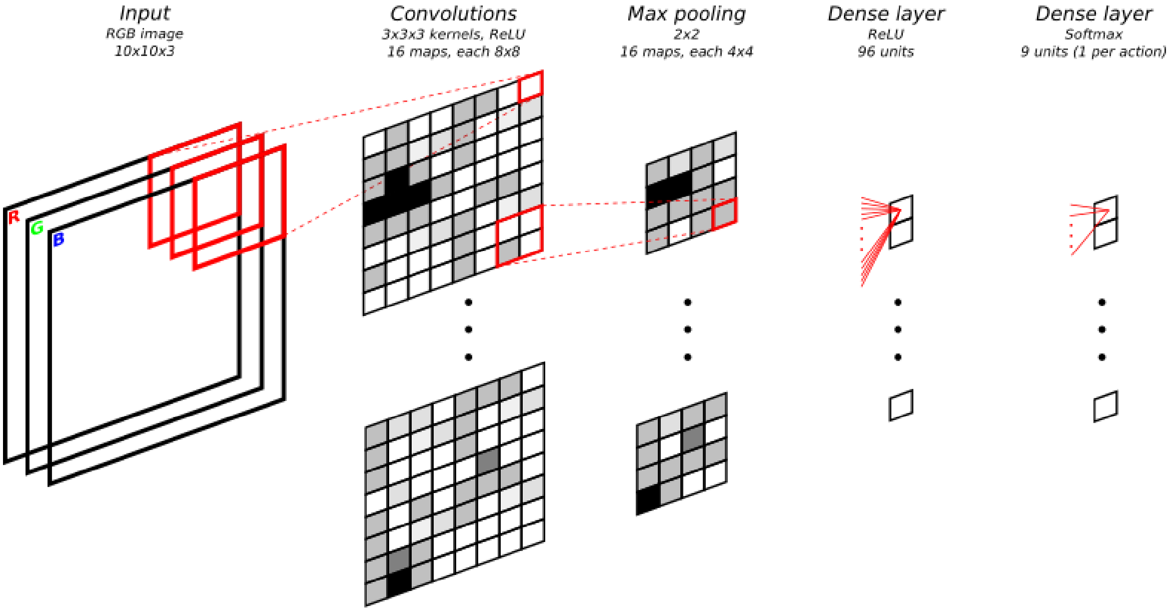


Fig. 4.4 The neural network architecture used for the actor. In red are shown example connections for each operation, illustrating how each operation relates to the output of the preceding one. Left to right: the **input** is a 10 by 10 RGB image. One **convolution** is performed, using 16 filters. Each filter is a sliding window, processing 3x3x3 slices of the image. Each filter produces one feature map of dimensions 8 by 8. In the **max pooling** operation, each feature map is reduced to a quarter of its size by dividing it into 2 by 2 windows and taking the maximal activation for each. Thus there are 16 feature maps of dimensions 4 by 4. These 16x4x4 = 256 units serve as input for a **first dense layer**. Finally, a **second dense layer** uses a Soft-max activation function, transforming the scalar preference for each action into a probability of taking that action. For networks used as critics rather than actors, there is no activation function on the last layer and only one output unit, encoding the estimation of the expected return.

- Actor-critic (see chapter 3, section 3.2.2): an on-policy algorithm which makes use of two elements, the critic and the actor. The critic works similarly to a Deep Q-Network, but evaluates only a state (rather than a state-action pair), and because of its on-policy nature cannot make use of experience replay. The actor encodes preferences for actions, and learns based on feedback given by the critic.

- A2C: a variant of the popular A3C algorithm of Mnih et al. (2016). It is also a type of actor-critic architecture. In contrast to a vanilla actor-critic algorithm, its updates are not done after each time-step, but by making “batches” over a certain buffer period. Additionally, A2C makes use of parallelism - several environments are run in parallel; like experience replay, this helps to stabilize learning by more closely approximating the iid setting on which Deep Learning works best; this also makes better use of parallel processing capabilities.

- Parallelism was also tested for Actor-Critic and Deep Q-Networks.

All algorithms were tested on a simple task: pecking a coloured spot, with the pigeon and the spot spawning in a different random location for each episode; episodes were aborted if still unsuccessful after 50 timesteps (for more details about the task and environment, see section 4.3). The results are shown in figure 4.5. All algorithms make use of the same network architecture; the actor-critic algorithms (actor-critic, parallel actor-critic, A2C) used one network for the actor, and one network for the critic, with only the final layer differing (as previously discussed in subsection 4.2.1). The other hyperparameters used are shown in table 4.1. Further details and discussion of these results can be found in appendix E.

	Learning rate (actor)	Learning rate (critic)	Parallel MDPs	Experience Replay
Q-learning	n/a	0.006	1	n/a
Parallel Q-learning	n/a	0.006	16	n/a
DQN	n/a	0.001	1	15
Actor-critic	0.02	0.002	1	n/a
Parallel actor-critic	0.01	0.001	16	n/a
A2C	0.0003	0.00003	16	n/a

Table 4.1 Hyperparameters used in the evaluation of six reinforcement learning algorithms (cf. figure 4.5). These values were obtained by manual optimization, with a focus on achieving high learning rates without causing instability. Experience replay refers to the number of replayed experiences (sarses) for each novel experience. The algorithms are implemented using as a basis the algorithms shown in chapter 3, sections 3.1.4 for Q-learning, and 3.2.2 for actor-critic; function approximation is done using the neural network architecture described in figure 4.4.

Choice of algorithm

The two best-performing algorithms were parallel Q-learning and parallel actor-critic (where "parallel" has the meaning presented in 3, section 3.2.1: multiple agents are trained in parallel to allow for training with mini-batches containing one sars per agent). The parallel actor-critic algorithm was preferred for three reasons:

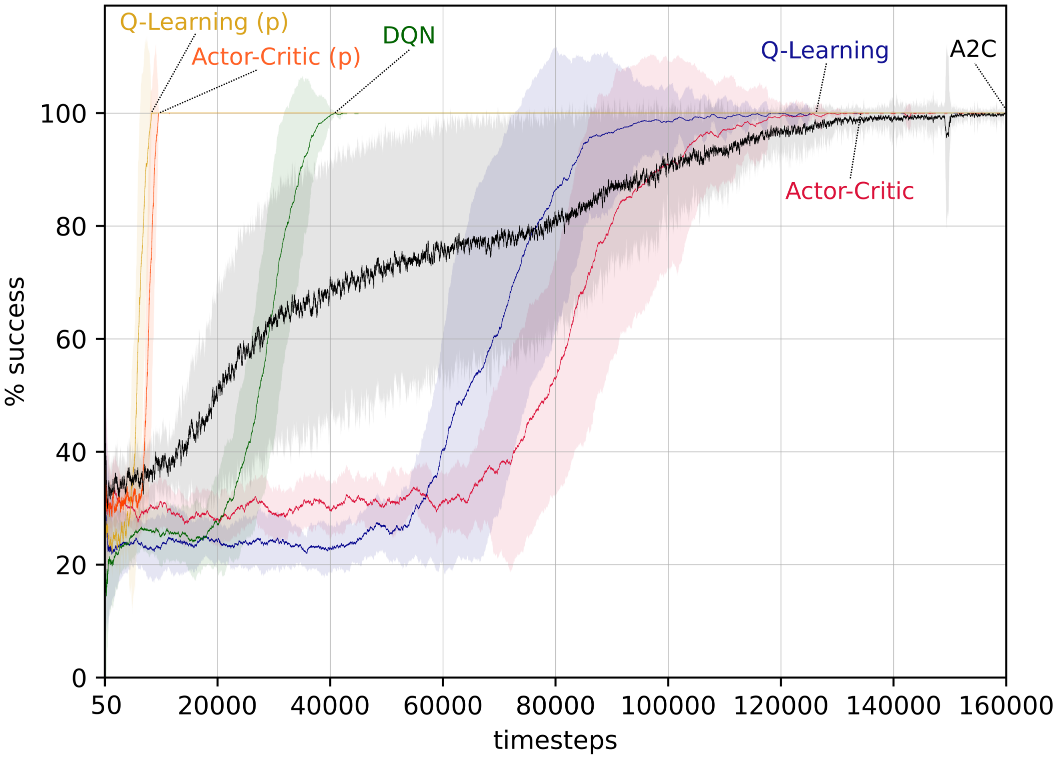


Fig. 4.5 Test of various RL algorithms on a simple Epstein-inspired MDP (“peck the spot”). “(p)” stands for “parallel”. The chance of success is the average over the last 100 episodes; standard error is shown. The curves are labeled where they achieve around 100% success. Although parallel algorithms made use of much more experience, all algorithms used roughly equal amounts processing time per time-step (except for A2C, which was slower). Thus parallel actor-critic and parallel Q-learning were the fastest algorithms. A2C was surprisingly slow to converge; its learning rate had to be kept low due to instability issues, arising possibly from the highly correlated nature of the batches (each episode had specific initial conditions and was highly internally correlated).

- Extinction, by withdrawing of reward, is likely easier to achieve with preference-based decision making (Soft-max), rather than maximization-based decision making (ε-greedy) (Sutton and Barto, 2018, p. 322).
- It was conjectured that agents that have a separate actor would be more stable in response to changes in the reward structure of the environment.

- Most hierarchical reinforcement learning models are also on-policy, including the one tested in this chapter; using an on-policy flat model makes for more informative comparisons.
- Finally, actor-critic algorithms are considered the most biologically plausible reinforcement learning algorithms (Sutton and Barto, 2018, chap. 15).

The actor-critic model uses the following standard (Sutton and Barto, 2018, chap. 13) updates, as seen in chapter 3, section 3.2.2. At each time-step, the temporal difference error is calculated as[7]:

$$\delta \leftarrow R + \gamma \hat{v}(S', w) - \hat{v}(S, w)$$

The critic parameters are then updated according to:

$$w_h \leftarrow w_h + \alpha_w \delta \nabla \hat{v}(S, w)$$

Finally the actor parameters are updated according to:

$$\theta \leftarrow \theta + \alpha_\theta \delta \nabla \log \pi(A|S, \theta)$$

Due to its relative simplicity, the flat algorithm could be applied “as is” to both shaping tasks and to the testing task; see section 4.3.3.

4.2.3 Hierarchical model

Epstein et al. (1984) and Epstein (2014) suggest that distinct, independent skills are learned. The pigeon may then pick and choose appropriate skills to use depending on its circumstances. This corresponds to a hierarchical architecture of behavior in which low-level behaviors are controlled by a low-level controller (or skill, or behavioral repertoire), while a high-level controller decides on which option (repertoire, skill) to activate.

There is a paucity of models for Deep Hierarchical Reinforcement Learning. The most promising and popular technique is the Option-Critic architecture of Bacon et al. (2017), which is fully differentiable and can therefore be trained using gradient descent; but this architecture has no clear mechanism for stipulating which option(s) ought to be trained from a given episode, and, left to its gradients, can produce options that bear little resemblance

[7] Recall that, in these equations, δ is the temporal difference error, R is the reward, γ is the discount rate, S and S' are two successive states, w is the parameter vector for the critic (which serves to compute the value function $\hat{v}(S,w)$), θ is the parameter vector for the actor (which serves to compute the policy $\pi(A|S,\theta)$), and α_w and α_θ are the learning rates for the critic and actor respectively.

to the intended repertoires[8]. Furthermore, the experiment used by Epstein has two distinct phases, a training/shaping phase during which the pigeons acquire skills, and a testing phase during which they transfer them to the insight problem. Thus I propose an adapted version of Option-Critic, designed to allow for the training of distinct skills specified by the experimenter in two distinct "shaping" and "testing" phases.

The architecture is shown in figure 4.6. It makes use of three networks for the test phase: a critic which doubles as a controller over options, and two option-actors. Additionally, two temporary critic networks were used during shaping to train the options. This model allows for training separately the policies corresponding to different repertoires (one actor network per option), and then choosing between these options using the controller over options. This architecture is heavily inspired from the Option-Critic architecture (Bacon et al., 2017). The main changes are the double use of the controller-over-option (as both a critic and an actor), and the prior training of the actor-networks in order to create the desired skills.

During the shaping phase, each actor-critic system is trained in the standard way, using updates identical to those of the flat agent (except for the learning rates; see appendix E):

$$\begin{aligned}
\delta &\leftarrow R + \gamma \hat{v}(S', w) - \hat{v}(S, w) \\
w_h &\leftarrow w_h + \alpha_w \delta \nabla \hat{v}(S, w) \\
\theta &\leftarrow \theta + \alpha_\theta \delta \nabla \log \pi(A|S, \theta)
\end{aligned}$$

After shaping is completed, the policies (the actor networks) are kept and a new critic is introduced, which evaluates value not merely based on states, but based on states and options. As a result the new critic can work simultaneously as a Sarsa controller deciding between options, and as a standard critic deciding between actions given a state and action. For this second phase, the following update equations are used:

$$\begin{aligned}
\delta &\leftarrow R + \gamma \hat{q}(S', O', w) - \hat{q}(S, O, w) \\
w &\leftarrow w + \alpha_w \delta \nabla \hat{q}(S, O, w) \\
\theta_o &\leftarrow \theta_o + \alpha_o \delta \nabla \log \pi(A|S, \theta_o)
\end{aligned}$$

Where: O and O' are two successive options, w is the parameter vector for the Sarsa agent-over-options, θ_o is the parameter vector for the actor for option o, and α_w and α_o are the learning rates for the controller over options and the option, respectively.

The full pseudo-code is given in algorithm 6, and a visual explanation is provided in figure 4.6.

[8] Suggestive evidence is provided in appendix E.3.

Algorithm 6: Hierarchical model B

```
Input    : A list shapingMDPs of shaping pairs: (MDP, stop-condition), plus
           another such pair for the final test
Require  : Interaction with shaping and final test MDPs using the functions
           observe, perform; learning rates $\alpha_1, \alpha_2, \alpha_3, \alpha_4$
Output   : Parameters for a hierarchical policy.
// Shaping phase: semi-gradient actor-critic learning
foreach $i, (\mathsf{mdp}, \mathsf{stop\text{-}condition}) \in enumerate(\mathsf{ShapingMDPs})$ do
    Initialize parameters $w_i$
    Initialize parameters $\theta_i$
    $\mathsf{s} \leftarrow$ observe()
    repeat
        $\mathsf{a} \leftarrow$ decide$(\mathsf{s}, \theta_i)$
        $\mathsf{r} \leftarrow$ perform$(\mathsf{a}, \mathsf{mdp})$
        $\mathsf{s}' \leftarrow$ observe$(\mathsf{mdp})$
        $\delta \leftarrow \mathsf{r} + \gamma \hat{v}(s', w_i) - \hat{v}(s, w_i)$      // Temporal difference error
        $w_i \leftarrow w_i + \alpha_1 \delta \nabla \hat{v}(s, w_i)$                    // Critic update
        $\theta_i \leftarrow \theta_i + \alpha_2 \delta \nabla \log \pi(a|s, \theta_i)$       // Actor update
        $\mathsf{s} \leftarrow \mathsf{s}'$
    until stop-condition is True
// Final test phase: interleaved Sarsa and actor-critic
mdp, stop-condition ← final-mdp, final-stop-condition
Initialize parameters $w_f$                  // Params. for final Critic/Sarsa controller
$\mathsf{s} \leftarrow$ observe$(\mathsf{mdp})$
$\mathsf{o} \leftarrow$ decide$(\mathsf{s}, w_f)$                                  // Select an option
$\mathsf{a} \leftarrow$ decide$(\mathsf{s}, o, \theta_o)$        // Select an action according to the selected option
repeat
    $\mathsf{r} \leftarrow$ perform(a, mdp)
    $\mathsf{s}' \leftarrow$ observe(mdp)
    $\mathsf{o}' \leftarrow$ decide$(\mathsf{s}', w_f)$
    $\mathsf{a}' \leftarrow$ decide$(\mathsf{s}', o, \theta_o)$
    $\delta \leftarrow \mathsf{r} + \gamma \hat{v}(s', \mathsf{o}', w) - \hat{v}(s, \mathsf{o}, w)$     // Temporal difference error
    $w \leftarrow w + \alpha_3 \delta \nabla \hat{v}(s, \mathsf{o}, w)$              // Critic/Sarsa controller update
    $\theta \leftarrow \theta + \alpha_4 \delta \nabla \log \pi(a|s, \mathsf{o}, \theta)$               // Actor/Option update
    $\mathsf{s} \leftarrow \mathsf{s}'$
    $\mathsf{o} \leftarrow \mathsf{o}'$
    $\mathsf{a} \leftarrow \mathsf{a}'$
until stop-condition is True
return $w_f, (\theta_0, ..., \theta_n)$
```

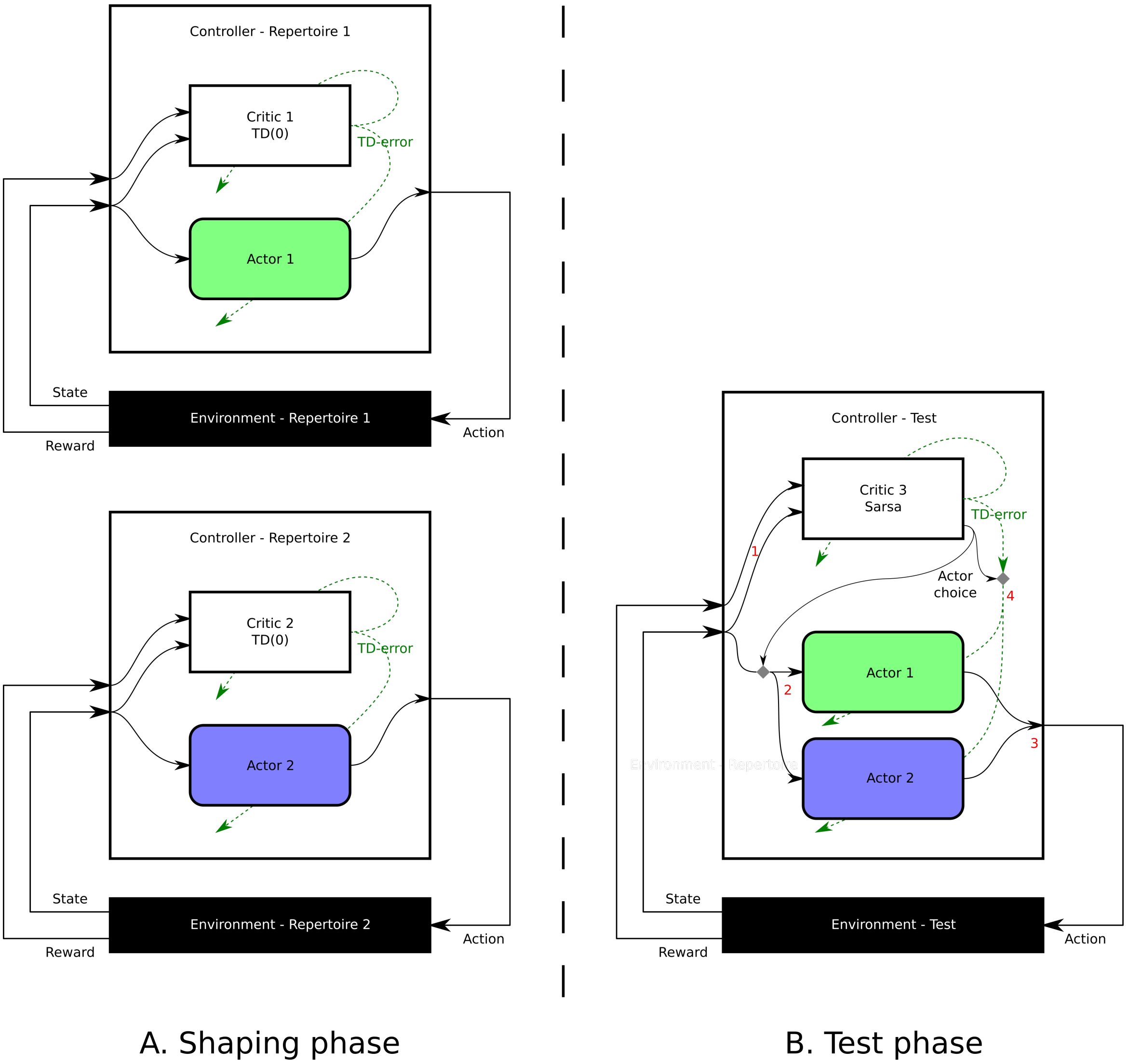


Fig. 4.6 Architecture for HRL model. During the shaping phase **(A)**, distinct behaviors are learned using a basic actor-critic architecture and the TD-error learning signal. Each actor learns the behavior associated with one shaping condition. During the test phase **(B)**, the agent architecture changes to allow for two levels of decision: a choice between which skill to use, made by the network Critic 3, and, given a skill, a choice of action, made by either Actor 1 or Actor 2. In the test phase, for each time-step processing goes as follows: **(1)** the state is forward propagated through Critic 3, which can then compute the value associated with each skill. A skill is selected based on these values (exploring with a probability ε). **(2)** Next, whichever skill was selected, Actor 1 or Actor 2, is used to select an action. **(3)** The agent performs the action, and the environment returns a state and reward, such that the temporal difference error can be calculated by Critic 3. **(4)** This is used to conduct learning at both the skill-selection level (Critic 3) and the action-selection level (whichever of Actor 1 or 2 was responsible for the action).

Biasing the hierarchical agent towards temporally extended skill-use

Usually, hierarchical agents are proposed specifically due to their ability to achieve *temporal abstraction* (Sutton et al., 1999). But the hierarchical agent, as described thus far, has no particular reason to adopt temporally extended behavior: why not pick a different option at every time-step? As a result, when exploring it has no motivation to persist in using an option long enough to achieve something worthwhile.

A similar issue arises in the Option-Critic architecture, such that the agent tends to degenerate towards single-step options, which immediately terminate (Bacon et al., 2017). The problem is solved by introducing an intrinsic negative reward for terminating an option, which biases learning towards the discovery of long-lasting options.

In the proposed hierarchical algorithm, an analogous solution is available, which consists in calculating the temporal difference error and the critic update as follows:

$$\begin{aligned}
&\delta \leftarrow R - \varepsilon g(O,O') + \gamma \hat{q}(O,S',O',w) - \hat{q}(O_p,S,O,w) \\
&w \leftarrow w + \alpha_w \delta \nabla \hat{q}(O_p,S,O,w)
\end{aligned}$$

Where O_p denotes the previous option, and $g(O,O')$ is a function that returns 1 if the options are different, and 0 otherwise, and ε is a small number, controlling the magnitude of the negative reward. Informally, this solution consists in introducing a small penalty every time the agent switches between options, thus encouraging the temporally extended use of options.

The above equations suggest that the previous option ought to be available as an input to the critic network, if only so that the network could "know" which option to persist with. However since the effect of the previous option on the final value is known in advance, it can be "hard-coded" into the critic network by simply adding the expected negative reward to the output values; i.e. the information is sent directly to the last layer of the network, with the effect of hard-coding a bias in favor of repeating the same option-choice. Mathematically, this just means isolating the contribution of the previous option on the expected reward:

$$Q(O_p,S,O,w) = Q(S,O,w) + \varepsilon g(O,O')$$

Thus, in this architecture, it is not necessary to propagate information about the previous option through the critic network; the intrinsic bias in favor of re-using the same option can be added directly to the output. The implementation of this technique in the neural network architecture is perhaps most easily understood via its illustration in figure 4.7, which shows the network used for the Sarsa controller-over-options in the HRL agent. Note that, on the last layer, a one-hot vector representing the previously active option serves to introduce a bias for selecting the same option again.

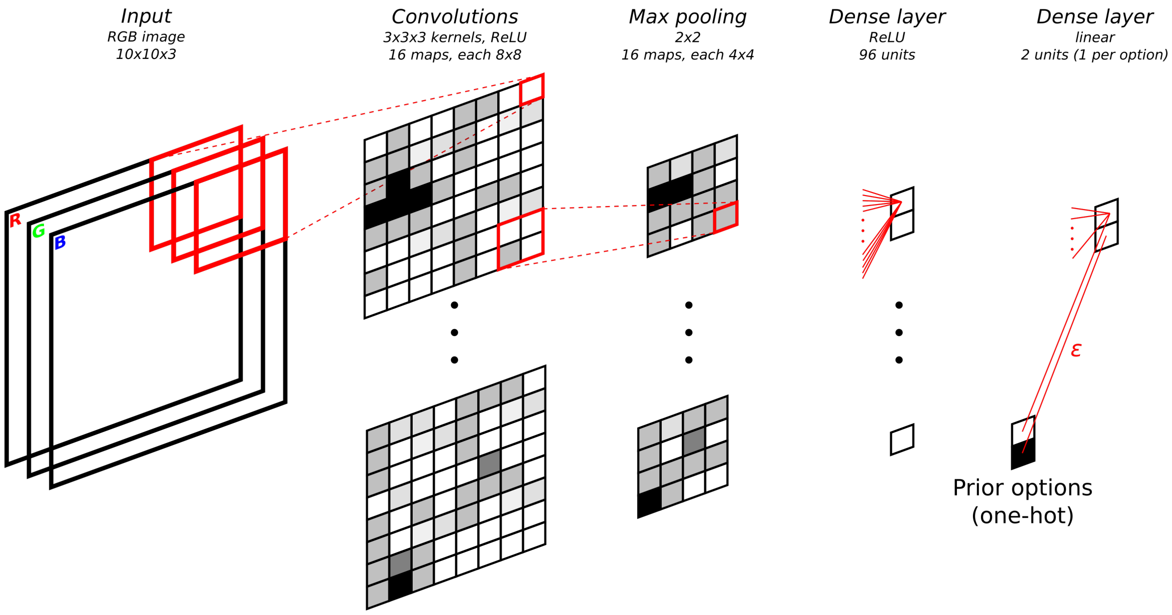


Fig. 4.7 The controller-over-options network (corresponding to "Critic 3" in figure 4.6). Note the additional input represented by the previous options with one-to-one connections with the output layer, which directly informs the value output of the additional cost, and thereby biases behavior towards consistency over time.

Transitioning from shaping to test

In early tests, it appeared that transitioning from the shaping regime to the test regime, thereby switching to a "naive" critic, led to instability in the option-networks. To avoid this the option-networks were frozen for 40,000 time-steps, such that the policy over options could stabilize. Only afterwards was learning resumed in the entire hierarchical agent at all levels. The procedure is somewhat reminiscent of the strategy used by Hinton and Salakhutdinov (2006), who trained neural networks layer-by-layer, freezing other layers. However in the present case it is "layers of behavior", rather than hidden layers in a neural network, that are frozen or trained in isolation.

Simpler version of the HRL model

In addition to the hierarchical model described so far, a simpler version was also tested. This model was trained identically during the shaping phase, using two separate critic and actors.

During the test phase, the original critics were kept, and whichever actor-critic was most "confident" (had the higher value for the current state) was given control.

This model had the advantage of being comparatively simple, but was more speculative, and lacked theoretical foundations. It proved to be unstable in practice (see section 4.4).

4.3 Methods

We have seen in section 4.1 how Epstein's experiment relates to the scientific literature on insight, and in section 4.2 how pigeon decision-making is implemented using deep RL and deep HRL techniques. In this section, I discuss the technical details of simulating the task.

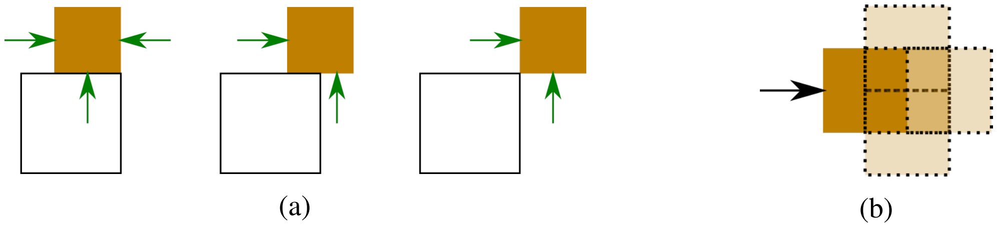
4.3.1 The Epstein environment

To reproduce the problem faced by Epstein's pigeons, who are not newborns but experienced laboratory animals, the interface between agent and environment consists of relatively sophisticated "elementary" actions (walking, pecking, pushing, and jumping) and situational awareness (position of objects in space).

Walking is straightforward (performing the actions "up", "down", "right", "left" allows the pigeon to move by one unit of distance in that direction unless there is an obstacle). Jumping allows the pigeon to get on top of the box if it is in contact with it, or off the box if it is on top of it. If the pigeon is in contact with an object, it can peck it towards certain directions; for instance, if the pigeon is to the south of an object, it can peck it "up" towards the north. The effects of pecking depend on both the presence of obstacles and chance. Figure 4.8 illustrates pecking rules and box movement.

Figure 4.9 shows the correspondence between the situation and the image sensed by the artificial agent: 10 by 10 by 3 (RGB), containing up to 3 objects - the pigeon, the box, and either the banana or the spot. At any time, the pigeon can perform one of 9 actions: moving in either cardinal direction, pecking at an object in either cardinal direction, and jumping. Figure 4.9 displays the representations for various states encountered while solving the problem.

Transforming the shaping know-how of the experimentalist into an algorithmic procedure proved an easy task from a technical standpoint (a substantial literature in AI investigates techniques resembling shaping, for instance Bengio et al. (2009); Florensa et al. (2017); Ng et al. (1999)), but an awkward task from a methodological standpoint. Ideally, one would like to closely imitate the experimental shaping program used by an animal experimentalist; however, it has proven difficult to obtain detailed information about the details of shaping


(a)

(b)

Fig. 4.8 **(a)** Pecking rules: which pecking actions can have an effect, depending on the respective positions of the pigeon (white square) and box (brown square). Note that this allows the pigeon to peck a box out of a corner.
(b) Box dynamics: pecking the box towards the right can result in 4 equiprobable outcomes (fewer if there are obstacles to the movement of the box).

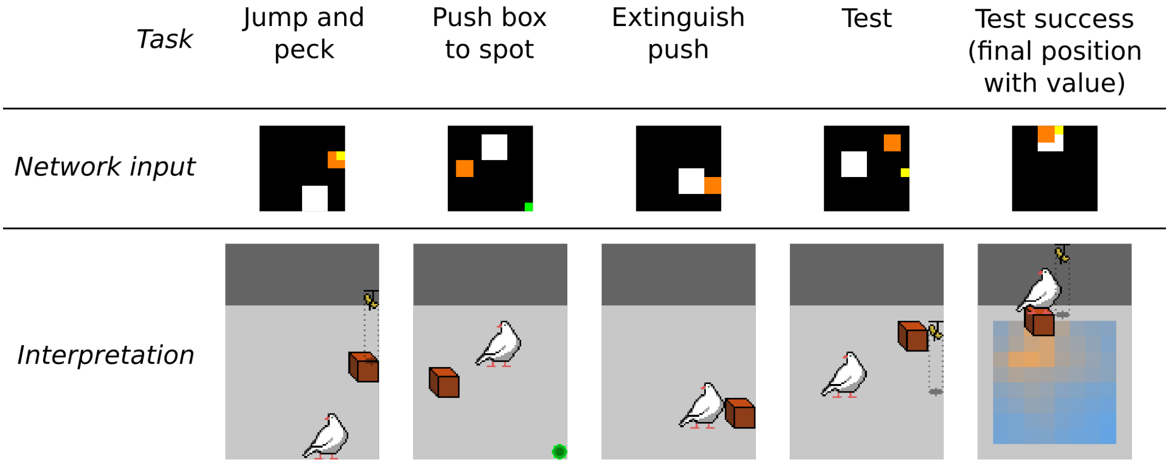


Fig. 4.9 Pigeon training tasks, example network inputs, and interpretations. On the last image, a value overlay (best viewed in color) is displayed - this overlay shows an estimation of the value for each position of the pigeon, considering its current policy, assuming the other objects stay in place.

procedures, and the exact means by which these procedures are chosen. Instead, I chose to use a simplified procedure, which leaves less room for optimization to the AI experimentalist.

Artificial shaping is conduced as shown in figure 4.10. The shaping program has therefore been simplified compared to those of Epstein et al. (1984) or Cook and Fowler (2014). Shaping is considered successful once the pigeon is successful on 50 successive trials (intermediary shaping stage) or 100 successive trials (final stage); in practice this corresponded to a success rate of about 92%.

Example shaping tasks were shown in figure 4.9. Figures 4.11, 4.12, and 4.13 demonstrates step-by-step how each task is solved in the simulated environment. Skill 2 is considerably more difficult, as it requires pigeons to make slow progress across the environment while pushing the box towards the target, and to position themselves in order to dislodge the

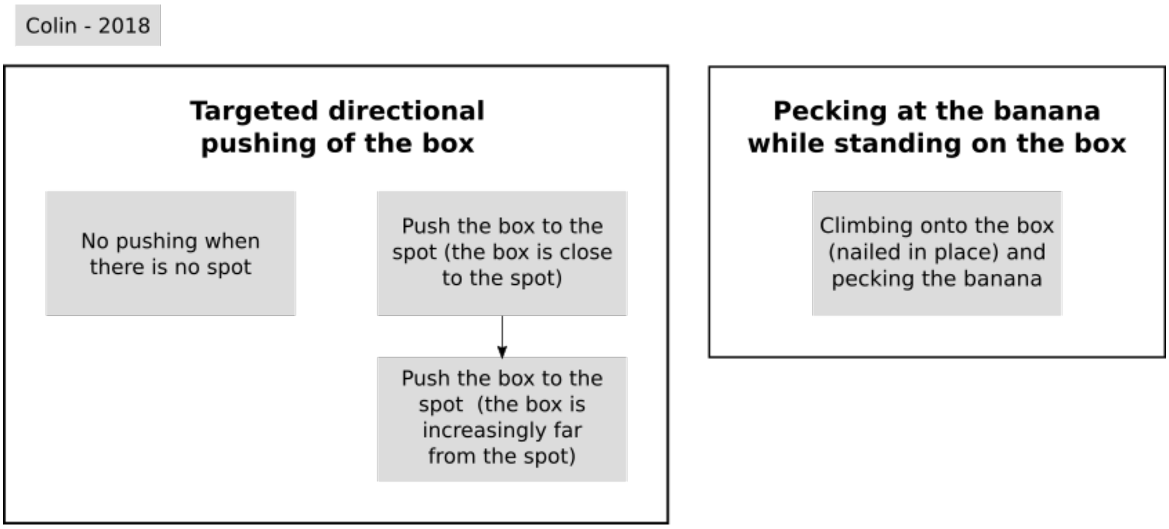

Colin - 2018
Targeted directional pushing of the box
No pushing when there is no spot
Push the box to the spot (the box is close to the spot)
Push the box to the spot (the box is increasingly far from the spot)
Pecking at the banana while standing on the box
Climbing onto the box (nailed in place) and pecking the banana

Fig. 4.10 Shaping/training program for the artificial pigeons.

box from corners when it gets stuck there. The test environment requires making use of both skills, with the added twist that skill 2 is trained with respect to a green spot, whereas in the test environment, the box must be pushed towards a yellow "banana" instead.

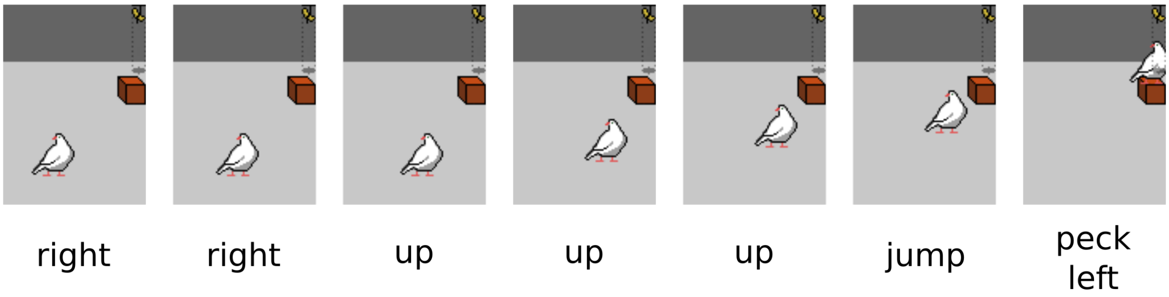


Fig. 4.11 Successive screenshots of an artificial pigeon solving the "jump and peck" shaping task. On top are shown the states, below are shown the corresponding actions taken by the artificial pigeon. The result of the action is visible in the next state. The last action successfully terminates the episode (the agent receives a positive reward).

4.3.2 Comparison with Epstein's banana-and-box task

The task closely matches the experimental set-up as described and videotaped by Epstein and colleagues (Epstein et al., 1984, 2007). In particular:

- The square environment is one of two used by Epstein[9];

[9] For practical reasons, I chose the square environment over the circular environment.

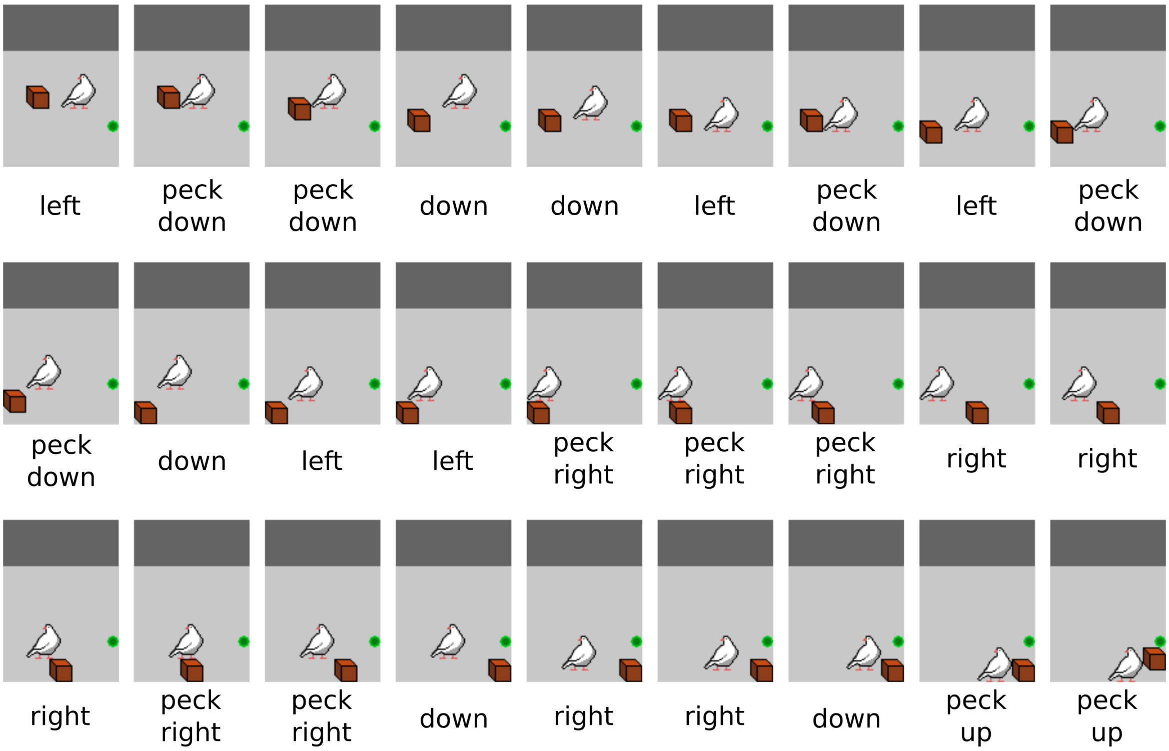


Fig. 4.12 Successive screenshots of an artificial pigeon solving the "push box to spot" shaping task. Note that the pigeon succeeded despite a sub-optimal policy (first pushing the box in the wrong direction). Also note the stochasticity of the "peck" actions: pecking actions have up to 4 different outcomes (less when it is constrained by the position of a wall or of the pigeon).

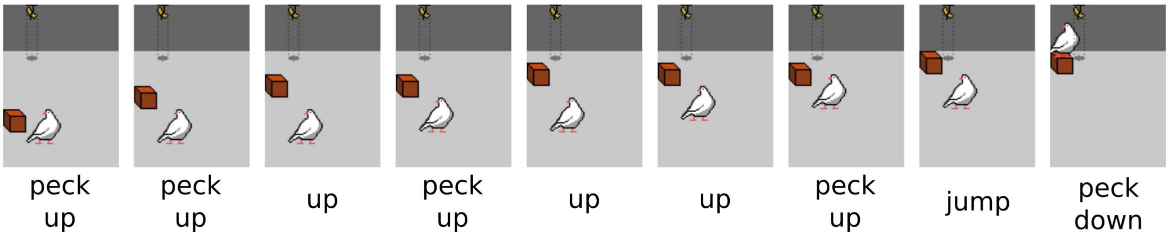


Fig. 4.13 Successive screenshots of an artificial pigeon finishing to solve the test task.

- The dimensions of the simulated objects, animals, and the effects of actions, are consistent with those used in the experiment. The room is 10 by 10 (41 by 41 cm for Epstein et al. (1984)), the pigeon 3 by 3, the box 2 by 2 (10 by 10 cm), the banana 1 by 1 (7 cm in length), and the spot 1 by 1 (4 by 4 cm)[10]. Likewise the stochastic movements of the box when pushed are consistent with those measured by Cook and

[10] That is, their relative dimensions are generally consistent with the relative dimensions of the objects and animals, as can be verified by comparing screenshots of the simulation to videos snapshots.

Fowler (2014) (3.5 cm on average) and observable in Epstein's video of a pigeon's performance.

- The effects of actions are consistent with those observed in the experiment. In particular, the pigeon pecks the box "directionally" to move it, rather e.g. than leaning its body against it; this allows pigeons to dislodge the box even if it is stuck to a wall or corner.

There are however some notable differences:

- A top-down view is used instead of the first-person view of the pigeon.
- The artificial pigeons do not need to learn the precise motor control involved in pushing the box or jumping on top of it, as their basic actions already accomplish these complex movements.

The use of the top-down view may seem like an unfair advantage for the algorithm. However, the algorithm is almost completely naive as the simulation begins, whereas pigeons have considerable prior experience of both real-world interaction and learning in laboratory tasks[11]. The choice of the top-down view thus helps "level the field" for the naive artificial pigeons, in comparison to the experienced actual pigeons; likewise for the structured actions.

4.3.3 Experimental procedure

Flat agents

For flat agents, the same actor-network learns to perform both skills. When the agents received prior training, training the two skills was interleaved, in proportions 1/8th and 7/8th for the (easy) jump-and-peck and (difficult) push-to-spot conditions, respectively.

A first cohort of 20 agents was directly given the test without any prior training; we call this *condition 1*.

A second cohort of 20 agents was given shaping training up to a performance of 90% completion within 50 time-steps, and then continued learning in the test condition; we call this *condition 2*.

A third cohort of 20 was given more extensive training (150,000 additional timesteps after meeting the criteria for condition 2); we call this *condition 3*. The expectation was that

[11] The importance of "learning to learn" should not be underestimated; in a classic article, (Harlow, 1949) notes that "(...) the full-brained monkeys make significantly better scores, but one should note that the educated hemidecorticate animals [who have been surgically amputated of half their cortex] are superior to the uneducated unoperated monkeys. Such data suggest that half a brain is better than one if you compare the individuals having appropriate learning sets with the individuals lacking them").

additional training would result in overfitting and render transfer more difficult (as observed for human insight by Wiley (1998)).

In all cases, the primary measurement is the rate of success: how likely each simulated pigeon is to succeed at its task within 50 time-steps. This is measured as a running average (cf. figure 4.14). Thus, summarizing:

1. Shaping: the agents train for both task environments in an interleaved manner
    - Condition 1: no shaping.
    - Condition 2: shaping continues until 90% of trials are successful.
    - Condition 3: shaping continues until 150,000 time-steps have passed beyond achieving a success rate of 90%
2. Test: the same agents train on the test environment

HRL agents

For HRL agents, different option-networks learn to perform each skill. Thus separate training is incurred by each network, until both achieve 90% completion of their task (jump-and-peck or push-to-spot) within 50 time-steps. Following this, the two types of HRL agents attempt the test scenario.

The HRL agents made use of neural networks at both levels of decision making (action-selection for the option-networks, and option-selection for the controller-over-options Sarsa network). During the test, they were first trained with "frozen" option-networks for 40,000 time-steps, in order for the policy-over-options to become stable. After this, the learning resumed at both hierarchical levels. Furthermore, these agents were tested both with and without the technique for encouraging temporally extended skill-use shown in figure 4.7. Summarizing:

1. Shaping: the relevant option is trained on the relevant task.
2. Test:
    - Condition 4: the HRL agents trains in two phases: first with frozen options (40,000 time-steps), then with learning enabled at all levels.
    - Condition 5: identical with condition 4, except that a bias towards using options for a longer duration is introduced.

4.4 Results

4.4.1 Shaping

In all conditions with shaping (2, 3, 4, and 5), the shaping procedure resulted in very fast learning compared to learning from scratch[12]. A comparison of learning without and with shaping for a non-hierarchical agent (condition 1 compared to conditions 2 and 3) is shown in figure 4.14. Shaping learning curves for hierarchical agents are shown in figure 4.18.

Figure 4.15 shows a single instance of shaping, revealing how performance changes as each step of shaping takes place: after the initial behavior is learned, successive increases in the difficulty of the task temporarily reduce agent performance. This curve shows that the agent is kept on a steep section of the learning curve, so to say, by the increasing difficulty of the environment.

Condition 3 can be seen in figure 4.14 (orange curve). Additional shaping led to, on average, slower and more variable performance in the test phase. Individual curves for conditions 1 and 2 are shown in figure 4.17.

4.4.2 Test task: flat agents

Flat agents could transfer knowledge from the shaping tasks to the test task. Indeed, the shaped flat agents (conditions 2 and 3) were faster learners than the naive agents (condition 1), even when one also considers the time previously taken by shaping.

Additional shaping was expected to produce over-fitting effects resembling the "fixedness" that gets in the way of solving insight problems, and is called the "Einstellung effect" in pyschology (Luchins, 1942). This expectation was verified, as shown by the lower performance and increased variance visible in figure 4.17.

Figure 4.16 shows what happens to subjective estimated value (by the *critic* part of the agent) and temporal difference errors (experienced by the agent) during the test. Value closely follows performance, with a sharp increase of estimated value corresponding to the sharp increase in performance.

4.4.3 Test task: hierarchical agents

Contrary to expectations, the hierarchical models (conditions 4 and 5, and simple model) did not outperform the flat agents of conditions 2 and 3 in the test task.

[12] As it turns out, the shaping procedure chosen is an instance of a general method (training at an increasingly far initial distance from the goal) that was independently proposed by Florensa et al. (2017) in the context of reinforcement learning for robotics, and which the authors call "reverse curriculum generation".

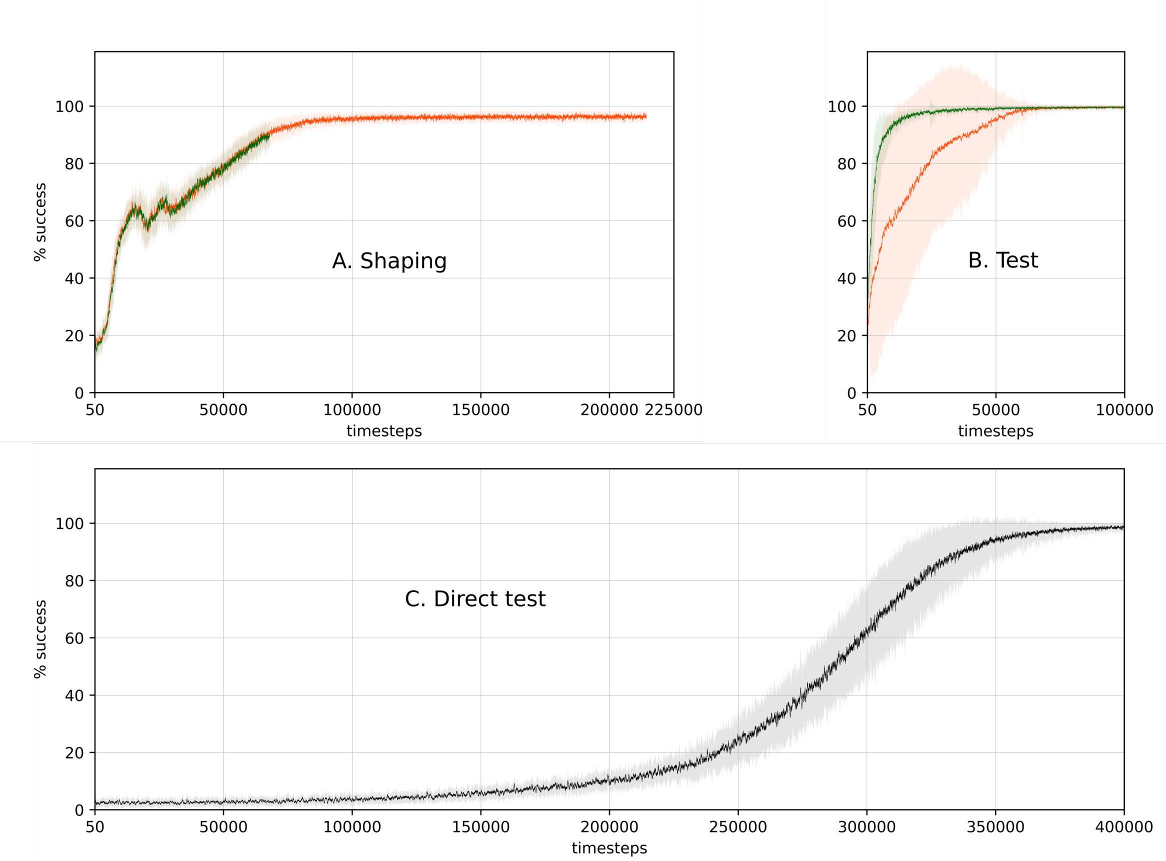


Fig. 4.14 Performance of the flat actor-critic (condition 1 in black, condition 2 in green, condition 3 in orange). All graphs show the success rate for 20 runs, smoothed over 100 time-steps. Graph A and B shows the performance when shaping is used; in condition 2, shaping is stopped as soon as performance is deemed good enough, whereas in condition 3 shaping is set to continue for another 150,000 timesteps. Note how the excess shaping time leads to worse performance on the test, with increased variance. Graph C shows performance if the agent must learn to solve the test problem from scratch. Note that shaping considerably reduces learning time, even when considering both shaping time and learning on the test problem.

In particular, the simple hierarchical model proved unstable (cf. figure 4.17). This could be due to the harsh disturbance of simultaneously having to adjust to the “test” task and to interference from the other repertoire.

The hierarchical model of conditions 4 and 5 could adjust to the final test and solve it, but its performance was not as good as that of the simpler actor-critic, though better than that of a naive agent (with no shaping).

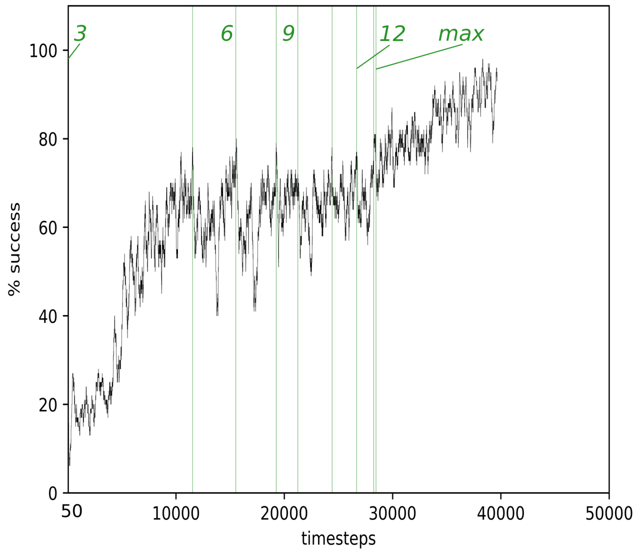


Fig. 4.15 The course of learning for a single agent undergoing shaping (condition 2). The task changes are shown in green - each change consisting in an increase of the maximal initial distance between the box and its target.

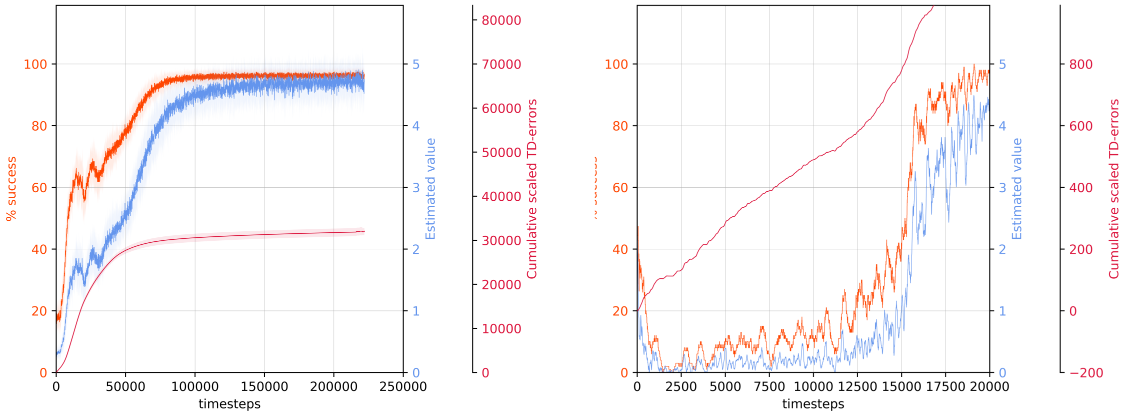
It was expected that using intrinsic rewards to bias learning towards fewer skill switches would produce very quick learning at the hierarchically more abstract level - but contrary to expectations this did not have a large effect (cf. figure 4.17, graph C.).

A summary of the performance achieved by all agents, in all conditions, is presented in table 4.2.

4.5 Discussion

The objective of this chapter (as presented in section 4.1.1) is to test several hypotheses about the suitability of Reinforcement Learning as a model for insight, through simulation. I begin by discussing the adequacy and limitations of the experiment as a simulation of Epstein's own study on live pigeons.

I then discuss whether the hypotheses presented in 4.1.1 are validated. The main hypothesis, the suitability of HRL for modeling insight, is not validated in this experiment (though it is not invalidated either). Other relevant features of RL in the context of modeling insight are validated: the ability to solve insight problems, positive and negative transfer, and the tracking of expected value.



(a) (b)

Fig. 4.16 **(a)** Success rate (orange), estimated value (blue), and cumulative temporal difference error (red) during the shaping phase (average and standard deviation over 20 runs, condition 2). Cumulative temporal difference error is calculated as $\frac{\delta^3}{|\delta|}$ (that is, a signed version of the squared error) because learning minimizes the squared, rather than absolute, temporal difference error.
(b) A particularly striking "insightful" learning curve on the test (condition 2). Note that there is a prolonged "impasse" phase during which the agent makes little headway; followed by a rapid "insight" phase during which the agent suddenly experiences a higher rate of success, a drastic rise in subjective expected value, and an increased average value of temporal difference errors.

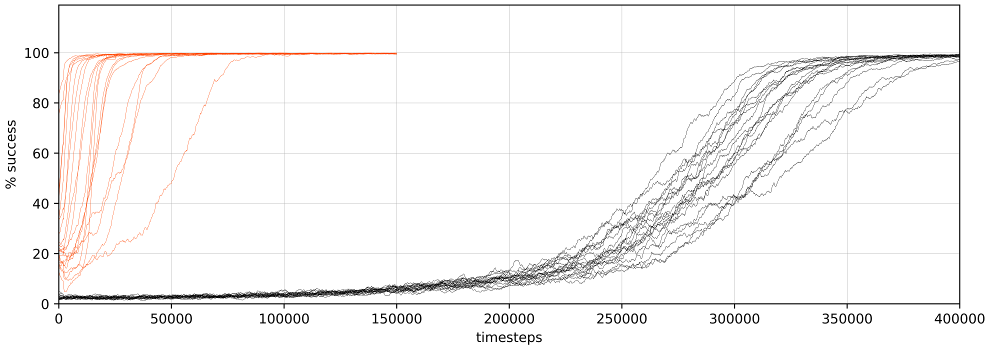


Fig. 4.17 All condition 3 curves (orange, left), compared to condition 1 (black, right), over 20 runs. (All curves have been smoothed for readability, showing the average over 4000 timesteps.)

4.5.1 Comparison between artificial and real pigeons

This experiment is motivated primarily by analogies between theories of animal learning and reinforcement learning techniques. Therefore a key consideration is the extent to which these

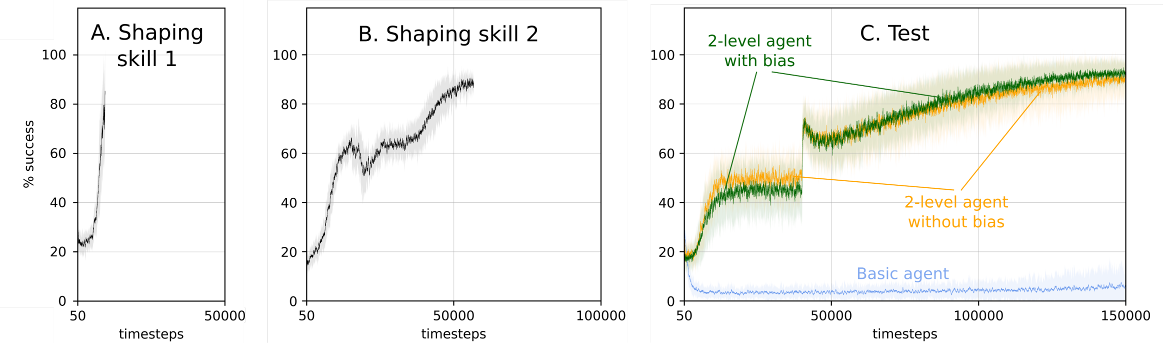


Fig. 4.18 Performance of hierarchical agents (in yellow, condition 4, in green, condition 5; shaping, in black, is identical for both conditions). All graphs show the success rate for 20 runs, smoothed over 100 time-steps. On graph C., note the performance spike at t=40,000, corresponding to allowing learning to take place at both levels simultaneously. The basic agent suffered from stability issues. Introducing a bias against switches between the two skills led to slightly worse performance when learning exclusively at the top-level, and later to slightly better performance when learning at both levels.

apparent similarities held up in the experiment. In at least two ways, the simulation reflected findings on actual pigeons:

1. The shaping techniques used on pigeons also proved highly effective for the reinforcement learning agents, leading to quick learning of the behavioral repertoires.
2. Transfer from the learned repertoires to the test task was successful, yielding fast learning on the test task.

However, the deep Reinforcement Learning agents differed in important ways, the most salient being the speed of learning. Whereas the pigeons learned from a relatively limited amount of experience, and solved the final problem in only a few minutes, the reinforcement learning agents required considerably more experience. This is despite the reinforcement learning agents facing a simplified problem. Nevertheless, it must also be stressed that the response of the real pigeons' is shaped not only by the training they have received, but also by all of their prior life experience.

4.5.2 Hierarchical RL was not insightful

In chapter 4, I conjectured that hierarchical reinforcement learning might underlie sudden insights. By offering discrete choices between extended strategies and associated representations, HRL allows for explorative jumps which, if successful, would resemble insight.

Condition	**Average time to 90% success:** **Shaping**	 **Test**
Real pigeons	~28 hours[a]	~2.5 minutes[a]
Condition 1 *(flat agent, no shaping)*	N/A	337
Condition 2 *(flat agent, shaping)*	68	7
Condition 3 *(flat agent, extra shaping)*	67[b]	39
Condition 4 *(HRL agent)*	67[c]	150
Condition 5 *(HRL agent, duration bias)*	67[c]	124

a: Estimates based on Epstein (1985) and Cook and Fowler (2014).

b: Shaping continued for 150,000 timesteps beyond 90% success.

c: Total of the timesteps for training both options (10 + 57).

Table 4.2 Shaping and test performance of artificial pigeons using different models and under different conditions, and of real pigeons. For all simulated pigeons, "times" are expressed in thousands of timesteps.

Furthermore, such exploratory jumps could in principle link strategy change with a change in representation (restructuring) as well as an “Aha!”-moment, in the form of a positive temporal difference error. Does this happen when an HRL algorithm is implemented and tested on an insight problem? The main objective of the work described in this chapter, as stated in section 4.1.1, was to verify this, using a well-established insight problem from the psychology literature. But the results do not appear to support this theory: the “flat” reinforcement learning agent outperformed the HRL agent; the HRL agent had to be constrained in order to preserve the stability of learning. In this subsection I discuss this result.

It is useful to summarize the task of the HRL agents. To solve the test problem, these agents had to:

1. Learn two skills, pushing the box and jumping on top of it;
2. Chain the two “skills” (discovering a policy over options);

3. Generalize from the spot to the banana (adjust the “push-to-spot” skill).

Learning the two skills was efficient. In the case of the push-to-spot task, the progressive shaping program allowed for quick learning and achieved reliably high performance; in that respect the simulation was a success.

Learning the policy over options was more difficult. In the original experiment, the training is episodic: each skill is trained separately. As a result, in the simulation there aren’t any opportunities to learn switching between skills; the agents therefore cannot learn a policy over options from the shaping phase. This policy over options must be discovered on-the-spot during the test. However, neural networks are subject to catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990) when encountering sudden changes in the training distribution. The transition from shaping to test is one such change; a switch between two options is also an abrupt change. This increased variance in updates presumably caused the skills to be “catastrophically forgotten”. To account for this, I introduced an adjustment phase during which the policy over options is learned without modifying the options. Unfortunately, during that phase it was not possible to adjust the learned options to the test task.

When learning is resumed at all hierarchical levels, one observes an abrupt jump in performance (cf. figure 4.18), plausibly corresponding to improvements made possible by the adjustment of the learned options. Nevertheless, after this jump learning resumes at a moderate pace, whereas flat reinforcement learning achieves high performance quickly (cf. figure 4.14). Overall, the behavior of the HRL agents had less in common with insight than that of flat RL agents (which displayed very accelerated learning on the test phase).

It is not clear whether additional engineering effort could resolve these difficulties. There is a vast range of techniques available to improve learning performance and reduce the risk of catastrophic interference/forgetting. For instance, learning at multiple time-scales (Smith et al., 2006) would correspond to the explanation in terms of fast learning during the test proposed by Epstein (2014). Alternatively, a Bayesian approach might protect against interference by keeping track of the certainty associated with the different parameters of the learned model (Ghavamzadeh et al., 2015). Grafting these techniques on the already complex HRL model, however, would amount to a multiplication of moving parts and hyperparameters, making the model increasingly unwieldy and inelegant as an explanation for insightful behavior.

In summary: the multi-level hierarchical agent faced instability issues, partly owing to its partition in two levels (options and policy over option), one of which could not be learned during shaping. The introduction of ad-hoc solutions to these issues (e.g. “freezing” the options until a stable policy over options was learned) made the algorithm more complex and less like psychological insight. Particularly, freezing one level of learning made it impossible to achieve good performance using the other, thus preventing the sudden discovery of a strategy that solves the problem. Not freezing this level led to instability.

4.5.3 Insight-like features of flat RL agents

In the rest of this section, I discuss insight-like features in the behavior of the flat agents.

Prior experience and transfer

In chapter 2, we have seen that a well-established aspect of insight is its sensitivity to prior experience. Some relevant prior experience is necessary for solving insight problems; too much experience may be detrimental (Birch, 1945; Wiley, 1998). A core reason for choosing Reinforcement Learning to model insight was to take into consideration the role of experience in insight problem solving.

This experiment further confirmed a vast number of results about the efficiency of transfer in reinforcement learning (Taylor and Stone, 2009): in all conditions in which they receive prior training, artificial pigeons solved the test task much faster than the naive agents of condition 1 (approximately 7,000 steps to reach 90% success for condition 2, compared to about 337,000 for condition 1). This makes the resolution of problems by trained agents very rapid, comparatively to naive agents: indeed, when viewed at the same time-scale, the increase in performance of condition 2 agents on the final test (green curve in figure 4.14) looks sudden compared to that of condition 1 agents (black curve).

Many insights present a characteristic lack of progress in the initial effort at solving the problem, which in the literature has often been discussed (Bilalić et al., 2008; Ohlsson, 1992) in relation with the Einstellung effect (Luchins, 1942). From a learning perspective, Einstellung is related to negative transfer. Condition 3, in which agents are made to overfit the training examples, was designed to elicit this effect. This has resulted in learning curves that are often sigmoidal, whereas condition 1 curves lack the flat initial phase of the sigmoid; furthermore, when one looks at individual learning curves 4.17, the sigmoid curve can sometimes be very steep.

Machine learning is concerned primarily with using existing data or prior experience in order to successfully process new exemplars or situations; thus I have argued in chapter 3 that it is well-positioned to study the complex role of prior experience (sometimes helpful, sometimes hindering the discovery of the solution) in insight problem-solving. These results confirm that a machine learning account of insight can help explain, even with "out-of-the-box" algorithms, the characteristic search-illumination-solution pattern of insight, including the Einstellung and fixation effects. Indeed, machine learning concepts such as transfer, negative transfer, and overfitting, which until now have not received substantial attention in the insight literature, appear to be closely related to these effects.

Value and reward

The concept of value plays a central role in Reinforcement Learning (see 3, especially section 3.1.3). It is also central in the present attempt at modeling insight on the basis of reinforcement learning.

The experiment allows for an illustration of a basic principle of reinforcement learning - as a successful strategy is implemented and meets with success, expected return – a subjective measure of progress – rapidly increases. However, this could not be attributed to a single large temporal difference error, but instead to an accumulation of smaller errors, both during the "impasse" phase and when the impasse is resolved (see figure 4.16). This suggests, firstly, that errors encountered even as no new progress is made may be an important first step towards enabling the discovery of actions that actually work. Secondly, it shows that the flat reinforcement learning agents did not experience a single large temporal difference error resembling an Aha!-moment, but instead a large number of small errors which, in interaction with the environment, allowed them to eventually solve the problem. Thirdly, there is indeed a rapid increase in the agent's subjective estimate of value as it solves the problem, corresponding to the "feeling of warmth" measured by Metcalfe and Wiebe (1987).

4.5.4 Summary: reinforcement learning and insight

Are the agents undergoing a phenomenon resembling "insight"? In chapter 2 I conducted a review of the insight literature, concluding with a list of consensus features of insight, which I summarize again here:

1. The insight sequence
2. Discontinuous change in "feeling-of-warmth" ratings
3. Restructuring (changes in representations, heuristics...)
4. Dependence on previous experience
5. The role of sleep as a facilitator for insight
6. The role of attention and executive control networks

Despite the lack of sudden discontinuity (particularly in the HRL case), there are several correspondences between psychological insight and the behavior of the systems tested in the present experiment. These similarities are made explicit in table 4.3. The table not only suggests that the artificial pigeons has captured some key properties of insight, it also suggests new interpretations for the phenomena observed in psychological experiments.

Biological insight	**Artificial Pigeons**
Search-Impasse-Restructuring-Verification (Ohlsson, 2011; Weisberg, 2015)	Impasse-Rapid learning-Slow improvement
Sudden change in feeling-of-warmth ratings (Metcalfe and Wiebe, 1987)	Rapid increase in subjective expected value
Changes in chunking (Knoblich et al., 2001), heuristics (Kaplan and Simon, 1990), constraints (MacGregor et al., 2001)	Changes in network weights underlying both representations and behavior
"Just right" amount of experience allows for best performance (Wiley, 1998)	"Just right" amount of experience allows for best performance

Table 4.3 Correspondences with some of the characteristics of biological insight.

About the *insight sequence* (first row in table 4.3), the following seems plausible: impasse is not unproductive, but rather corresponds to learning from failure; whereas the rapid learning of "insight" corresponds to learning from success. The verification phase might also be characterized as a slow improvement phase, during which the sighted solution is refined in a process resembling "simple" search (e.g. via dithering) within the representational context of the current policy.

About the sudden change in feeling of warmth, the present experiment suggests that insight can be in some sense incremental, similar to non-insight problem-solving (by contrast with Metcalfe and Wiebe (1987)), but in a very accelerated manner, made possible by prior learning, which looks like a step-function only by comparison with an otherwise slow learning process.

Finally, the experiment suggests that restructuring (changes in representations/chunking, heuristics, and constraints) might be the outward signs of an internal structure in which these categories (chunking, heuristics, constraints) have little relevance. In the case of this experiment, this structure is a neural network which encodes all of this based on the weights associated with its connections; but of course other statistical learning models might display similar behavior despite using different structures.

Could insight be “merely” accelerated learning? Although biological brains are computationally powerful due to their massively parallel nature, they are also slow to propagate information. Because of this, the kind of accelerated learning seen in the artificial pigeons, which involves thousands of iterations and tests, seems biologically implausible. I believe something is still missing from this flat actor-critic model of insight: restructuring events and “Aha!”-moments that are not merely *accelerated* learning, but are truly *sudden*, in the manner of a step function.

4.6 Conclusion

4.6.1 Summary

This chapter sought to model “insight in the pigeon” - inspired by a classical experiment aiming to show insight could be explained by fundamental associative learning mechanisms. The experimental set-up was simulated, and several reinforcement learning agents were subjected to a shaping procedure closely modeled on that undergone by the actual pigeons. These agents were designed to instantiate the concepts used in the psychological descriptions of the behavior of the pigeons: “repertoires”, “conditioned responses”, and so on.

Some artificial pigeons were designed to model all of behavior as part of a single neural network (the actor network), whereas others were designed to separate behavior into distinct “repertoires” which could later be chained. Contrary to expectations, the first kind performed substantially better, and had more “insight-like” learning curves. In particular, these agents:

- Succeeded at solving the test task;
- Displayed a very steep learning curve during the test task (conditions 1 and 2), compared to trying to solve that task from scratch (condition 3);
- Performed better with an appropriate amount of experience (condition 1), as opposed to no experience (condition 3) or too much experience (condition 2). This corresponds to findings for insight on humans and animals (Birch, 1945; Wiley, 1998).
- Had a fast-rising subjective estimate of value, as computed by the critic, thus resembling the reports of sudden increases in “feelings-of-warmth” as measured in Metcalfe and Wiebe (1987).

Nevertheless, the simulated pigeons were much less data efficient, and their faster learning during insight was still seems very slow in comparison to insight as it occurs in human beings and animals. Furthermore, whereas biological insight occurs *prior to testing the solution*, in this experiment the agent was continuously interacting with the task even as the "insight" took place. The simulated pigeons did not experience sudden insights resembling step-functions, but instead accelerated learning over thousands of time-steps of trial and error.

Although the main hypothesis motivating this chapter (the suitability of HRL as a model for insight) could not be validated, other hypotheses were corroborated. Particularly, the artificial pigeons could solve a difficult insight problem; the role of experience in the simulated pigeons corresponded to the effects of experience for biological insight-problem solvers; and the artificial pigeons' subjective estimate of value seemed to correspond with the "feeling of warmth" ratings measured by Metcalfe and Wiebe (1987).

This experiment suggests that there is something right about the reinforcement learning picture of insight, but that something is still missing. The hierarchical architecture has promising properties; but it also has limitations. Tentative solutions to these limitations are discussed in the next chapter.

4.6.2 Contributions

- A hierarchical reinforcement learning model, adapted from the the option-critic architecture, with negative results by comparison with "flat" RL. The model made use of two techniques which, to my knowledge, are novel: the use of the controller-over-options as simultaneous critic for the options, and Sarsa controller; and the method for biasing the network against option-switches.

- Experimental evidence in favor of a flat reinforcement learning model of insight on an adapted version of an animal insight experiment.

4.6.3 Bibliographical remarks

- **Shaping and chaining (Psychology):** Shaping is just the application of operant conditioning techniques to behavioural change – as a result it is perhaps more a subject of applications than of fundamental theory. Shaping and chaining have found applications in techniques such as applied behavior analysis, used in the treatment of autism for example (Fisher et al., 2011). Illuminating discussions of the theoretical foundations of shaping and "learning by consequences" can be found in the work of Skinner (1975, 1981). The term "chaining" is used in the behaviorist research programme in a non-technical, ambiguous way; it can refer to chaining reflex responses or temporally extended ones.

- **Shaping and chaining (AI):** Shaping and chaining have been studied and implemented variously in AI settings. Notable implementations include a symbolic RL approach of shaping and chaining by Touretzky and Saksida (1997) (which is explicit about psychological inspirations) and an HRL approach by Konidaris and Barto (2009) (which isn't). In a seminal paper, Ng et al. (1999) demonstrates how to use shaping rewards without affecting the optimal solution; however the method of supplying "guiding" rewards has largely been supplanted by the equivalent technique of using a learned initial value function (Wiewiora, 2003), as is done for instance in Singh et al. (2010). For a more in-depth discussion of the nature of shaping in Reinforcement Learning, readers are encouraged to consider the insightful article by Erez and Smart (2008).

- **Transfer (Psychology):** "Transfer" is not a technical term in psychology. Various other concepts take its place, most notably, in the context of insight problem-solving, the concept of *mental set/Einstellung* (Luchins, 1942; Wiley, 1998). Interestingly, "mental set" is generally used for positive transfer and "Einstellung" for negative transfer (as in the expression "Einstellung effect"); this likely reflects the conflicting approaches of (American) behaviorist and (German) gestalt psychologists. As in AI (see below), the idea of transfer is related to too wide a swath of psychology for a comprehensive review of connected sub-disciplines - for instance, much of developmental psychology is directly or indirectly related to the concept of transfer.

- **Transfer (AI):** The hierarchical learning algorithm proposed in this chapter is properly speaking a transfer algorithm: it transfers relevant learning from several source tasks, to a target task. There is a rich literature on transfer in both machine learning writ large and reinforcement learning specifically. For the reinforcement learning setting, Taylor and Stone (2009) offer a review of work up to 2009, whereas Lazaric (2012) offers an illuminating taxonomy of transfer settings and techniques. Also relevant is the notion of multi-task learning (Tanaka and Yamamura, 2003). These various settings (transfer, multitask learning), however, are somewhat artificial/unrealistic discretizations of a more general setting: reinforcement learning in changing MDPs. This setting, which edges closer to the full problem of artificial intelligence, can be approached for instance from the angle of temporal coherence (Koop, 2008) and from the angle of learning to learn/lifelong learning (Thrun, 1998). A proper review of learning to learn techniques would require considerably more space than is available here - for instance, techniques as simple and omnipresent as momentum in gradient descent could be understood as "learning to learn".

Chapter 5

The Actor-Critic-Intention Architecture

In the previous chapter, I used existing techniques (shaping, transfer, HRL) from the RL literature in order to produce effects similar to insight. Some of these techniques functioned as expected (shaping, transfer) for a model of insight. The artificial pigeons solved the test task faster than if they were learning from scratch; they did best when prior experience was neither too much, nor too little; their success was accompanied by an increase in the value function. However, their learning lacked the immediacy of psychological insight: though they succeeded much faster than when learning from scratch, they still necessitated a considerable amount of trial and error.

In this chapter, the issue of temporally extended exploration is studied in greater detail, with the aim of finding a new algorithm that resembles insight more closely than existing techniques. A theory of intentions as a basis for insight is laid out, which allows for a closer model of insight than the transfer and overfitting phenomena observed in chapter 4.

This chapter begins in section 5.1 with an extended discussion of the twin challenges of temporally extended exploration and exploration of representations. Section 5.2 presents an architecture which tackles both challenges, called the Actor-Critic-Intention (ACI) architecture. Section 5.3 discusses the manner in which this architecture can reproduce the characteristic properties of insight.

5.1 Background and motivation

With regards to exploration, insight is characterized by two traits. Insight seems to involve the simultaneous discovery of:

1. a new course of action or strategy, extending from the present state to the solution: temporally extended exploration.

2. a new representation of the problem, changing the perception of which features characterize the problem, and how they function together to afford a certain solution.

In this section I discuss these two challenges in turn, reviewing or proposing tentative solutions to each of them, and explaining why they fall short of the objective. In the next section, I propose the Actor-Critic-Intention architecture, which is capable of both exploratory traits.

5.1.1 Temporally extended exploration

The problem of temporally extended exploration (sometimes called deep exploration, e.g. by Osband et al. (2016)), is perhaps best presented using a simple simulation: a biased random walk. Imagine an agent walking on a line; the agent can only go east, or west. The best policy discovered so far is to go east. What is the probability of the agent ever finding out what lies at the end of the western path? This depends on the manner in which the agent explores. If the agent is ε-greedy, we can easily simulate the corresponding Markov process. The list below shows how many agents, out of a million, reach a distance of at least n west-wards:

- $n \geq 0$: 1,000,000
- $n \geq 1$: 110,902
- $n \geq 4$: 150
- $n \geq 7$: 0

The take-away of this simple experiment is that ε-greedy exploration, for all practical purposes, is not capable of exploring far into the unknown, so long as it remains possible to simply turn around and get back to the safety of a more usual trajectory.

Because this effect is dependent on the scale of the actions used (smaller actions mean greater distances, in terms of the number of exploratory steps), this causes RL algorithms to depend on the "scale" of the actions used. Either one uses finer, more precise actions, which enable a finely tuned policy; or one uses larger or more complex actions, which enable fast exploration. The efficiency of the algorithm, at any given task, depends on the programmer's ability to pick the right trade-off between precision and exploration (McGovern and Sutton, 1998); and/or of hand-coding appropriate high-level actions. For instance, in the simulation MDP of chapter 4, the artificial pigeons were "given" appropriately-sized actions for the task, such as pecking or moving around.

Thus an RL algorithm using ε-greedy exploration will either be efficient at executing its overall strategy, or capable of discovering original strategies. Using an analogy from the historical *age of exploration*, ε-greedy exploration will most likely manage many small optimizations on the eastern route to India, but it will not cross the Atlantic and discover America.

There are several ways to alleviate this issue (for instance, many artificial curiosity techniques seek exhaustive exploration). However, these techniques address only part of the problem: they encourage active discovery, but they do not learn appropriately-sized actions. Ultimately, artificial curiosity techniques still rely to some extent on stochastic exploration - and are thereby affected by the problem of action scale and temporally extended exploration.

Randomized networks

How could we introduce variation to our exploration? Osband et al. (2016, 2017) used several neural networks with different initialization parameters and switched the networks between episodes. In this manner, agents collect diversified experience by making use of different exploratory biases. Continuing the *age of exploration* analogy: west might seem more attractive than east for at least one set of parameters; thus using the experience from multiple different starting biases can allow the agent to see more of its environment.

This approach has three limitations.

1. It does not learn appropriately sized actions; it depends instead on the biases of a set of neural networks. This suggests that the method would not scale as the size of the environment to explore increases: there may be a great variety of "interesting" exploratory biases, but each new bias requires a new network.

2. The change between networks is episodic - it is not clear how to adapt such a method to a non-episodic, continuing setting.

3. The networks might pick up new biases during the initial phases of learning owing to properties of the MDP: for instance, if only east is accessible at the beginning of the experiment, any western bias in the initial parameters might have been erased by the time a western path becomes available.

Successive exploratory steps

An alternative is to encourage temporally extended exploration understood in a literal manner, by increasing the chance of temporally extended exploratory sequences. It is possible to achieve this while insuring that only a certain percentage of actions are exploratory, in the manner of ε-greedy exploration.

First, let us compute, as a baseline, the proportion of its time that an ε-greedy agent spends in "deep" exploration. Define the depth of exploration as the length of the uninterrupted sequence of explorative actions. Call e_t the length of the current exploratory sequence at time-step t (1 if the agent is exploring for just one action, 2 if the agent is on its secondary consecutive exploratory action, and so on). On a random time-step t, the probability that the agent is on its n^{th} consecutive explorative action is given by:

$$\Pr(e_t = n) = \begin{cases} 1-\varepsilon & \text{for } n = 0 \\ \Pr(e_t = n-1) \times \varepsilon & \text{for } n > 0 \end{cases}$$

$$\text{By recurrence, } \Pr(e_t = n) = (1-\varepsilon)\varepsilon^n \tag{5.1}$$

Thus for ε-greedy, $\Pr(n)$ is a geometrically converging sequence as n becomes large, with convergence rate ε. Increasing ε slows down the convergence (at the cost of exploiting less often); but a linear convergence rate remains very rapid for even unreasonably large ε (cf. figure 5.1).

One may increase the proportion of temporally extended exploration as follows: let the probability of exploration depend on the current length of the exploratory sequence. Of course, the agent ought to avoid getting stuck in exploration; ideally exploration should still be limited to a proportion ε of all actions. One way to achieve this is the following[1]:

$$\Pr(\text{exploration on } n^{\text{th}} \text{ action}) = \begin{cases} \dfrac{\varepsilon}{\zeta(\upsilon)(1-\varepsilon)} & \text{for } n = 0 \\ \left(\dfrac{n+1}{n}\right)^{-\upsilon} & \text{for } n > 0 \end{cases} \tag{5.2}$$

where $\zeta(\upsilon) = \sum_{n=0}^{\infty} \frac{1}{n^{\upsilon}}$ converges[2] for any $\upsilon > 1$. The constant υ controls the relative importance of deep versus shallow exploration (small υ allows for deeper exploration).

This change to ε-greedy action selection results in the agent exploring a proportion ε of the time, as in ε-greedy; but exploration is now distributed in a manner that favors temporally extended exploration at the expense of shallow exploration. This is shown in figure 5.1. More experimental results are presented in appendix E, section E.4.1.

However, such exploration is by nature purely random. For instance, in the East vs. West MDP, it encourages long random walks, rather than directed sequences of westwards

[1] Other converging series can also be used.

[2] This is called the Riemann zeta function; most scientific software contain tools for computing this efficiently for any υ.

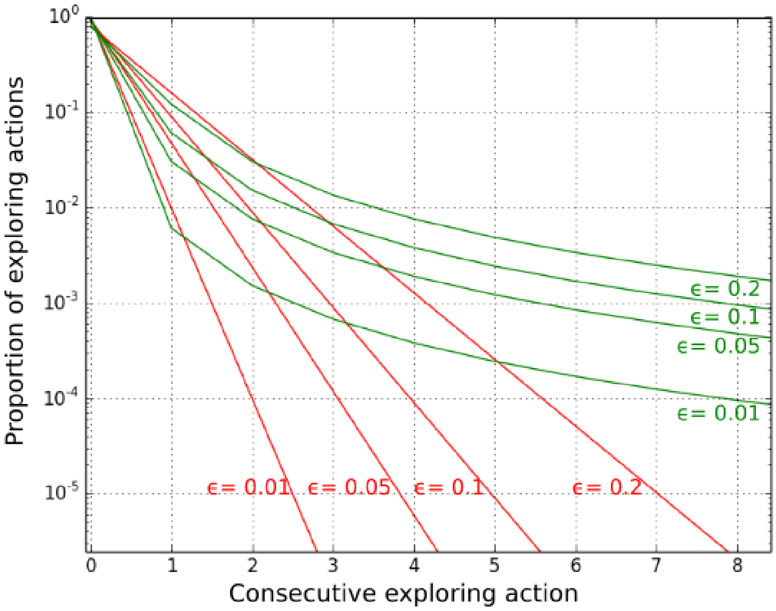


Fig. 5.1 How much time does an agent spend exploring, and at what depth? ε-greedy exploration (red, straight lines) explores superficially: the probability of a long exploratory sequence is very low. For instance, at $\varepsilon = 0.01$, only a proportion 10^{-4} of actions are exploring at depth 2; this becomes 10^{-6} at depth 3, and 10^{-16} at depth 8 (outside the graph). In contrast $\varepsilon - \upsilon$ greedy exploration (green) trades off short exploratory sequences for long exploratory sequences. It explores approximately a proportion 10^{-4} of its actions at depth 8, at $\varepsilon = 0.01$. On all $\varepsilon - \upsilon$ curves (green), $\upsilon = 2$.

exploratory steps. Furthermore, the choice of the atomic actions may bias the random walk: for instance, if there are two actions that go east and only one that goes west, a random walk based on uniformly selecting actions would have an eastern bias. Thus while this method offers a computationally cheap way to increase temporally extended exploration for off-policy reinforcement learning algorithms, its effectiveness is limited for real-world problems.

Summary

Temporally extended exploration ought not only to produce “long” exploratory sequences; but these sequences must also:

- Have the correct length: different MDPs demand different lengths of exploratory sequences. The deep exploration of Osband et al. (2016) explores only at the length of a full episode; the $\varepsilon - \upsilon$ exploration presented above is controlled by the parameter υ. Neither method discovers the correct scale for exploration based on interaction with the MDP itself.

- Be directed: explorative sequences should correspond to tentatively adopting a bias in action-selection. The $\varepsilon - \upsilon$ exploration replaces the exploitative bias with a random walk, which is itself biased by the choice of atomic actions. In the end, $\varepsilon - \upsilon$ always explores according to the same bias. Osband et al. (2016) explore according to a predefined number of neural networks using different biases, and learning from experience; however, nothing prevents these networks from all converging to a similar bias over time. Neither method learns a variety of interesting exploratory biases.

5.1.2 Exploring representations

Insight does not require only a new temporally extended solution for the problem at hand, but also representational change - a sudden change in the representation of the problem. How may one explore for new representations?

Generate and test

One approach is to generate and test new features, as proposed for instance by Mahmood and Sutton (2013). Starting from an input vector, one can generate an expanded set of representations corresponding to relationships between inputs. For instance, from a binary input vector $\{\phi_0, \phi_1\}$, one may consider the feature set $\{\phi_0,\ \neg\phi_0,\ \phi_0 \wedge \phi_1,\ \phi_0 \vee \phi_1, \cdots\}$ etc. The number of such features grows combinatorially with the size of the binary vector.

The testing part of generate-and-test requires time and evidence. Indeed, testing must be done by first learning how to use the feature in order to choose actions. But insight appears to be swift and abrupt, rather than protracted and iterative: generation and testing seem immediate. Can this be achieved through other means?

Stochastic neural networks

In deep learning, it is common to think of the activation of successive hidden layers as intermediate representations (Bengio et al., 2013a); such that each successive layer corresponds to a more abstract or more task-orientated representation of the input. To discover good representations, then, is to discover useful patterns of activation on the intermediate layers. Should activation be stochastic, we may speak of exploration of representations. There are various kinds of stochastic neural networks, the most popular example being dropout (Srivastava et al., 2014). In this chapter we focus on binary stochastic neurons, sometimes called Bernoulli-logistic neurons.

In a standard logistic neuron, the output of the neuron o is given by:

$$o(p_0, ..., p_i) = \sigma(\sum_i w_i p_i)$$

where σ denotes the sigmoid function, w_i denotes the weights, and p_i denotes the inputs to the neuron. This is shown graphically in figure 5.2, in the “Standard sigmoid neuron” box.

Such deterministic neurons can be augmented with a random number generator producing scalars uniformly on $[0,1]$, and denoted by $\mathcal{U}[0,1]$. One then compares the output r of the random number generator with the activation a of the sigmoid neuron. If $r < a$, the final output is 1; else it is 0. Thus the output of a Bernoulli-logistic neuron is:

$$o(p_0, ..., p_i, r \sim \mathcal{U}[0,1]) = \mathbb{1}_{r<\sigma(\sum_i w_i p_i)}$$

A Bernoulli-logistic neuron is illustrated in figure 5.2.

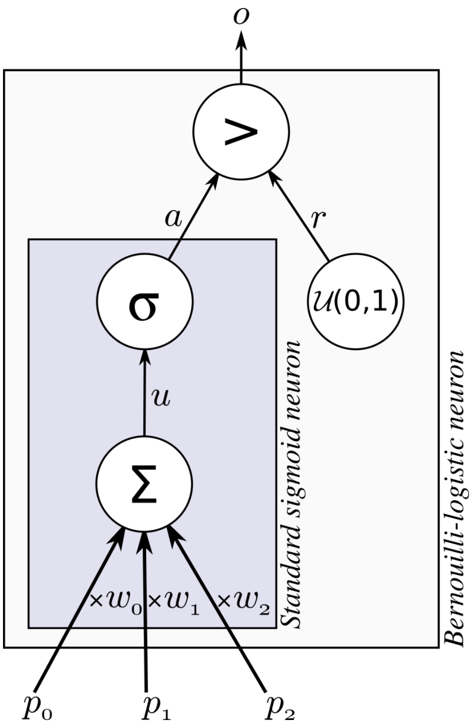


Fig. 5.2 Bernoulli-logistic neuron. For the standard sigmoid neuron, the inputs are denoted p_i, the weights w_i, the weighted-sum of the inputs u, and the sigmoid activation $a = \sigma(u)$. This is further processed by comparing a with a random number $0 \leq r \leq 1$. The output o is binary, either 0 or 1, depending on the result of the comparison.

When integrated into a neural network, stochastic neurons can be said to explore for representations. If each neuron encodes the presence or absence of a feature, then the neuron's activation or lack thereof tests whether it helps to consider that feature as present, or absent. Each neuron's trial-and-error activity can result in different patterns of activation in subsequent layers, leading up to the choice of output. This effect is illustrated in figure 5.3.

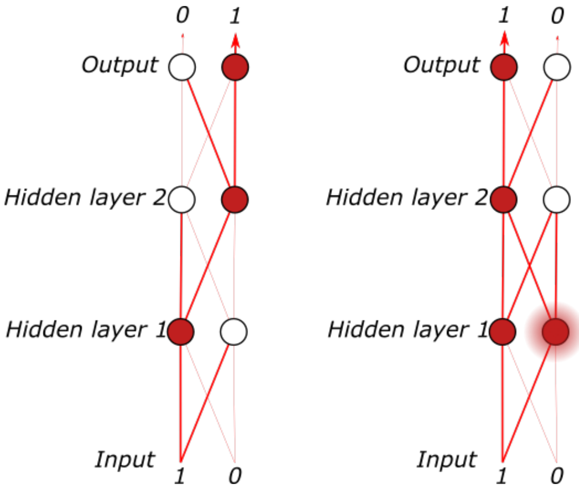


Fig. 5.3 Stochastic network example; neurons filled in red are activated (output 1), whereas neurons filled in white are not activated (output 0). The output connections of activated neurons are shown using bolder lines, in order to represent the flow of information from sensation to action.
Note that even using the same input, different activation patterns may occur due to the use of random number generators. Furthermore, because the activation of each layer depends on the activation at the layer preceding it, change in just one of the randomly generated numbers, on an early hidden layer (in hidden layer 1 in this case), can potentially affect activation on all subsequent layers.

As a result, Bernoulli-logistic neurons can learn to predict *structured output* (Bengio et al., 2013b; Gu et al., 2015) - meaning statistical relationships between the output variables. This is in contrast to deterministic backpropagation neural networks, which by design learn the statistical relationship between the input and the output variables, with the output variables considered independently.

This is an important feature for exploration in reinforcement learning scenarios with a large vector-valued action. For instance, consider the problem of learning to control the 800 muscles of a human body. By construction, a deterministic neural network seeks to independently find the best output for each muscle. But this is inappropriate when the value of the chosen action depends on the pattern of the output. For instance, the scissors-jump and the Fosbury flop require separate patterns; it is not clear that there is a path for gradient descent from one to the other, if each output (muscle) is being treated independently. In contrast, using Bernoulli-logistic neurons allows for learning statistical relationships between outputs, or patterns of action, and allowing them to vary together in the context of exploration[3].

Thus if one interprets hidden layers as intermediary representations, it follows that stochastic networks can undergo something resembling restructuring (or as Ohlsson (2011) puts it, "redistribution of activation"), leading to different intermediary representations, and complex differences in the output. For instance, if a unit corresponding to a certain feature is, by chance, activated, the following layers may encode the situation differently because their own input differed. We may thus have a cascading effect where the unusual interpretation of one detail of the observation can lead to a radically different activation pattern for the whole scene.

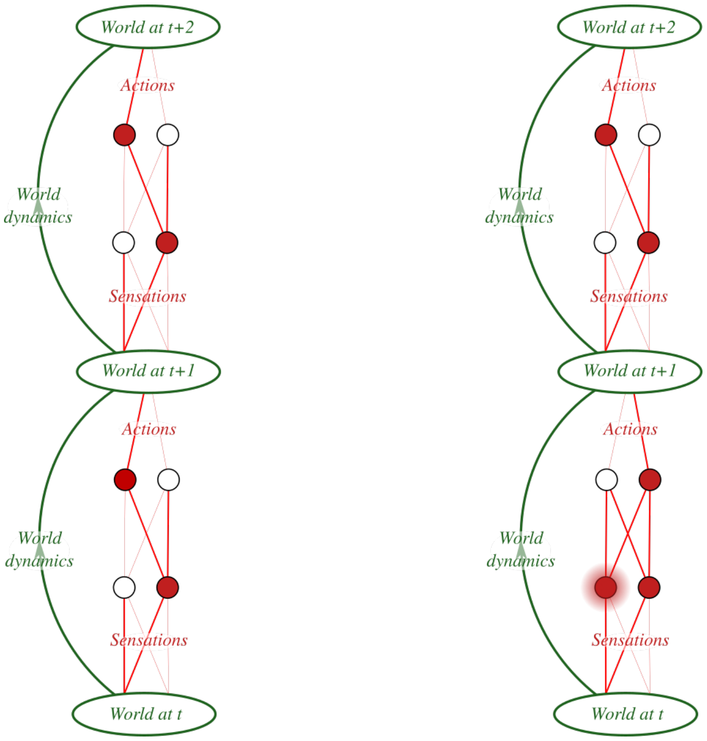


Fig. 5.4 Two actor networks using Bernoulli-logistic neurons (example). Neurons filled in red are activated (output 1). The output connections of activated neurons are shown using bolder lines, in order to represent the flow of information from sensation to action.
The stochastic neurons allow for the trial and error activation of different features in the hidden layers.
However, in the absence of a mechanism to carry changes over time, perception is affected for only one time-step: an exploratory representational change is unlikely to be maintained for a sequence of decisions over time.

[3] In practice, many RL algorithms reduce the action space to a small number of dimensions, or even to a small number of digit actions; there is relatively little work on structured output or large action spaces in RL, although their importance is recognized e.g. in Sutton and Barto (2018), pp. 385, 405, 407.

But the “cascading change” of restructuring made possible by stochastic networks is limited to one time-step. On the subsequent time-step chances are that the features will be interpreted in the most probable manner once again (unless a single different action was sufficient to land the agent in a radically different region of state-space). This is illustrated in figure 5.4.

5.1.3 Options

The previous two subsections have discussed the difficulties associated with both temporally extended exploration, and exploration of representations. Options seem to be capable of both. If an agent uses options whose policies are encoded with neural networks, a switch between two options can induce simultaneously a change of representation (different internal representations), and the exploration of a long-term strategy (a different policy), in a temporally extended fashion. This is why, in the previous chapter, options were investigated as a model for insight.

But options have several limitations:

Parameters: Most implementations of options do not allow for options that take parameters. As a result, two different options would be required even for behaviors that closely resemble each other. Indeed, an infinite number of options would be required to encode behaviors that vary on a continuous spectrum (e.g. executing the same action at different speeds).

Overlapping: Options connect with one another through abrupt switches; options do not “ease” into one-another, nor is it possible for two options to be active simultaneously. For instance, an option to move a robot’s right arm in a certain way could not be active simultaneously with an option that activates a robot’s left leg.

Generality: Option-discovery algorithms lack a rationale for capturing temporal structure about behavior[4], which would make the options “general”: likely to be efficiently re-usable.

I believe these limitations are serious, and may explain the lower performance of options compared to flat-reinforcement observed in chapter 4. They suggest that an alternative approach is needed.

The first challenge is to make options continuous and vector-valued (an option vector influences behavior) rather than discrete and one-hot (only a single option is active at any

[4] Sometimes they discover structure in the MDP, e.g. bottleneck states; however this typically leaves out reward information.

one time). The former method allows for combining the characteristics of multiple options, thus allowing for both overlapping and parameterizability. However the option framework (Sutton et al., 1999) and the SMDP theory were not made to accommodate such extensions.

The second challenge is to make options, in some sense, more “general”. Outside of a bayesian framework, it is difficult to formally demonstrate that an algorithm achieves better generalization than another in principle. However, an argument will be offered that the ACI architecture, presented below, captures temporal structure in a manner that is more general than existing option-discovery algorithms.

5.1.4 Summary

This section has presented a range of ways in which exploration can be conducted in a temporally extended manner, considering techniques for extending exploration in time or for introducing biases in action selection. Two methods for exploration of representations were introduced, illustrating the multiple aspects of the problem of exploration, and the promises as well as the limitations of existing techniques. Finally, I discussed the promise and limitations of the option framework.

In the next section, an architecture is introduced that combines some of the ideas from the techniques introduced here, while seeking to overcome their limitations.

5.2 The Actor-Critic-Intention architecture

The proposed architecture makes use of two existing methods (Bernoulli-logistic units and an actor-critic architecture) plus a novel component of self-prediction. In this section I describe the architecture, beginning by an informal discussion of the main idea (using self-prediction to detect patterns and suppress redundant processing), then gradually showing how the other components of the architecture function together.

5.2.1 Predictor network for self-prediction

To achieve temporally extended exploration, the agent must deviate decisively from its exploitation trajectory over multiple consecutive time-steps. With respect to both representations and actions, temporally extended exploration is difficult because each *exploitative* decision risks undoing what a prior *explorative* decision had achieved.

Examples

For instance, a stochastic network may represent its input differently for one time-step, but this is reversed the next. Likewise, an agent might go west instead of east on one time-step, but on the next time-step returns east, undoing prior explorative progress. In many cases, this reversal occurs due to redundant processing. A reinforcement learning agent re-processes “from scratch” a state, even when its features were predictable based on the previous state and action, to arrive at a predictable action, through a sequence of predictable intermediary stages of processing. There is mutual information between successive decisions, whether one considers the final action, or the intermediary stages of processing. Therefore one may predict how the network processes the information at one time step based on what took place on the previous time-step.

Continuing the analogy with the age of exploration: to a large extent, it is possible to deduce an explorer’s intention based on decisions made early on in the journey: what quantity of supplies was taken along, the harbor of departure, the direction adopted for the first few nautical miles, and so on. Prior decisions predict future ones: one can use regularities in behavior to predict future decisions (e.g. explorers do not take more or less supplies than reasonable with respect to the distance to be covered). If the explorer acts contrary to expectations, it would be most likely because something unexpected has occurred: a storm, a disease, a discovery on the way, rather than due to a reconsideration “from scratch” of the whole enterprise. This contrasts with reinforcement learning algorithms, which treat each step as an independent decision, for which all available information is brought to bear. Most RL agents treat each situation “from scratch”: because of this, extended exploration cannot occur.

Making independent decisions

If the difficulties of temporally extended exploration are caused by *redundant* decision-making, a logical solution is to make successive decisions *independent*. Removing redundancy is equivalent to ensuring that changes to the agent’s course of action should be made solely based on unpredictable, relevant events; predictable events should be ignored. How could this be approximated in a deep learning architecture?

To ensure that decisions are mutually independent, I seek to remove relevant, but redundant information from stochastic decisions. This is achieved by *replacing* the redundant information by the prediction of that information. The activations at each layer of a neural network correspond to an increasingly abstract and relevant representations of the input:

thus I seek to predict the activations within the network at the current time-steps, using information from prior time-steps.

By making predictions of future decisions and actions and using these predictions, in place of making these decisions from scratch at every time-step, an agent avoids the redundant processing of information which restricts temporally extended exploration.

The predictor network

To achieve this, in the ACI architecture the actor network is complemented with a predictor or intention network (the connection between the state of the predictor network and the concept of "intentions" is discussed in more detail in the next chapter). The predictor network learns to predict future decisions. As a result, predictable network activity is predicted by the predictor network, such that the actor only needs to react to unexpected information in order to maximize performance. The structure of a small such network is shown in figure 5.5.

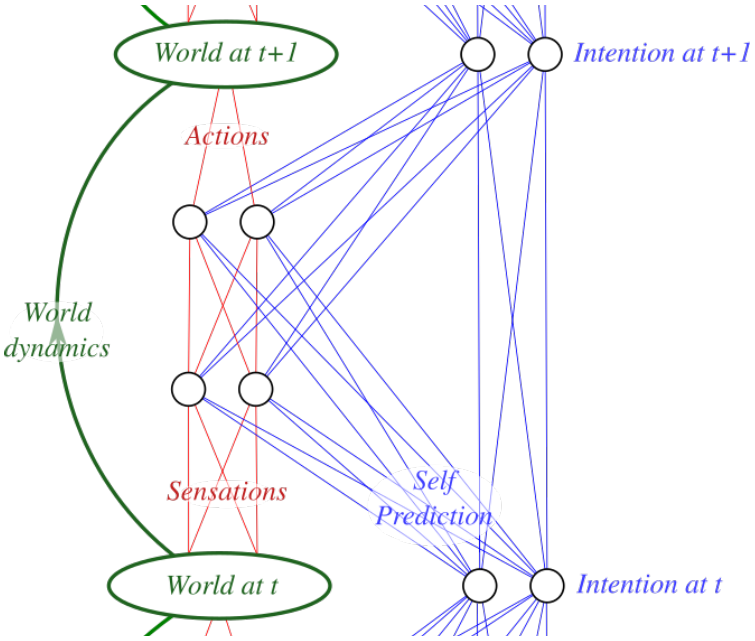


Fig. 5.5 Intention network (actor and predictor). In red, a regular actor network with two hidden layers, interacting with the world via sensations and actions. In green, world dynamics that do not depend (directly) on the agent. In blue, the predictor network, which uses recurrent units to predict activation in the actor network. The predictions (of pre-activation values) are also inputs to the neurons being predicted.

How might we "transfer" the predictable part of decision making from the actor network network (which can learn by reinforcement) to the predictor network? The key idea is that the actor network will automatically adjust, via reinforcement, if the input to its neurons overshoots optimal performance. Recall that, per the architecture proposed in figure 5.5, the predictor network provides inputs to the stochastic actor. If the predictor network correctly predicts the total input to neurons in the actor network, the actor network will learn not to overshoot, by abstaining from providing the relevant input itself. Thus we may want to use an update resembling:

$$w_p \leftarrow w_p + \alpha_p \delta \nabla \sum_n (u_n - \sum p_{p,n} w_{p,n})^2$$

where α_p is a learning rate, $u_{d,n} = \sum p_{d,n} w_{d,n} + \sum p_{p,n} w_{p,n}$ is the preactivation value of the actor network neuron n, based on this neuron's input vectors from decision neurons $p_{d,n}$, and from predictive neurons $p_{d,n}$, multiplied with the corresponding parameters (weights) $w_{d,n}$ and $w_{p,n}$.

However $(u_d - \sum p_p w_p) = \sum p_d w_d$, so this update would not transform the weights of the predictor network. (In other words: because the predictor is predicting a summation that includes its own output, any update would move the target, and thus would not reduce the prediction error.) To deal with this issue, one may use a separate *target network* mimicking the predictor network with weights w_t, as in Double DQN (Van Hasselt et al., 2016). This leads to the following update rule:

$$w_p \leftarrow w_p + \alpha_p \delta \nabla \sum_n (p_{d,n} w_{d,n} + \sum p_{p,n} w_{t,n} - \sum p_{p,n} w_{p,n})^2$$

The network is recurrent; thus this gradient ought to be estimated "through time" in order to learn long-term dependencies. There are several (often compatible) ways to achieve this, the most popular being backpropagation through time (BPTT) using some variation of long-short-term memory (LSTM) units. In the figures (e.g. figure 5.6), only standard neurons with recurrent connections are shown: this is theoretically sufficient for learning to occur.

5.2.2 Actor network for representational exploration

The predictor network can propagate decisions, such that they affect not merely the current action, but many subsequent decisions. This actor network makes use of the Bernoulli-logistic neurons discussed in section 5.1.2 in order to explore for representations.

The weights w_d of the actor network[5] are learned by updating the weights in the direction of increased performance $\nabla J(w_d)$:

$$w_d \leftarrow w_d + \alpha_d \nabla J(w_d)$$

where α_d is a learning rate, and $J(w_d)$ is the return. In a deterministic actor this can be computed as (cf. chapter 3):

$$w_d \leftarrow w_d + \alpha_d \delta \nabla \ln \pi(A_t|S_t, w_d)$$

where δ is the temporal difference error computed by a critic.

For a stochastic actor however, the gradient $\nabla \ln \pi(A_t|S_t, w_d)$ does not exist. Instead, one may use the gradient of the expectation, $\nabla \mathbb{E}[\ln \pi(o|S_t, w_d)]$, where o denotes the probability for each neuron to have produced the output that it did (e.g., if a neuron will output 0 with probability 0.9, but outputs 1, the $o = 0.1$ for that neuron). This results in the following update:

$$w_d \leftarrow w_d + \alpha_d \delta \nabla \ln \pi(o|S_t, w_d)$$

To compute this gradient, one approach is to use REINFORCE (Bengio et al., 2013b; Williams, 1992). The idea is to treat each individual neuron as an individual decision maker seeking to figure out the best policy to maximize a reward signal, based on its input; however the reward signal is shared by the whole network. A consequence of treating each neuron individually is that the update signal has a large variance (is very noisy), because the relationship of each neuron with other neurons (as encoded by weighted connections) is not taken into account, and because the contribution of each neuron to the whole is both small and weakly related to the outcome.

Another approach called "straight-through" (Hinton, 2012) is to backpropagate *as if* the network was not stochastic: treating the neurons as if they were deterministic sigmoid neurons for the purpose of backpropagation. This does not compute the correct gradient, but an approximation of it that has the merit of taking into account the relationship between neurons (whereas using REINFORCE, neurons do not take into account information about their connections). This method considerably outperforms REINFORCE (Bengio et al., 2013b).

In this chapter I combine REINFORCE (for calculating gradients) with straight-through (to propagate them through the network), in order to achieve sufficient performance.

[5] I depart from the traditional notation of using θ for the actor parameters and w for the critic parameters. Since there are now three components with distinct learning rules – actor, predictor, and critic – their respective parameters, or weights, are denoted w_d (for 'decision'), w_p, and w_c.

5.2.3 Critic network for strategy evaluation

In the previous subsection I treated the critic as a given. However, it is no longer possible to use a standard critic with this architecture. This is because the performance of the agent depends not only on the state of the world, but also on the long-time plans of the agent encoded in the current state of the predictor network – its “intention”.

To account for this, and to enable agents to learn to act based on their intentions, the critic must receive the agent’s intentions as an additional input. Thus it estimates expected return based on a state of the world that includes both the internal state of the agent, and the external state of the environment - $\hat{v}(s,i)$ instead of $\hat{v}(s)$.

With this minor modification, the critic parameters are updated as in chapter 3 section 3.2.2, by using the temporal difference error, with the sole addition of intentions:

$$w_c \leftarrow w_c + \alpha \delta_t \nabla \hat{v}(S_t, I_t, w_c) \tag{5.3}$$

5.2.4 Summary

This section presented the ACI architecture. It consists of three components: a decision network or actor, an evaluation network or critic, and a prediction/intention network or predictor.

The central idea is the use of prediction of one’s own future “decisions” to perform temporal abstraction. This ensures that explorative decisions are made exclusively based on new and relevant information: the predictor network learns to predict everything else.

Another important component is the Bernoulli-logistic network used for the actor. This enables the actor to explore representations, while also allowing it to learn to produce structured outputs, which is important e.g. for intentions and actions to be consistent with one another.

The algorithm below makes use of existing methods to achieve a fairly efficient version of the architecture, at the cost of following a biased estimate of the gradient. Ongoing improvements in structured prediction and generative networks may allow for performance increase.

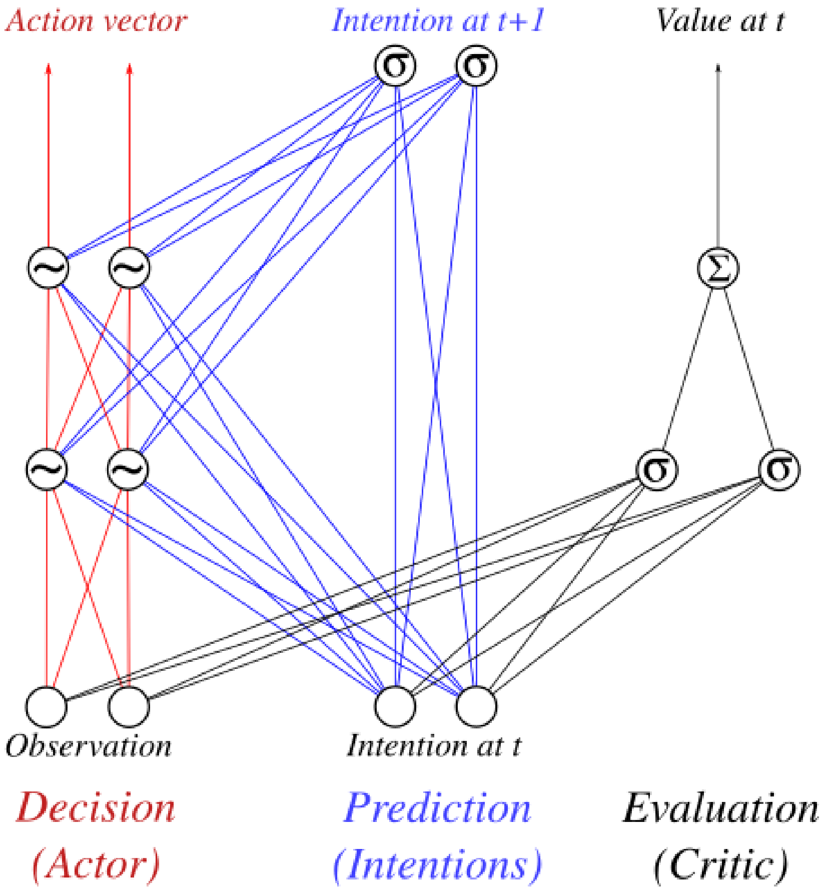


Fig. 5.6 The full ACI network architecture (using a small number of units). Bernoulli-logistic units are denoted with a "∼", linear units (without an activation function) with a "Σ", and non-linear units (using for instance a sigmoidal or ReLU activation function) with a "σ".

Algorithm 7: Intention architecture - algorithm

```
Input      : A network set up according to the architecture shown in figure 5.6. The
             policy function $\pi$ corresponds to feedforward along the predictive and
             actor parts of the network. whereas the value function $\hat{v}$ corresponds to
             feedforward along the critic part of the network. In addition to $s'$ and $i'$,
             $\pi(s,i)$ outputs $d$ (the preactivation value to decision neurons from other
             decision neurons), $o$ (the activated/output value of decision neurons), $p$
             (the preactivation value to decision neurons from predictive neurons),
             and $t$ (the corresponding outputs of the target network).
Input      : An initial state and an initial intention.
Require    : Interaction with a (potentially continuing) MDP using functions
             observe(), perform(action)
Parameters : Step sizes $\alpha_d, \alpha_p, \alpha_c \in (0,1]$, and randomized initial parameters
             $w_d, w_p, w_c$, for the decision (actor), prediction/intention (predictor) ,
             and evaluation (critic) networks respectively; plus step-size $\alpha_t$ for the
             target network.
// Initialize action, intention, and target network:
$\mathsf{a}, \mathsf{i}' \leftarrow \pi(\mathsf{s}, \mathsf{i})$
$\mathsf{i} \leftarrow \mathsf{i}'$
$w_t \leftarrow w_p$
// Learning:
repeat
    $\mathsf{r}, \mathsf{s}' \leftarrow$ perform(a)                     // Perform action and observe consequences
    $\mathsf{a}, \mathsf{i}', \mathsf{d}, o, \mathsf{t}, \mathsf{p} \leftarrow \pi(\mathsf{s}', \mathsf{i})$                                   // Make new decision
    $\delta \leftarrow \mathsf{r} + \hat{v}(\mathsf{s}', \mathsf{i}') - \hat{v}(\mathsf{s}, \mathsf{i})$                                // Compute TD error
    $w_c \leftarrow w_c - \alpha_c \delta \nabla \hat{v}(\mathsf{s}, w_c)$                                     // Update critic
    $w_d \leftarrow w_d + \alpha_d \delta \nabla \mathbb{E}[\log \pi(o|\mathsf{s}, \mathsf{i})]$                              // Update actor
    $w_p \leftarrow w_p + \alpha_p \nabla \sum (\mathsf{d} + \mathsf{t} - \mathsf{p})^2$   // Update predictor.  Using BPTT here allows
      for finding long-term dependencies.
    $w_t \leftarrow w_t + \alpha_t (w_p - w_t)$                                        // Update target
    $\mathsf{s} \leftarrow \mathsf{s}'$
    $\mathsf{a} \leftarrow \mathsf{a}'$
    $\mathsf{i} \leftarrow \mathsf{i}'$
until Forever
```

In this section I have focused mostly on describing the ACI architecture. Although I have mentioned insight in several places, the focus was on understanding how intentions relate to exploration and temporal abstraction, concepts from the field of Reinforcement Learning. What was not discussed, so far, is whether or not the ACI architecture is a good model of insight, and why.

5.3 The ACI architecture as a model of insight

The ACI architecture was designed to emulate psychological insight; in this subsection I discuss the manner in which it does so. There are three key capabilities of the model which make it, in principle, capable of “insight”. These are:

- Sudden discovery of a complete solution (a single exploratory step discovers what may be called a strategy, including multiple steps, contingency plans, etc.),
- Restructuring (the representation of the problem, its important features, changes radically and for the duration of the solution),
- “Aha!”-moment (the agent experiences a sudden subjective leap in its “feelings-of-warmth”, i.e. a subjective rating of its closeness to the solution, as first operationalized by Metcalfe and Wiebe (1987), also presented in chapter 2).

5.3.1 Strategy discovery

Because intentions confer the network with an internal state, they enable decisions to have a long-term effect. Notably, a single explorative decision can push the agent “off-course” for several subsequent actions. This is illustrated in figure 5.7.

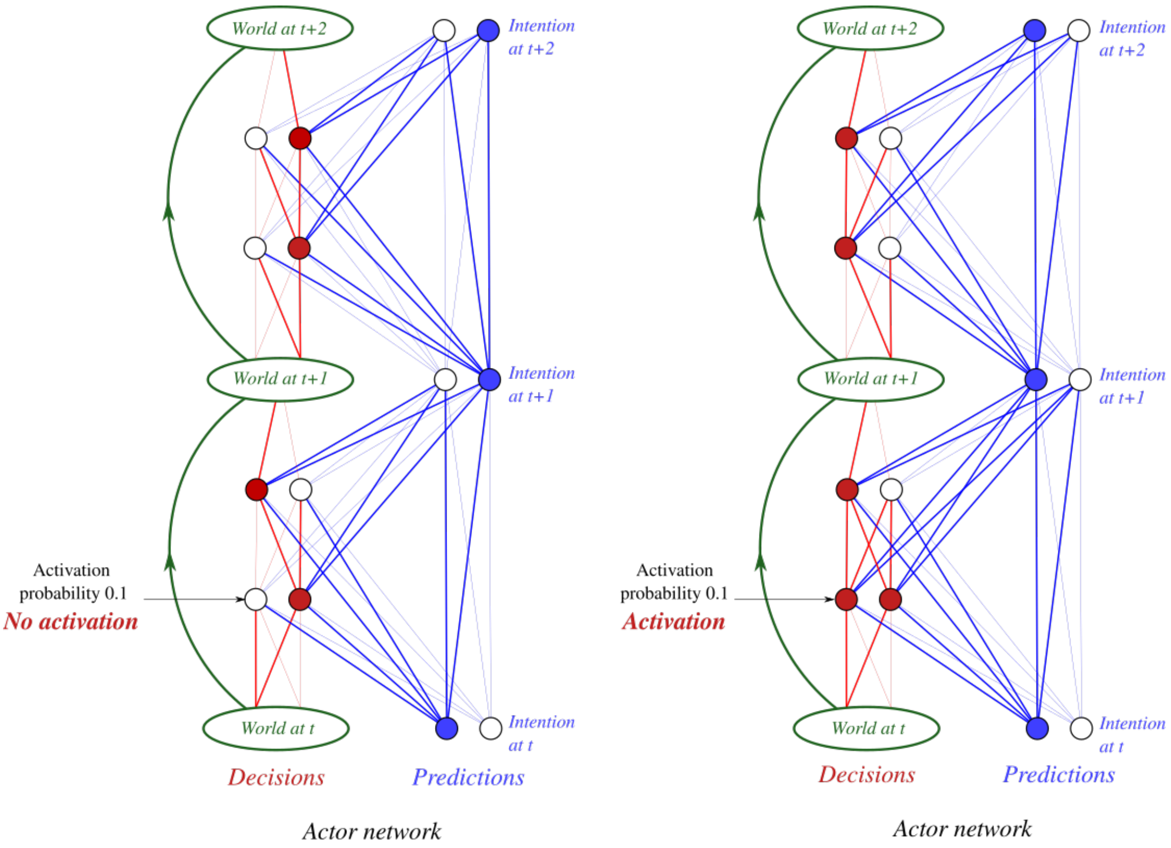


Fig. 5.7 Two actor networks using Bernoulli-logistic neurons and a predictor (intention) network. Neurons filled in red are activated (output 1) whereas neurons filled in white are not activated (output 0). The output connections of activated neurons are shown using bolder lines, in order to represent the flow of information from sensation to action.
Strategy discovery: Explorative decisions have long-term effects. These effects correspond to how the agent has statistically behaved after making similar decisions - thus insuring that the long-term effect correspond to a coherent course of action.
Restructuring: The temporally extended effects of an explorative decision correspond not only to actions, but also to intermediate/hidden layers of the neural networks, i.e.: representations. This corresponds to restructuring.

Furthermore, this deviation is not purely random, but is guided by prior learning: one decision can affect the manner in which the agent seeks to solve the problem in the long term, based on prior experience. One may view this as an analogy stretching over time: the agent's decision consists in trying to consistently apply a previously acquired strategy onto the current problem, step-by-step.

5.3.2 Restructuring

Intentions affect the external behavior of the agent, but also the internal processing leading to that behavior. This internal processing can be interpreted as decision-making (in that it affects behavior) or perception (in that it determines which features are activated on each hidden layer); one way to acknowledge this ambiguity is to call the stochastic activation of a hidden unit a “perceptual decision”. Such perceptual decisions cause extended changes in an agent’s strategy; they also cause changes in which hidden units are activated, which patterns are detected; thus which representations are used.

The persistent change in internal representations is most useful for explaining insight. This is achieved by the ACI architecture, as shown on figure 5.7. Due to the recurrent connections, an agent’s change of representation at time t is propagated forward in time, based on which features of future observations are made relevant by the current decision, and whether or not they are predictable.

5.3.3 “Aha!” experience

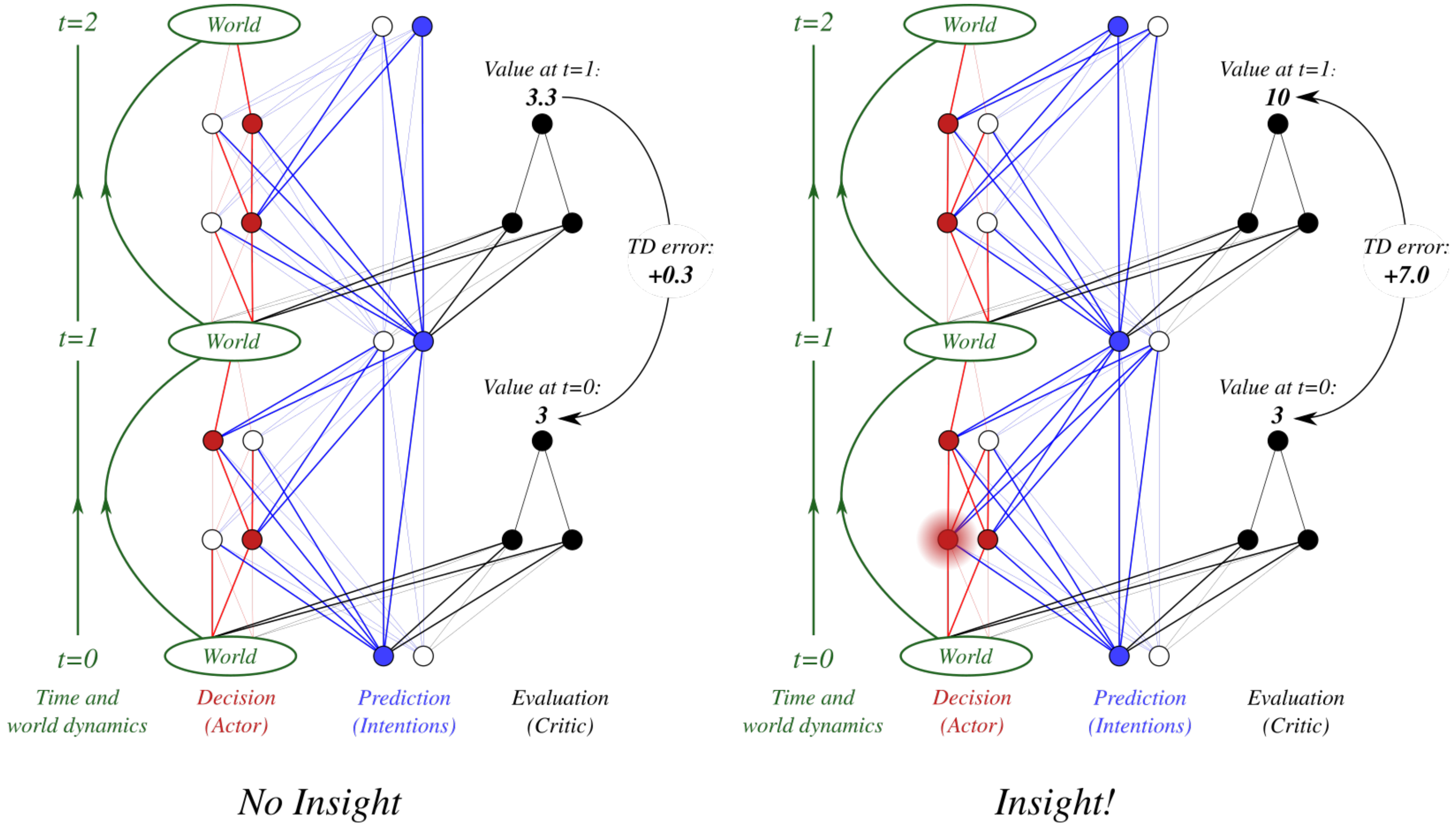
Recall that in the ACI architecture, the critic no longer judges the situation based solely on the state of the environment, but also takes into account the agent’s internal state: its intention for the future. A consequence of this is that a change of intentions can have an effect on estimated value, for instance if the critic estimates that the new intention is very likely to succeed.

If the agent undergoes restructuring, a sudden and large change in expected return may occur. If this change is negative, this means the agent is trying out a strategy that is, per its own evaluation, unlikely to succeed. If on the other hand this change is positive, then the agent is trying out a new strategy that it evaluates positively. This corresponds to the affective aspect of the “Aha!” moment: the sudden increase in “feelings-of-warmth” ratings, occurring suddenly, and before the anticipated solution has actually been carried out, or worked out in detail.

5.4 Conclusion

5.4.1 Summary

The ACI architecture is designed to remedy the limitations of existing HRL methods with regards to exploration. To achieve this, it makes use of:



(a) **No insight.** The most probable perceptual decisions are made, and the actor behaves in a manner that produces predictable results. The critic network measures a small time-difference error as a result of this decision (perhaps due to stochasticity in the world dynamics). No "insight" occurs: things are happening in a largely expected manner. The agent is presumably slowly getting closer to its objective.

(b) **Insight.** A single (in this case) improbable perceptual decision is made (highlighted in red). This affects many subsequent activations in the actor network, both on the same time-step and in subsequent time-steps through intentions, in a manner that is statistically consistent with the perceptual decision based on prior experience. The change of intentions is evaluated by the critic based on experience. In this case, the new intention is judged more likely to succeed, resulting in a large, positive time-difference error: *"Aha!"*

Fig. 5.8 No insight vs. Insight using the ACI architecture

- An intention or predictor network, whose role is to predict the pre-activation values of the actor.
- A critic network which computes value and temporal difference errors.
- An actor network consisting of Bernoulli-stochastic units, whose role is to do structured prediction of the correct actions/intentions, and to explore representations.

Agents using this architecture adjust their behavior only on the basis of new, relevant, unpredictable information: anything else should be predicted by the intention/predictor network. As a result, such agents are likely to avoid redundant processing, and therefore to remain true to their explorative decisions.

The architecture can in principle model insight, because (like options) it is capable of:

- Strategy discovery;
- Restructuring;
- "Aha!" temporal difference errors.

Unlike options, the ACI architecture is based on learning temporal structure, and is therefore more likely to generalize well – as is needed for analogical strategy adaptation. Because the intentions in the ACI architecture are vector-valued and continuous, they may be more flexible than options, notably in their capacity for recombining temporally extended patterns.

5.4.2 Contributions

This chapter contains the main contribution of this thesis, the ACI architecture.

- The $\varepsilon - \upsilon$ exploration technique presented in section 5.1.1;
- The ACI architecture, which is:
  - A novel approach to temporal abstraction in reinforcement learning;
  - A model of insight;
  - A model of intentions.

5.4.3 Bibliographical remarks

- **Information theoretic concepts in Reinforcement Learning:** Information theoretic concepts have been used in a variety of contexts within RL. The most common application has been seeking to *maximize* entropy of action-selection (Lee et al., 2018; Mnih et al., 2016), or the overall Markov Process entropy rate (Savas et al., 2018), in both cases for the stated aim of increasing exploration. Another approach has been to seek to minimize information loss during learning, thus improving stability and hopefully maximizing the policy's relevance to the whole learning domain. This is achieved by minimizing the Kullback-Leibler divergence of the policy during learning updates, in two algorithms called Relative Entropy Policy Search and Trust Region Policy Optimization (Peters et al., 2010; Schulman et al., 2015). This has also been used in a hierarchical context (Daniel et al., 2016). A (technical) discussion of the relationship between these two uses of entropy is proposed by Neu et al. (2017). The treatment that is closest to ours, and which has most inspired the ACI framework proposed here, is that of Tishby and Polani (2011).

- **Self-prediction in brains and in artificial neural networks:** The idea that neural networks ought to predict their own future activations has considerable history both within deep-learning and in psychology and neuroscience. Indeed, some have sought to explain all of biological cognition as based on prediction (Adams et al., 2013; Friston and Kiebel, 2009); whereas others have focused on the context of perception (Rao and Georgeff, 1991). In either case, predictions at multiple levels of abstraction (where sensory data or atomic actions form the least abstract level) lead naturally to self-prediction in intermediate levels (Clark, 2013, p. 183): "One key task performed by the brain, according to these models, is that of guessing the next states of its own neural economy". In AI, similar ideas have been successfully applied, for instance, to video prediction (Lotter et al., 2016).

Part III

Intentions, Insight, Creativity

Chapter 6

About Intentions

The previous chapter introduced the Actor-Critic-Intention (ACI) architecture, which is intended as a model of insight and as a technique for temporal abstraction and deep exploration in reinforcement learning. However, little justification was given for the name "intention". Almost no consideration was given to the rich interdisciplinary literature on intention, which ranges from topics such as "intentional" action, intention as related to beliefs and desires, or "prospective memory", among others.

This chapter seeks to correct this, acknowledging sources of inspiration, highlighting similarities and contrasts, and suggesting new directions for research on intention. Thus this chapter temporarily turns away from the insight literature to discuss instead what underlies insight[1]: intentions. I propose an extensive review of related concepts in relevant disciplines: philosophy, psychology, and artificial intelligence.

In the introduction, I expressed hope that investigating the insight phenomenon might provide answers relevant to both psychology and AI. This chapter investigates the manner in which the ACI architecture relates to the psychology and AI literature - both with respect to precedents, and with respect to potential future applications. Throughout this discussion, intentions as used in the ACI architecture will be compared to other accounts of intentions.

6.1 In philosophy

This section must begin with a disambiguation. In the philosophical literature, the concept of *intention* (introduced[2] into contemporary philosophy by Anscombe (1957)) differs impor-

[1] If the theory proposed in this thesis is correct.

[2] There are of course prior mentions of intentions, but Anscombe's book is the first extended and focused treatment of intentions.

tantly from that of *intentionality* (introduced by Brentano (1874/2014), and more recently by Searle (1983)). According to Searle (1984):

> Intentionality is that feature of certain mental states and events that consists in their (in a special sense of these words) being *directed* at, being *about*, being *of*, or *representing* certain other entities and states of affairs. (...) The obvious pun on "Intentionality" and "intention" suggests that intentions in the ordinary sense have some special role in the theory of Intentionality; but on my account intending to do something is just one form of Intentionality along with belief, hope, fear, desire, and lots of others (...)

The two ideas, intention and intentionality, are related. However discussions and disputes about "intentionality" (see for instance Chalmers (1996); Dennett (1989); Searle (1983)), which range from metaphysics and epistemology to philosophy of mind[3] are only tangentially related to the features of intention relevant here. Comparatively, discussions and disputes about "intentions" are more focused within the subfield of the *philosophy of action* (key texts include Anscombe (1957); Bratman (1987); Davidson (1963, 1978)). "Intentionality" is therefore outside the scope of this thesis. I will be interested specifically in *intentions* rather than in intentionality. According to the Stanford Encyclopedia of Philosophy (Setiya, 2015) (emphasis mine):

> Philosophical perplexity about intention begins with its appearance in three guises: **intention for the future**, as when I intend to complete this entry by the end of the month; the **intention with which someone acts**, as I am typing with the further intention of writing an introductory sentence; and **intentional action**, as in the fact that I am typing these words intentionally.

Although much of the philosophical literature on intention treats the "three guises" together, some have argued for treating intentional action separately (Holton, 2009; Mele and Moser, 1994). For now, I focus on the first two guises, which relate to the philosophy of mind, at the exclusion of "intentional action", which relates more closely to moral philosophy[4].

Within this reduced scope, there remains a rich variety of views. Below I summarize some of the leading accounts in the literature, with a special emphasis on the work of Bratman (see Bratman (1987), or Bratman (1990) for a summary), whose aim aligns best with those of an AI scientist.

[3] Cognitive scientists and artificial intelligence researchers will perhaps be most familiar with the Chinese Room argument of Searle (1980).

[4] Under what description an action is intentional, or unintentional, is important for assigning moral responsibility for its positive or negative consequences.

6.1.1 Beliefs, Desires, and Intentions

In philosophy, an agent's (mental) life is often understood as populated by beliefs; desires; sensations and perceptions; actions and intentions; and sometimes other things such as qualia. In particular, the decisions made by an agent, and the reasons for these decisions, have typically been understood using beliefs (or their probabilistic counterpart, credences), desires; and intentions (Chandler, 2017; Steele and Stefánsson, 2016).

Are all three (belief, desire, and intention) necessary? It is tempting to view agents as harboring only *beliefs* and *desires*; a view often attributed to Hume (1738/2000) (see Cohon (2018)). One may speculatively consider how a Belief-Desire framework might be applied to Reinforcement Learning agents: the state and its abstract representation would presumably be analogical to "belief"; whereas Q-values (in a critic-only architecture) or preferences (in an actor-critic architecture) would be analogical to desire[5], with rewards being the cause of primary desires. In non-hierarchical reinforcement learning, there seems to be no need for an extra component corresponding to intentions.

However philosophers starting with Anscombe (1957) have argued for the importance of intentions. There has since been debate about whether it is possible to give a reductive account of intentions in terms of beliefs, or desires, or some combination thereof (see e.g. Ridge (1998); Setiya (2007); Velleman (1989) for such reductive accounts). The predominant view is that intentions are not so reducible or are, at any rate, sufficiently specific to deserve a separate treatment (Bratman, 1987; Holton, 2009).

Indeed, an intention cannot be a *mere* desire: desires may be inconsistent (I may want/desire to go to two incompatible events, such as a conference in North America or a wedding in southern France); whereas intentions seem constrained by a consistency requirement (I may intend to go to the conference, or to the wedding; but it would be irrational to intend to do both). Beliefs also must be consistent. This suggests intentions might be beliefs - in this view *intending X* would be synonymous with *believing that you will X*. However, like desires, but unlike beliefs, intentions have something like a motivating or normative character (Brand, 1984; Lumer, 2013; Percival, 2014): if I intend to go the conference, then I *ought* to make the corresponding travel arrangements. Thus intentions are neither quite like special desires, or quite like special beliefs.

Let us briefly relate these characteristics of intentions discussed in philosophy with those of intentions in the ACI architecture: like philosophical intentions, these intentions correspond in some sense to a belief about future actions (this is how they are learned). However, they are distinct from memories in that they are the product of decisions, and they

[5] Typically, beliefs, desires, and intentions are understood as propositional in nature, whereas RL systems are not. See Dennett (1989) for discussion on propositional attitudes as made of non-propositional components.

have a motivating character in that they influence future decisions. Finally, ACI intentions allow agents to make decisions “once and for all” about mutually incompatible choices, thus displaying the consistency characteristic that philosophers have noticed in intention.

6.1.2 Intentions and practical reasoning

Whether irreducible components or composite constructs, then, intentions are of particular interest to philosophers. In the influential account of Bratman (1987), intentions are a psychological state, separate from beliefs and desires, that plays a key role in cognitive processes such as practical reasoning. On a conservative (Humean) account, practical reasoning consists in weighing beliefs and desires (inputs) to decide on actions (outputs). But most (see e.g. Castañeda, 1975; Thompson, 2008) see an important role for intention as both inputs and outputs of practical reasoning.

Intentions, characteristically, are future-directed[6]. Bratman (1990) asks: *“Why bother with future-directed intentions anyway? Why not cross our bridges when we come to them ?”*. Part of the answer, for Bratman, is that we are cognitively limited - having only so much time for deliberation at our disposal, it is essential to preserve and retrieve prior deliberations. Intentions, then, form partial or abstract plans, extended courses of action, the intermediate result of a prior deliberative effort. Furthermore, they allow for deliberation over time, with further intentions building onto previously formed intentions. Recall the wedding vs. conference example: your intention to go to the conference guides you to new decisions and sometimes to dilemmas – buying plane tickets, declining the wedding invitation – which you will solve by acting on them or by forming newer intentions.

A particular difficulty is the place of reconsideration. For instance, finding that you do not have the heart to decline the wedding invitation, or perhaps having seen an advertisement for holidays on the Côte d’Azur, you might decide to renounce your intention to go to the conference, and instead attend the wedding in France. When exactly is it rational to reconsider an intention? Bratman discusses the issue extensively, weighing limited cognitive abilities against the potential for improving on an earlier decision.

How does this account match with intentions in the ACI architecture? These intentions allow for a reduction of redundant processing, and thus perhaps help reduce the cost of computation; however this is not the theoretical justification, and there most likely are cases in which their use result in more overall processing (when adding the computation cost for both decisions and predictions). Nevertheless, intentions in the ACI architecture are both inputs and outputs for practical reasoning, via the recurrent hidden layer. They are future-

[6] Recall that I am not dealing with intentional action, for which this is not necessarily true.

directed - an agent's intentions affect its future decision making, and thus the formation of future decisions, representations, and intentions. This suggests that ACI intentions form a plausible substrate for deliberative reasoning over time, a process by which an agent refines its decisions through several consecutive processing steps (I will discuss this idea further in the upcoming discussion of intentions in psychology and AI).

The view of intentions as "abstract plans" is also compatible with intentions in the ACI architecture. Indeed, because the ACI architecture seeks to predict whichever future decisions are predictable, while leaving unpredictable ones to future decision-making, they constitute a kind of "abstract plans", or plan outlines (as opposed to the fully specified plans that one obtains from sequential predictions of the future e.g. in decision-time planning (Sutton and Barto, 2018, p. 181) including tree search techniques as in e.g. AlphaGo (Silver et al., 2016)). The ACI architecture also offers an alternative answer to the question of reconsideration: intentions should be updated[7] when *unanticipated relevant evidence* is observed.

6.1.3 Intentions and theory of mind

Before concluding this review of the philosophical literature, a word should be said on the ontological status of intentions. Are intentions a thing "in the brain", and if not, what are they? Philosophers usually attribute both a *causal* and a *justification* role to intentions (intentions can be the *cause* for actions or further intentions, as well as the *reason* for them). Davidson (1963) has argued that the two are compatible. In contrast, Dennett (1989) argues that "mental states" including intentions, are usually conceived as propositional, but that a neuroscientific approach will likely not be. Hence he concludes that intentions are best understood as emergent *patterns* in an agent's behavior, which can be harnessed to predict that agent's behavior. This is applicable both to other agents (theory of mind) and to oneself (self-understanding, self-awareness). According to Dennett, it would be excessive to assign to intentions a causal role in the generation of that behavior.

Intentions in the ACI architecutre, of course, have a causal role in decision making, and are thus aligned with the views of Bratman or Davidson against those of Dennett.

6.1.4 Summary: intentions in philosophy

I have summarized a rich literature on intentions in philosophy, focusing especially on the work of Anscombe (1957), Bratman (1987), Davidson (1963, 1978), and Dennett (1989).

[7] Note that rather than be "reconsidered from scratch", as if there was a meaningful intentional "blank state", intentions in the ACI architecture only ever get modified/updated.

The picture that emerges is of the compatibility of the ACI architecture presented in this thesis with the philosophical literature on intention.

The ACI architecture can in theory support the many functional roles given to intentions in philosophy, including the faculty of internal deliberation and the creation of abstract plans, while displaying the same characteristics of being future-directed, of serving as both input and output of reasoning, and of maintaining internal consistency.

Finally, the ACI architecture proposes answers to at least two of the philosophical questions set by the philosophical literature: first, it suggests that reconsideration is not an all-or-nothing process, but rather that intentions are constantly updated, to the extent needed, based on new relevant information; secondly, it suggests that biological intentions are not merely an emergent phenomenon useful for theory of mind, but have a genuine causal role in controlling behavior.

6.2 In psychology and neuroscience

By comparison with philosophy, the concept of intention has received relatively little attention in psychology (but see Holton (2009), chap. 1, for a discussion of the psychological evidence relative to intentions). When intentions are discussed in a psychological context, it is often in the context of theory of mind[8], in a manner resembling Dennett's treatment of intentions presented in the previous subsection. In neuroscience, most discussions of "intention" appear to focus on "intentional action" often with a view to consciousness, free will, and personal responsibility. Intentions as *future-directed decisions* are curiously rare in the psychology and neuroscience literature.

However, the concept of prospective memory is closely related: prospective memory consists in "the realization of delayed intentions" (Brandimonte et al., 1996)[9]. Psychologists and neuroscientists investigating prospective memory have mostly treated it, as the name implies, as a form of memory. The special role of intentions in planning, decision, reasoning, or reasons for action, recognized in the philosophical discussion of intention, is usually not considered here. Hence research on prospective memory focuses on *retrieval* and *recall*, rather than on the *formation* or *reconsideration* of intentions, or the re-use of prior intentions to form new ones; see e.g. Brandimonte et al. (1996); Simons et al. (2006); Volle et al. (2011); but also Burgess et al. (2011); Poppenk et al. (2010) for work that considers the

[8] In psychology, theory of mind refers to the ability to attribute mental states (such as beliefs, desires, emotions, intentions) to others or to oneself. Young children struggle or are unable to deduce or keep track of another individual's mental states. See e.g. Wimmer and Perner (1983).

[9] Perhaps owing to this difference in terminology, it appears that related research in philosophy, psychology, and neuroscience has seen little cross-fertilization.

formation of such “memories”; Cona et al. (2015) reviews much of the relevant literature. Nevertheless, some studies consider the manner in which such memories interact dynamically with contextual cues; see e.g. Scullin et al. (2013).

What are the neural correlates of intentions as prospective memories? Based on the literature on biological insight and reinforcement discussed so far, one would expect activations in the PFC, especially in the ACC. Indeed, prospective memories are mainly correlated with activations of the PFC compatible with a role of the executive control network in maintaining intentions, and thereby (at least superficially) compatible with the reinforcement learning theory presented in this thesis. On the other hand the ACC, although also among the regions related to prospective memories, does not have an obviously dominant role in that process (Burgess et al., 2011).

Overall, there is a paradigm difference between intentions in the ACI architecture and prospective memory: the former are not memories of past decisions (akin to memories of events) that need to be “stored” and “retrieved”, but decisions for the future (see section 6.4). This makes it difficult to discuss in detail how the ACI architecture relates to results from the psychological and neuroscientific literature.

6.3 In artificial intelligence

In contrast to psychology, the concept of intention has been applied to AI with some success in the Belief-Desires-Intention framework. However, it took a surprisingly different direction than the ACI architecture (especially considering their common relationship with Bratman’s work). The difference is mostly due to the respectively subsymbolic and symbolic nature of the ACI architecture and the BDI framework.

Within AI, I also offer some discussion of matters which have rarely or never been described as involving intentions: machine learning approaches to planning (section 6.3.2) and to reasoning (section 6.3.3). These two sections contain speculative, high-level considerations relevant to almost all of AI. They are presented here, without experimental validation, because they constituted part of the motivation for the ACI architecture.

6.3.1 The BDI framework

The account of intentions proposed by Bratman (1987) overlaps with artificial intelligence considerations. The influential BDI model of rational agents (Rao and Georgeff, 1991) is a fleshed out account of intention for rational planning agents, directly inspired by Bratman’s work. This model/framework consists of a modal logic[10] in which intention is formalized based on a branching-time possible-worlds model.

In particular, most implementations of BDI models can make decisive choices between mutually incompatible goals, and can intend not an entire course of action, but only certain aspects of it. For instance in the marriage/conference dilemma that I have used throughout this chapter: choosing the course of action/forming the intention of “going to the conference” does not imply intending to “hurt the feelings of the newlyweds” even though one *believes* this is an inevitable side-effect of going to the conference.

Although the BDI architecture is close to the ACI architecture with respect to the intentional element, it is radically different in several other respects. In particular, it is a logic-based framework, which depends on the symbolic representations given to it; whereas the ACI architecture proposed in this thesis belongs to the statistical machine learning paradigm. As a result, the ACI architecture can account for representational phenomena such as restructuring, which are not accessible to BDI. This paradigmatic difference makes it difficult to compare BDI and ACI. It may be, however, that the combination of symbolic and associative reasoning (afforded to human beings by language, and discussed in AI e.g. in Besold and Kühnberger (2015); Garnelo et al. (2016); Medsker (2012); Sun and Alexandre (2013)) may allow a convergence between BDI and ACI; but this discussion is beyond the scope of this thesis.

A characteristic of the BDI architecture is that it treats intentions as dealing with possible worlds. In contrast, the ACI architecture maintains only one intention at a time - it does not seek to entertain and compare different intentions. This leads to perhaps the most glaring departure of the ACI architecture compared to human insight: it is possible, in the intention architecture, to “achieve” “negative insights”, in which the agent adopts a strategy, evaluates it as a bad idea, and proceeds with it anyway. In contrast, human beings are seemingly capable of modifying their behavior immediately based on the output of the critic. Could the intention architecture be modified to allow the actor to react directly (rather than slowly, through learning) based on feedback from the critic - as if it was “entertaining” a course of action, and abandoning it, or continuing with it, depending on its “gut feeling” as provided by the critic? Perhaps this could be achieved using techniques similar to those of Wang et al. (2016). This, again, is beyond the scope of this thesis.

[10] Modal logic extends classical logic with modality operators. Such operators include alethic operators (modalities of truth): rather than only p, we might have *possibly p* or *necessarily p*. Modal logic thus extends to making deductions about possible worlds, which is useful in situations of uncertainty, or when an agent is considering possible courses of action.

6.3.2 Intending and planning

In contemporary AI research, the sub-symbolic, statistical approach of machine learning is preponderant. But it has not been clear how to do planning or reasoning over complex, multidimensional representations. To the question "how to make decisions about the future?", a tempting response is to try to do, in the statistical setting, what worked well in a symbolic setting. That is: predict the possible next states for of each successive action, and from each of these states continue unfolding the graph of possibilities.

However, notably in the reinforcement learning context, it has proven difficult to learn models capable of looking far into the future. Although symbolic planning methods continue to be efficient for domains where the dynamics are known (e.g. chess and Go; see Silver et al. (2016, 2017)), they cannot compete with "model-free" algorithms[11] for domains where state spaces are large and initially unknown, such as Starcraft II (Vinyals et al., 2017, 2019) or Atari games (Mnih et al., 2016). Why is it difficult to use step-by-step predictive planning in (more) realistic environments? A major reason is that the realistic environments are computationally expensive to represent adequately. Thus what is tractable when dealing only with an observation of the present situation, becomes intractable when seeking to model a whole decision graph; and this complexity explodes when uncertainty arises from either stochasticity, or ignorance of the real dynamics, or a large range of possible decisions. One ends up modeling an intractably large number of possible future worlds.

The ACI architecture, or some of the concepts developed as part of it, can help tackle this problem in three ways. Firstly, ACI intentions predict the features that are activated in making a decision - that is, only those features of the world that are *relevant for action* according to the agent's current strategy[12]. Thus ACI intentions help reduce the computational cost of prediction, by focusing on predicting *relevant* aspects of the world. A second, additional way to reduce computational expenditure is to avoid unnecessary updates of these predictions, instead taking into account only novel (relevant) information: this is achieved by un-learning reinforcement learning activations that could be predicted. Finally, the third way consists in learning probable patterns of temporally extended behavior, allowing for reducing the branching factor by trimming unlikely patterns.

Although some of the ideas behind ACI intentions seem relevant for planning, the algorithm presented in chapter 5 does not plan in the classic sense of building a graph of sequences of actions and their consequences. What it does, instead, is use a critic network to estimate an expected return using the agent's current situation (combining observation and

[11] Which do not use a model at all.

[12] In a loose sense, intentions help solve a problem related to the "frame problem" (Dennett, 2006; Hayes, 1981): how is an agent to keep track of which features are relevant for the problem at hand?

intended strategy). One may view this operation as solving the role of planning in a more direct manner: rather than first building a graph, then evaluating its nodes based on some input data, one directly seeks to obtain a numerical estimate from the input. This remains true for other model-free architectures; it has been convincingly argued (Wang et al., 2016, 2018) that such "meta-learning", arising from model-free learning, might explain apparently model-based decision-making in human beings. Although this can occur for a range of RL architectures, in the ACI architecture the input to the critic is enriched with *relevant* predictive information.

The ACI architecture was inspired in part by concerns, within AI, relative to planning. Its approach is to limit prediction to relevant features only, and to update these predictions only when new relevant information is obtained. As such the architecture can be viewed, especially when considered in tandem with the critic, as a form of planning. Whether this idea would work experimentally could be tested in the meta-learning architectures used by Wang et al. (2018).

6.3.3 Intending and deliberating

The ACI architecture allows for an uncommon type of deliberation in RL systems. Indeed, non-reflex behavior in RL takes two main forms (besides Options):

1. Various planning or model-based techniques, where the agent either learns or is given knowledge about transitions, and uses this to improve the policy. See for instance Sutton (1990); Tamar et al. (2016).

2. Techniques for tackling partial observability (Sutton and Barto, 2018, p. 467), such as Partially Observable Markov Decision Processes (POMDPs)[13], where the agent learns to estimate a hidden state of the world.

These methods escape reflex behavior by way of modeling the world: planning techniques make use of transitions, and techniques for POMDPs seek to learn about a hidden state. In contrast, the ACI architecture seeks to identify patterns in an agent's successful behaviors, involving both the external world state and the agent's internal state. As a result, an idle ACI agent (perhaps with a "wait" or "think" action) might make successive adjustments to its intentions on consecutive time-steps, despite constant external input. This would occur without an explicit world model or tree-like expansion. In that sense, such an ACI agent performs internal deliberations different from existing RL approaches.

[13] An alternative that is conceptually closer to the ACI architecture, Predictive State Representations (Jaeger, 2000; Singh et al., 2012), departs from the POMDP framework, but focuses on predicting observations.

6.4 Intentions or memories?

Above I have sketched an account of “intention”. There are obvious similarities with the options framework (Sutton et al., 1999): options also commit an agent to an abstract “course of action” in a revisable manner (via termination of the option); they allow for temporally extended behavior; they permit abstract planning; higher-level options can serve as reasons or causes for more concrete actions.

However, within the option framework, it is unclear how one might learn general options - options that correspond to temporally extended patterns of behavior (for instance, the option-critic architecture cannot learn temporal relationships between decisions). Furthermore, the framework considers patterns of behavior that are discrete and exactly hierarchically structured. This limits the kinds of temporal patterns that are possible. Specifically, patterns on the same hierarchical level cannot overlap; and the maximum number of simultaneous patterns that can be encoded is equal to the number of hierarchical levels.

In reinforcement learning, eligibility traces[14] are traditionally understood (Sutton and Barto, 2018, chapter 12) in two ways:

- The forward view seems to suggest the future is already known, but it allows for seeing, all at once, all the updates that affect a given decision. Mathematically, this makes it easy to see why the combination of updates is sound.
- During the execution of an algorithm, an agent must wait for the sequence of future states, actions, and rewards to unfold before it knows the size and direction of each update. The “backwards” view of eligibility traces makes it clear how to write an algorithm that performs updates as information becomes available.

Reinforcement Learning algorithms using eligibility traces must use the backwards view, but they produce updates that are identical to those of the forward view.

In this section, I describe a related change of perspective. In the case of eligibility traces, switching views enables one to see how to *learn* across time-steps. In the case of intention, a similar switch allows one to see how to *act* across time-steps.

6.4.1 The memory view

The common way to envision action-across-time in reinforcement learning is as a transfer of information from the past to the present; typically, this is used to solve Partially Observable

[14] Eligibility traces are used in algorithms such as TD(λ) to propagate learning from TD-errors more than a single time-step in the past. Thus when a TD error is encountered, the values all states in the eligibility trace are updated, with the most recent ones receiving a proportionally larger update.

Markov Decision Processes (POMDP), in which at every time step t, instead of being capable of observing the state s_t, the agent can only make an incomplete "observation" o_t. The objective of these approaches is to understand, at a time-step t, what is the non-observable current *hidden* state of the world s_t. In other words, the objective of memory is to maintain an internal picture of what the world *really* is like despite ambiguous appearances. If one were able to deduce the real state of the world, or at least whichever aspects about the world are relevant to making a decision, then one would "simply" have to solve the resulting MDP.

Figure 6.1 illustrates the mindset underlying these views. At every time step, the agent must decide on its next move; but because the current observation does not contain all the relevant information, it is necessary to look back in time, and retrieve additional relevant information. Because the past is no longer available, the agent must learn to collect a summary of relevant past observations to use for future decisions.

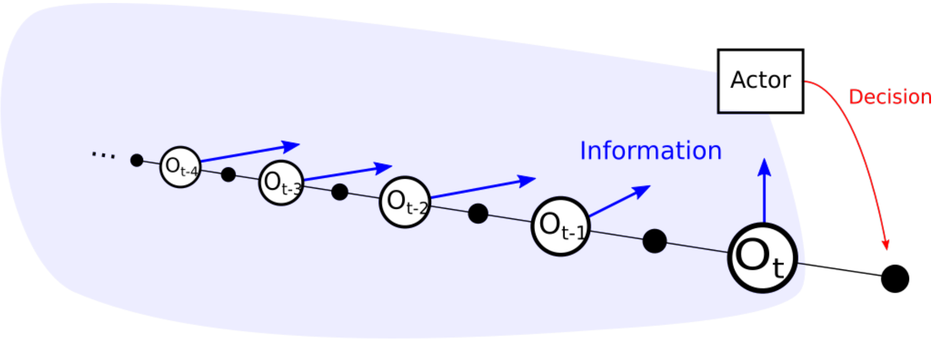


Fig. 6.1 The memory view of temporal relationships between states. At each time-step, the agent processes all the relevant available information from current and past observations $o_{t-n}...o_t$, in order to make the best possible decision for the current time-step as a function of the entirety of the available data.

6.4.2 The intention view

I propose an alternative view[15], referred to as the intention view. Each state provides the agent with new information; the agent must use that information to *decide* on what to do next. That is, our perspective shifts from recording information about the past to forming intentions about the future.

This is not a cosmetic change. A first consequence is that temporal relationship between states become interesting even in a fully observable MDP. Indeed, in the memory view past observations are useful to the extent that the current observation is incomplete. In the

[15] Of course, the memory view is still necessary for POMDP. The intention view goes alongside the memory view, serving a different purpose, rather than replacing it.

intention view, a state t can contain information relevant to the action at $t+2$; thus it is possible to anticipate future actions. This view is illustrated in 6.2.

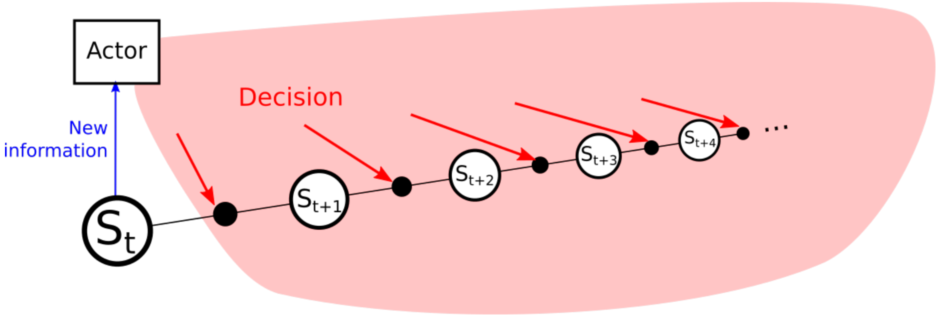


Fig. 6.2 The intentional view of temporal relationships between states. At each time-step, the agent processes new relevant information from the current observation, in order to make necessary adjustements to the current plan for action.

The benefits of the intention view may not be immediately clear: after all, if the environment is an MDP, then it is possible to behave optimally using only the current state. However, doing so would generally imply *redundant* decisions, that is, repeated arbitrations based on the same information, processed over and over. Since it is always possible to know the *best* decision, nothing stops us from successively choosing the right arbitration; but this seems wasteful. Worse, an exploring agent might explore in a self-contradicting fashion, interpreting the same information in contradictory ways in successive decisions. By interpreting information as soon as it becomes available, making a decision at that moment and sticking to it thereafter, the agent can ensure coherent explorative behavior.

6.5 Conclusion

6.5.1 Summary

An interdisciplinary literature review reveals considerable work on intention from different angles - with limited communication (and limited mutual awareness) between them.

In philosophy, intentions are seen as having the following characteristics:

1. In the “realist” view (adopted for instance by Bratman (1987)), intentions resemble or imply beliefs (or credences – Holton (2009)) about one’s own future behavior,

2. Intentions can act *both as the input and the output* of deliberative processes, enabling reasoning over time,

3. Intentions constitute flexible, revisable *commitments*, or partial/abstract plans;

4. Intentions ought to be *consistent* with one another (thus resolving conflicts between desires);

5. One reason to make use of intentions is the *reduction of cognitive costs*.

The ACI architecture presented in this thesis either implements, or is likely capable of all of these functions - justifying the use of the term “intention”, and the analogy between the internal network state in the ACI architecture with intentions in the colloquial use (“doing something with/in view of doing something else”).

In psychology and neuroscience, the concept of intention is often used in the sense of intentional action or with respect to metaphysical concerns about free will or consciousness. The exception is the study of prospective memory. Unfortunately, research in prospective memory tends to assimilate intentions to delayed actions - although some recent work has began to stress the dynamic nature of prospective memory (Cona et al., 2015).

Finally, in AI the work of Bratman gave rise to the BDI architecture, which is difficult to compare to the ACI architecture due to the difference of paradigms, and which may in some ways be complementary to it. Within Reinforcement Learning, there has been little interest in intentions, despite many possible applications. In particular, the concept of intention can be informatively compared to the common usage of memory in RL algorithms, and could serve as a basis for emulating complex cognitive faculties.

6.5.2 Contributions

This chapter consists mostly of a literature review of intention across disciplinary boundaries, revealing relationships between disciplines that function largely independently; the main exception is the link between the BDI architecture in artificial intelligence (Rao and Georgeff, 1991) and the work of Bratman (1987) in philosophy. Some speculative suggestions are made about potential applications of the intention architecture as an alternative to planning and as a basis for reasoning.

Chapter 7

From Intentions to Creativity

In the course of this thesis I have explored the enigma of creativity, and throughout the chapters I have singled out sub-problems. These sub-problems appears increasingly removed from the original question; but their solutions condition the answer to the problem of creativity. In chapter 1, I explored the problem of creativity, and singled out insight problem-solving. In chapter 2, a review of the literature on insight revealed reinforcement learning lying in wait. In chapters 3 and 4, I investigated a reinforcement learning theory of insight – but there was still a deeper, more profound riddle. Finally, in chapter 5, I proposed a theory of intentions, further discussed in chapter 6. Intentions, in turn, offer fascinating challenges, which are beyond the scope of this thesis (but not of ongoing, related work).

In this chapter, I methodically put these sub-solutions back together starting from the last, each time discussing how the latest sub-solution contributes to answering a larger problem, all the way back to the beginning of this investigation: creativity. In the first section, the key contribution of this thesis (the ACI architecture) is discussed in the context of reinforcement learning. In the second section, intentions and reinforcement learning, taken together, are presented in the context of insight. Finally, in the third section, the investigation comes to an end as intentions, reinforcement learning, and insight, are situated in the context of creativity.

7.1 The ACI architecture: a theory of temporal abstraction in Reinforcement learning

How does the ACI architecture fit inside the reinforcement learning paradigm?

Intentions in the ACI architecture are primarily concerned with exploration: they allow for introducing variability to a policy, in such a manner that the policy explores efficiently. Intentions are about “carving good behavior at its joints” - such that explorative behavior can be segmented and interconnected in a promising manner. How does that compare with other reinforcement learning approaches to exploration?

The ACI architecture can be distinguished from two extreme cases:

- **Articulating exploration at each time-step:** The first limit case is exploration in action space. In flat RL using ε-greedy or Softmax action selection, exploration is typically articulated independently at each separate time-step, for each separate action. This results in algorithms for which efficient exploration is heavily dependent on the size of the time-step and on the actions chosen by the developer (McGovern and Sutton, 1998), and which will presumably perform poorly if different actions with varying time-lengths are required.

- **Articulating exploration at each episode:** The other limit case is to explore in parameter space. This can be done in the episodic setting, or with an arbitrary time-limit. For instance, genetic algorithms are one way of exploring in parameter space: introduce mutations in the parameters controlling an organism, let it live its life, and preferentially keep modifications that led to improved performance. A variation of this strategy was proposed by Osband et al. (2016) under the name of "bootstrapped DQN", consisting of using agents with different initial parameters to collect experience.

The ACI architecture finds a principled middle ground: exploration is conducted based on learned, temporally extended patterns. These patterns are likely longer than a single time-step and shorter than an entire episode; the strength of the ACI architecture is that it detects and learns these patterns, and explores for a temporal duration that is not based on prior assumptions made by the human engineer. Instead, the agent discovers structure inherent to the problem, from its own experience.

HRL methods such as options also fall in between these two limit cases; below I discuss how intentions in the ACI architecture relates to other methods that learn temporal abstractions.

7.1.1 Exploration in HRL

Various temporal abstraction approaches have been published, among which the main frameworks are the hierarchical abstract machines of Parr (1998), MAXQ (Dietterich, 2000a), Dynamic motion primitives (Ijspeert et al., 2003), skills (Konidaris and Barto, 2009), feudal networks (Dayan and Hinton, 1993), and options (Sutton et al., 1999). This section focuses on Options, which have become the leading formalization for HRL.

Within the Option-discovery literature, one may distinguish between methods that exploit only the structure of the environment such as Machado et al. (2017); McGovern and Barto

(2001); Metzen (2013); Şimşek et al. (2005) (e.g. by seeking to use bottleneck states or other interesting states as goal-states for options), from those that also take into account the reward such as Bacon et al. (2017); Konidaris et al. (2012). The former methods have tended to be more principled, but at the cost of not factoring in which options are typically relevant to the agent's interests (thus seeking to maximize "empowerment" that allow movement in state space, but not focusing on "empowering" those abilities that are relevant to the agent's goals). One may speak of an "empowerment frame problem": how to focus empowerment on the *relevant* skills? For instance, there is a large number of different ways for a curious agent to injure itself, but a healthy agent ought to investigate other matters, which are more likely to be useful in obtaining positive rewards.

The other class of methods also takes into account rewards. But these methods tend to be more heuristic, lacking an organizing principle which ought to lead to options which generalize well. For instance, the option-critic architecture climbs the gradient of performance, using a small penalty for option changes to encourage temporal abstraction (Bacon et al., 2017): it is not clear why such an architecture, which uses independent gradient descent updates for adjacent time-steps, should produce generalizable options.

A more general issue with the option framework is that, in a straightforward implementation, options of a given hierarchical level are applicable strictly one at a time. That is, having learned the skill of singing and the skill of walking, an option-based agent would still need to learn a new option for walking while singing. In contrast, human beings are capable of combining options not only in temporal succession, but also in temporal conjunction.

A related issue is that of parameterized options. One may think of options as a parameterization of behavior: such that being in a certain option is an input to a general action network (see for instance Bacon et al. (2017)). But why limit oneself to only one parameter? Certainly, having learned to perform a skill at a certain speed or magnitude, it ought to be easy to learn to perform the same skill at greater speed or at a greater magnitude, without having to learn an entirely different option. Thus one would like to have multiple parameters, some specific to an option, some shared across options.

All these diverse considerations form the difficult specifications for the ideal temporal abstraction for RL. The ACI architecture presented in this thesis tackles many of these problems at once by using vector valued intentions (which therefore consist in a range of parameters, and may in certain cases be combined), and by learning temporal patterns. Table 7.1 shows how some of the different existing approaches tackle these specifications.

The ACI architecture seeks to implement these specifications by using not a rigid hierarchy of behaviors, but instead a vector-valued internal state. This can be seen as a

	Feudal Networks	MAXQ	Skills	Option-critic	Intentions
Learnable options	No	No	Yes	Yes	Yes
Return-based	N/A	N/A	No	Yes	Yes
Generalizable	N/A	N/A	No	No	Yes
Online learning	N/A	N/A	Yes	Yes	Yes
Sequential combination	Yes	Yes	Yes	Yes	Yes
Overlapping combination	No	No	No	No	Yes
Parameterizable	No	No	No	No	Yes
Off-policy	Yes	Yes	Yes	No	No

Table 7.1 A comparison of some approaches to temporal abstraction in Reinforcement Learning, including the Feudal Networks of Dayan and Hinton (1993), MAXQ (Dietterich, 2000b), Skill chaining (Konidaris and Barto, 2009), and Option-Critic Bacon et al. (2017). This table is not intended to be exhaustive, but merely to show some of the variety of existing approaches, and the many constraints they seek to satisfy. (For Feudal Networks and MAXQ, all or part of the option decomposition must be provided either by a human or by a different algorithm.)

parameterization of the agent's behavior, allowing for generalizing between similar behaviors.

But perhaps the most important feature of the ACI architecture is the possibility to learn to form intentions "on-the-go", with very limited input from the developer. In the ACI architecture, there is no need to guess the number of options required, the number of hierarchical levels, or to set different learning or exploration parameters for different levels of the hierarchy. There is also no need to explore the state-space before beginning to learn options. Learning intentions can take place online, alongside problem resolution.

The ACI architecture, at least in the context of exploration, promises a range of advantages over existing frameworks: it seeks to carve temporally extended behavior at its joints without falling into the edge cases of parameter search or action search, it allows for exploring the conjunction of skills or the usage of differently parameterized skills, and it can learn online.

7.1.2 Other approaches to exploration

An alternative approach to exploration (previously covered in chapter 3) is intrinsic motivation, including artificial curiosity. The proposed ACI architecture is most likely compatible with techniques such as intrinsic motivation, artificial curiosity, or empowerment: these techniques are based on modification of the reward function or of the value function, whereas intentions transform the policy. For instance, one may use the ACI architecture with an initial value function learned by means of a genetic algorithm (like that used by Singh et al. (2010)) or provide additional rewards when surprise and learning occur (like those used by Oudeyer and Kaplan (2008)).

Another principled approach to reinforcement learning is Bayesian reinforcement learning (see the review by Ghavamzadeh et al. (2015)). This class of methods has been left out of the discussion until now. I believe that the fundamental ideas of the ACI architecture (tracking temporal regularities in behavior; deciding based on new relevant information only) can in principle be used in conjunction with a Bayesian learning scheme, but a full discussion of this question is outside the scope of this thesis.

7.1.3 Temporal and structural credit assignment problems

The credit assignment problem consists in distributing credit for success among the many decisions that may have been involved in producing it.

This problem was introduced in chapter 3; it was then explained how the value function helps in solving the temporal credit assignment problem, by allowing the agent to estimate whether an action has improved its situation based solely upon its immediate consequences. In chapter 3, the commute MDP was used to illustrate these ideas; I return to this example here.

Structural credit assignment is distinct but related to temporal credit assignment. Say, for instance, that an agent's action space is multi-dimensional. For instance, the agent might sit or stand, or read or wait. Thus in the bus, the commuting agent from chapter 3 would have a total of four choices in the bus: sit and wait, sit and read, stand and wait, stand and read. A naive solution is to treat each combination as a separate action(cf. figure 7.1). However, this quickly becomes intractable for large action spaces; for instance, the human body has around 800 muscles, leading to 2^{800} possible actions in the (simplified) case of treating the control of each muscle as a binary decision.

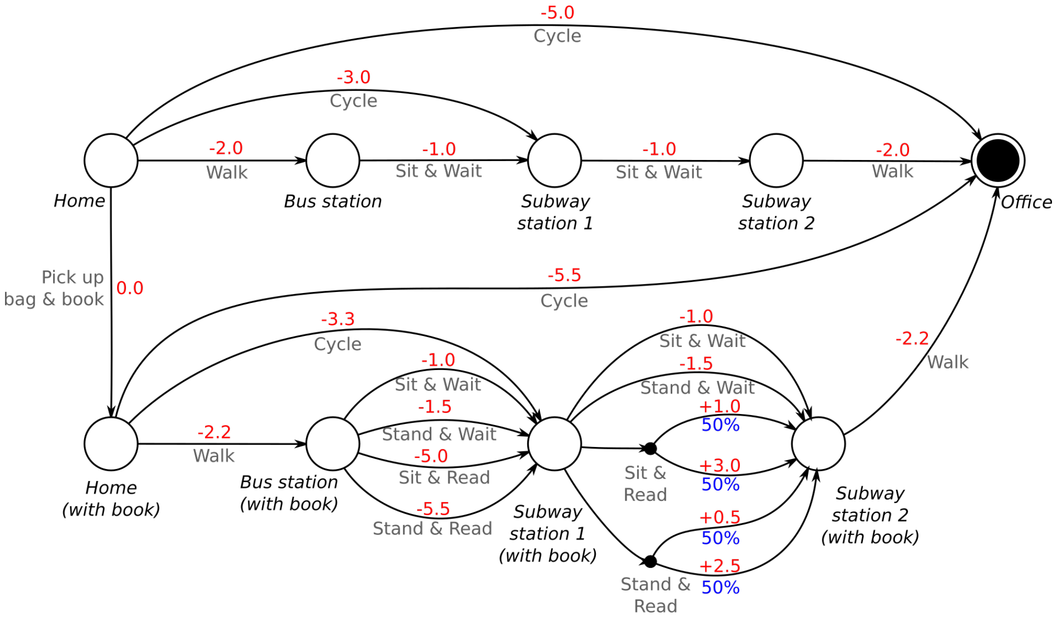


Fig. 7.1 The commute optimization MDP from chapter 4, augmented with the option to stand in the bus and subway.

The structural credit assignment problem consists of identifying which actions in a combined decision have led to success (Sutton and Barto, 2018, p385).

If the actions have mutually independent effects, the structural credit assignment problem is relatively easy to solve. For example the stand/sit and read/wait actions in the commute MDP have independent effects: standing is always worse than sitting by 0.5, and (aside from this) reading is just as enjoyable in either position. This allows for easy learning: just learn separately whether sitting or standing, and reading or not-reading, is better. However, other actions combine in interesting ways. For instance, the Fosbury flop cannot be obtained by optimizing independently each aspect of the scissor or straddle jump: it relies on a completely different and more effective pattern of muscle contractions. Knowledge of these patterns allows for more effective credit assignment: if a different pattern was used, the credit ought to go to the whole pattern taken as a whole, whereas if multiple uncorrelated changes were made, it is not clear which of the changes in behavior was responsible for the change in performance. This is also relevant for exploration: by using such patterns, the agent may reach areas of state-space that are not accessible by using gradient descent on the individual actions[1].

[1] Consider a minimal example: a size two binary action vectors and a single state. Action ‘00’ has value 1, ‘11’ has value 2, and ‘01’ and ‘10’ have value 0. If the agent gets stuck in ‘00’ early on, the gradients will continue to reinforce this local minimum. (This problem is related to XOR, but is more serious; it cannot be fixed by adding a deterministic hidden layer.)

In the ACI architecture, temporal credit assignment and structural credit assignment become intertwined.

To illustrate this, consider the smaller version of the commute MDP, shown in figure 7.2. Here, assuming a critic is present which is capable of estimating the temporal difference error, the agent may learn that the action of reading is better than the usual, and therefore reinforce it. Can we do better?

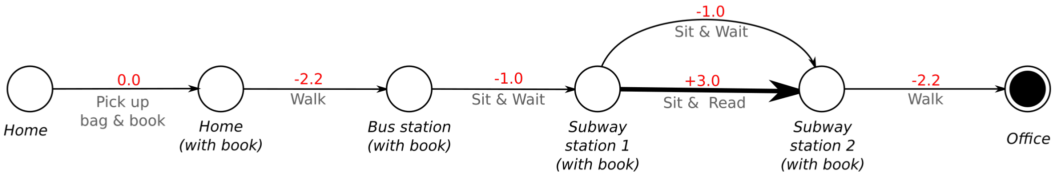


Fig. 7.2 A smaller version of the commute MDP. In green are shown the values of the states, assuming a policy which never reads. The bold arrow represents an exploratory reading action.

As naive observers of the agent, we might say that the agent picks up the book with the intention to read it on the subway. Indeed, if it uses an intention-based architecture, the agent might have previously noticed the following pattern: picking up books is usually followed by taking other actions leading to reading them. Thus when choosing to pick up a book, the agent would also modify its internal state to intend to read the book in the subway. The two apparent actions (picking up the book and, later on, reading it) are linked as part of one pattern, and internally are treated as just one. Therefore they are reinforced jointly: the action of picking up the book gets credit for the subsequent reading of the book.

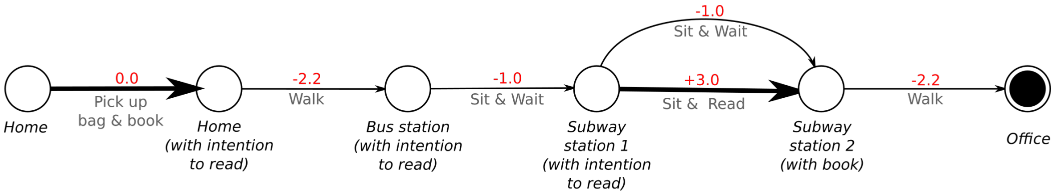


Fig. 7.3 Intention version of the commute MDP. The two related actions (picking up the book and reading it) are treated as a single one, and reinforced jointly.

7.1.4 On-policy and off-policy learning

Recall that off-policy algorithms allow for learning with a policy other than the one currently considered best. Off-policy learning therefore allows for re-using past experiences (obtained with a policy no longer considered best) thus considerably improving the sample efficiency. Unfortunately, converging off-policy algorithms tend to be more difficult to find than their on-policy counterparts, especially when function approximation is used (Sutton and Barto, 2018, chap. 11).

The ACI architecture is an on-policy algorithm. Obtaining an off-policy version seems a daunting task, because intentions are obtained from self-prediction using the on-policy thread of experience. It is not clear whether an off-policy version of ACI is possible.

7.1.5 Summary

The ACI architecture is relevant to several problems in RL. In particular, it is a novel approach to temporal abstraction or hierarchical reinforcement learning. In the context of temporal abstraction, intentions have a range of desirable properties (learnable online from experience, based on reward, can be combined both sequentially and in an overlapping fashion, parameterizable).

The focus of this thesis is the use of intentions for exploration, leading to creative insight. The ACI architecture discovers the temporal structure of behavior, and constrain learning to occur within this already discovered structure, thus allowing the agent to "learn to explore".

A non-obvious related benefit is improved temporal credit assignment. Because intention-based agents make decisions ahead of time, they can reinforce the whole pattern of behavior, rather than the first surprising component (provided an accurate value function). Furthermore, intentions suggest a convergence of the two problems of temporal and structural credit assignment.

7.2 Intentions and RL: a theory of insight

In the previous section, I discussed intentions and the ACI architecture as a theory of temporal abstraction in the context of reinforcement learning. However, this architecture was also intended as a theory of insight. In this section I review the evidence in favor of the RL/intention theory of insight, based both on the experiments presented in this thesis and on the wider literature on insight in psychology and neuroscience.

7.2.1 Reinforcement Learning as a unifying paradigm?

In chapter 2, I have argued that much of insight research has been part of larger controversies over psychological paradigms: for instance between conditioning and gestalt psychology (Köhler, 1921), gestalt psychology and behaviorism (Epstein, 1983), and behaviorism and cognitivism (Ohlsson, 1992). Insight may have been the center of many controversies because none of these paradigms could offer a satisfactory explanation of restructuring:

- Gestalt psychologists observe characteristics of restructuring without providing a computationally or biologically plausible explanation;
- Problem solving methods involving search can succeed in a sudden, surprising manner; but it is not clear how such success could be accompanied by representational change.
- Representation learning and associative learning techniques, typically statistical and/or connectionist, tend to be slow, requiring substantial amounts of data and processing before a good representation is progressively discovered.

The tension is most evident between a *representational change* perhaps best explained nowadays by connectionist learning approaches (Rumelhart et al., 1987), and a *sudden success* best explained by symbolic approaches (e.g. Newell and Simon, 1972). These two approaches also correspond roughly to Type 1 processes (automatic, unconscious and associative) and Type 2 processes (deliberative, conscious and analytic) (Kahneman and Egan, 2011).

In recent insight literature, the controversy has been articulated in terms of "routine/business as usual" Type 1 explanations versus "special process" Type 2 explanations (Bowden et al., 2005; Gilhooly et al., 2015). However, recently there has increasingly been interest in finding out whether insight might emerge from an *interaction* between Type 1 and Type 2 processes, and what that interaction might consist in (Gilhooly et al., 2015; Hélie and Sun, 2010; Weisberg, 2015).

The Reinforcement Learning paradigm is at the crossroads of these influences. It combines exploration of a state space using operators, in the manner of Newell and Simon, and parallel processing of representations in the manner of Rumelhart and McClelland. Furthermore, it offers tools of striking relevance for the study of insight problem-solving, such as the value function (reminiscent of "feelings-of-warmth" ratings), and a central role for surprise in the guise of temporal difference errors (reminiscent of the positive "Aha!" feeling at the time of insight). It therefore offers a theoretical basis from which it is possible to study not the awkward interaction of two radically different processes, but the smooth working of a single system, driven throughout by optimization. Reinforcement Learning is thus uniquely suited to unify the two aspects of insight problem-solving under a single, consistent explanation.

7.2.2 Psychology and neuroscience

Experimental evidence for biological Reinforcement Learning in insight

Part of the appeal of RL is its interdisciplinary credentials: it is both a leading field for decision making in artificial intelligence, and a theory of animal and human decision-making.

RL explains many aspects of classical and operant conditioning in psychology and can be viewed as an extension of associative learning models, such as the Rescorla-Wagner model (Rescorla et al., 1972). Furthermore, RL has biological plausibility: dopamine signals in the brain are thought to implement the temporal difference learning signal used in critic-only and actor-critic algorithms in RL (Montague et al., 1996; Schultz et al., 1997); see Sutton and Barto (2018), chapters 14 and 15, for a summary.

Unfortunately, there have been no studies on a potential role of dopamine in insight. However, the neuroscientific evidence consistently shows that insight is associated with activations of executive control networks in the cortex (PFC, including ACC) typically associated with reinforcement learning (Sprugnoli et al., 2017) and specifically hierarchical reinforcement learning (Holroyd and Yeung, 2012). There is also evidence of activation of the midbrain structures associated with dopaminergic pathways (Tik et al., 2018) during insight, such as the thalamus, VTA, substantia nigra, and striatum - all consistent with a reinforcement learning explanation. There is thus a considerable body of evidence from neuroscience and psychology that is consistent with a reinforcement learning account of psychological insight.

Contemporary theories of biological RL are not sufficient to explain all of the neural phenomena associated with insight (e.g. in the temporal lobes and arguably the hippocampus). Nevertheless, they provide a framework for understanding the role of the PFC and several midbrain structures, which have often been neglected in theoretical investigations of insight, in favor of associative cortices in the temporal lobes (see Bowden and Jung-Beeman (2003a); Shen et al. (2017)).

Experimental evidence for temporal structure learning in insight

The theory presented in this thesis, however, differs substantially from existing reinforcement learning models, in that it treats temporal abstraction as a self-prediction problem (cf. chapter 5). In doing so, it connects RL views with for instance predictive coding (Friston and Kiebel, 2009; Rao and Ballard, 1999), which has also received attention in the context of insight Friston et al. (2017).[2]

The predictive aspect of the intention theory presented in this thesis is interestingly compatible with studies associating insight and sleep. Indeed, sleep[3] has been observed to lead to increased fluency in motor responses, e.g. in learning a finger movements: people instructed to tap their fingers repeatedly in a sequence (e.g. 4-2-3-1-3-3-1-4) were faster and

[2] Whether a convergence of these views is possible remains to be explored - in this thesis I use a simple gradient-based approach to learning, rather than a more theoretically grounded bayesian one.

[3] The precise roles of REM and non-REM sleep are not fully understood, with sometimes inconsistent results (Fischer et al., 2002; Walker et al., 2002).

committed less errors after a night of sleep compared to the same amount of waking time, or to no time passing. Sleep is also associated with a greatly enhanced proportion of insights (from 24% to 59% according to Wagner et al., 2004) for a task consisting in decoding a message, where the last part of the sequence could be decoded faster by noticing that all messages were symmetric – that is, the second half of the message was predicted by the first half. The theory of insight presented in this thesis is loosely consistent with both these results, as it requires transfer of temporally structured information from a reinforcement network to a predictive network, which might take place during sleep (Louie and Wilson, 2001).

There is therefore a substantial body of evidence from psychology and neuroscience in favor of a reinforcement learning theory of insight, and specifically consistent with the temporal-abstraction based intention theory proposed in this thesis.

7.2.3 New evidence presented in this thesis

Reinforcement Learning in AI and insight

In chapter 4, I have investigated the behavior of deep reinforcement learning algorithms on a classic insight problem. Four main results emerged:

1. Prior relevant experience obtained by reinforcement learning led to sudden (by comparison with learning from scratch) learning in the test.

2. Excessive prior experience (overfitting) led to a period of impasse, which was nonetheless followed by sudden success.

3. The value estimates calculated by the agents reflected their large and rapid increase in performance, thus resembling the sudden change of "feelings-of-warmth" reports measured by Metcalfe and Wiebe (1987).

4. The option-based architecture used did not perform as well as "flat" reinforcement learning.

However, moments of rapid learning still required a prolonged period of trial and error (rather than immediate, all-or-nothing success as in psychological insight). The ACI architecture was presented as a tentative answer to the limitations of option-based temporal abstraction.

7.2.4 Objections: disinterested insights and "Uh-oh!"-moments

Are there any obvious limitations to the ACI/intention theory of insight? The theory presented here has focused on insights in a problem-solving context, where agents are promised some kind of reward for solving the problem. But there are situations in which people experience a feeling resembling insight where no obvious extrinsic reward is available - let us call these *disinterested insights*. Another non-positive sort of insight is the “Uh-Oh” moment that we experience when we suddenly realize our situation is worse than expected. Below I discuss the two phenomena in relation to the proposed theory of insight.

Insight is commonly associated with disinterested understanding, for instance of a natural phenomenon. It is not clear, however, that such insights are really disinterested: after all, researchers derive professional advantages and self-satisfaction from achieving improved understanding of the phenomena they study. Whereas there is no extrinsic primary reward, such as food, there may be secondary rewards, such as an improved social status or a perceived personal improvement, which themselves are predictors of future primary rewards; this is well established in psychology under the name of “conditioned stimuli” or “secondary reinforcers”, and is a basic feature of all temporal difference and dynamic-programming reinforcement learning algorithms. Another possible explanation is the presence of *intrinsic* primary rewards (Kaplan and Oudeyer, 2007). There is a substantial literature on combining reinforcement learning with intrinsic motivation rewards (see e.g. Barto (2013)), which I believe would be entirely compatible with the ACI architecture.

Perhaps more challenging for the present theory are negative insights (Gick and Lockhart, 1995) , or “Uh-Oh moments”. In negative insight, subjects undergo a change of perspective accompanied by a negative surprise. This contrasts with the insights discussed in this thesis, which are positive - corresponding to the discovery of an unexpected solution, rather than of an unexpected problem. In fact, the ACI architecture presented in this thesis allows for bad exploratory moves, which would trigger negative insights. However, I do not believe this is a good explanation for “Uh-Oh” moments: indeed, when human beings form a bad intention and become aware that it is bad, they simply switch to another strategy.

The difference can, I believe, be explained by contrasting plans with memories. Whereas plans are chosen, memories are not: one cannot intentionally forget something[4]. A plan is a choice, a decision, whereas a memory is not. Thus “negative insights” corresponds not to the discovery of a strategy, but to the discovery of a hidden state of the present situation which subsequently enters memory. They are not the negative form of insights in problem solving, but simply bad surprises, and (per the present theory) they rely on different neural mechanisms.

[4] At least, not directly via the normal, non-intoxicated operation of the nervous system.

7.2.5 Special-process or business as usual?

A recurrent debate in psychology concerns the process underlying insight, by contrast with analytic problem solving. Is there a special process for insight? Or does it occur as part of analytic problem solving (Weisberg, 2015)?

The proposed architecture suggests that insight problem-solving does not fit neatly in either the special-process or business-as-usual categories. In the view suggested in this work, the same cognitive construct, intentions, underlies all problem-solving; there is no special process dedicated to insight as opposed to analytic problem-solving.

There is nevertheless a distinction between analytic and insightful problem solving: in analytic problem solving, the agent follows an established strategy which solves the problem step-by-step; little to no surprise occurs, and little to no learning takes place. In contrast, insight implies a positive surprise. One may describe problem-solving as more or less insightful on a graded basis, by considering the extent to which representational change (restructuring, or redistribution of activation) and evaluative surprise ("Aha!" or large positive temporal difference errors) occur.

7.2.6 Summary

The previous five chapters of this thesis have consisted in preparing and conducting an interdisciplinary investigation of the insight phenomenon. I have proposed that deep reinforcement learning is the right framework for explaining insight, or at least, the best so far: this was supported by the characteristics of deep reinforcement learning, by a review of psychological and neuroscientific evidence, and by A.I. simulation experiments. I have further suggested that insight requires a new type of temporal abstraction, which I call intentions. This is supported (in a negative manner) by a failed effort to achieve insight with a more traditional option framework, and reflection about its causes. The proposed solution has theoretical support from information theory and is consistent with evidence from sleep studies of insight. Finally, I have proposed explanations for the manner in which phenomena similar to insight are accounted for by the theory.

7.3 Insight, Intentions, and Creativity

This thesis has so far focused on insight; however, since chapter 1, insight was seen as interesting not only for its own sake, but also for its relationship to creativity. In this subsection I connect the findings on insight to the more general study of creativity, presenting its findings within the context of various established frameworks of creativity research.

7.3.1 Intentions, information theory, and creativity

The ACI architecture is in part based on seeking value, and in part on making (self-)predictions. Its heuristic justification is to avoid redundancy in decision making, an information-theoretic measure. Thereby the ACI architecture, and the theory of intentions it embodies, find their place within a body of work on creativity and general artificial intelligence that includes for instance the “Formal Theory of Creativity and Fun and Intrinsic Motivation” of Schmidhuber (1991, 2010), and the “Information Dynamics of Thinking” model of Wiggins (2018); Wiggins and Forth (2015). The ACI architecture, like Wiggins’ model, attributes a central role to time, and is motivated by information-theoretic aspects of prediction. Compared to it, however, the ACI architecture does not have a symbolic/linguistic component; this symbolic component might be what gives human creativity an edge over its animal counterparts.

7.3.2 Insight within several models and taxonomies of creativity

The components of creativity have been labeled “4 Ps” (person, process, product, press) by Rhodes (1961) and “5 As” (actor, action, artifact, audience, affordances) by Glăveanu (2013), and most recently “7 Cs” (creator, creating, creation, collaboration, context, consumption, curricula) by Lubart (2017). My focus has been decidedly on the “process” (respectively action, affordance, creating in the other frameworks) element of creativity. Arguably I have also considered the artifact or product, if one thinks of a motor behavior or skill as a creative product. But what has received the least attention are the social elements of creativity: the press, the audience, collaborations, consumption. The agents presented in this thesis are, from their point of view, alone in the world - they discover and create efficient solutions by themselves, for themselves, without acknowledging any other agents. They behave in novel/original and valuable/appropriate ways in a self-centered sense.

Such acts of creativity are common in what are called “mini-c” and “little-c” types of creativity (Kaufman and Beghetto, 2009): that is, creativity in the course of normal development and education, and in reaction to the problems of daily life. Many of us “reinvent the wheel” for our own usage, so to speak, when facing problems for which there is a known solution of which we have never heard. This even takes the form of a carefully organized and monitored process in the course of education. Children, teenagers and young adults are made to arrive at creative insights in a guided manner, through the gradual, systematic acquisition of skills and by applying them on a range of problems. This occurs for instance in a highschool mathematics class or during a test, but arguably also in literature or arts classes in which skills are taught in order to be used for “problems” that allow for an open-ended class of “solutions”. The model presented in this thesis focuses on the acquisition of such skills and their explorative use. It is therefore relevant to these creative behaviors, whether encouraged and controlled within a school environment, or spontaneous in the course of daily life.

In chapter 1, I have offered some justification for a distinction reminiscent of that between little-c/mini-c creativity and eminent big-C or H-creativity. I explored the meaning of “creativity” in terms of the ambiguity of most definitions of creativity with respect to context (novelty/originality) and norms (value/appropriateness). I chose to focus on self-centered creativity because whereas one can imagine individual-centered little-c existing in the absence of society-centered big-C, the opposite seems less likely.

But this dichotomy between little-c and big-C creativity[5] has been criticized on the grounds that it divides what are profoundly related phenomena (Runco, 2014a). In what way are they related? Firstly, there is little to no evidence of two distinct cognitive abilities - one for “little-c” and one for “big-C” creativity – as pointed out by Merrotsy (2013); Plucker and Beghetto (2003); Runco (2014a)). The latter says: “[Big-C and little-c] are different, but not because of the creativity involved. They all depend on the same creativity. They differ mostly in things that occur after the creative act.” This suggests at the very least that if the two are in some way cognitively distinct, they nevertheless have much in common.

If indeed we consider creativity not merely in the context of post-enlightenment liberal societies (which have made of creativity a slogan), but in the broader context of all human societies, we find that creative practices are generally rooted in the traditional acquisition of skills which form much of a society’s culture (Ingold, 2014; Rosaldo et al., 1993). From the Chinese calligrapher to the jazz musician (Ingold, 2014), creativity finds its source in acquired skills; one might also mention the PhD student acquiring the practical skills of scientific research. But in focusing on the creativity displayed by the finished product, we tend to ignore the role of the routine cognitive processes at their source: *“the obsession with novelty implies a focus on final products and a retrospective attribution of their forms to unprecedented ideas in the minds of individuals, at the expense of any recognition of the form-generating potentials of the relations and processes in which persons and things are made and grown”* (Ingold, 2014, p. 124). Thus anthropologists, by putting contemporary western society in the wider context of human societies everywhere, show how strong the link is between the creativity of the “genius” artist and that of the child or even the animal. Both are rooted in learning, doing, imitating, and improvising. Intentions, conceived as learned from a predictive relationship between past and future decisions, put the roots of the sudden

[5] See Merrotsy (2013) for an investigation of the genesis of these two terms and concepts.

creative leaps of insight in the everyday practice of even the “genial” scientist and artist, and thereby connect everyday creativity with the originality of genius.

In this subsection I have discussed the relationship between the work undertaken in this thesis, and some of the major frameworks offered by the creativity literature. Unsurprisingly, this work is most directly relevant to investigations of little-c, everyday creativity; but parallels extend to “big-C”, eminent or historical creativity... which ultimately must be based, at least for its purely cognitive component, on the same processes as little-c creativity.

7.3.3 Intentions, discourse, and creativity

In the discussion of intentions so far, I have steered clear of issues of “intentional” action: really “meaning” to do something. This was largely because it is not clear that this is relevant with respect to insight. Furthermore in the context of philosophy and psychology, such arguments are held with views to speculative topics such as free will or moral issues such as personal responsibility (see e.g. Haggard and Libet (2001)), which are far removed from the main contributions of this thesis. Nevertheless, the intention construct presented here is related to the common-sense meaning of intentions also in this respect: after introducing an internal state for the agent which defines its long term behavior, one may ask whether the results of an agent’s actions are, or are not, consistent with their long-term intentions. This might provide a basis for distinguishing a stroke of luck from a stroke of genius[6].

To illustrate the matter of intentional action, I will recount a famous example from philosophy. In her book *Intention*, Anscombe (1957) offers the following example: a person is replenishing a house’s water supply with a poisonous liquid. If the person is not aware that the liquid is poisonous, we might say that they are intentionally replenishing the water supply, but not that they are intentionally poisoning the occupants of the house. Intentions, by introducing a mental state corresponding to the actions, which has implications for how the agent would deal with contingencies (e.g. remarking that the liquid has a peculiar smell), offers a basis for making the distinction. Thus the same action, described in different ways, can be intentional or unintentional.

Anscombe’s interest is related to considerations of moral responsibility; likewise Bratman discusses the (un)intentional collateral damages of strategic bombings. This is nonetheless related to creativity: in the same way that one may ask whether somebody intentionally committed a crime, one may also ask whether an artist or scientist intentionally created

[6] A lucky agent would be constantly surprised by its own success as its strategy unfolds, whereas for a genius, work following the initial insight ought to be fairly routine (see e.g. Poincaré’s discussion of insight in appendix B).

a masterpiece[7]. Indeed, a common concern with computationally creative algorithms is their supposed lack of intention, motivation, and purpose, and the question of whether the "intentionality" behind the work is that of the program or that of the programmer (Charnley et al., 2012; Colton et al., 2011; Cook and Colton, 2011). Such concerns have led to producing creative agents which are capable of communicating about their creations (Augello et al., 2016; Lopes et al., 2016), thereby "proving" that they made them with understanding or intent and encouraging the perception of creativity in a human spectator or consumer of the creative product.

Intentions offer a technical solution to this issue: a creative agent can be said to intend the final product *under a certain description*, to the extent that the intention was relative to achieving the features emphasized in that description. This is just a sketch: there is a great deal of innovation required before a general reinforcement learning agent is capable of producing creative artwork or science at the level of the specialized artists of computational creativity. Nevertheless, intentions as described in this thesis offer a promising basis for answering questions of artistic authorship and merit in computational creativity.

Thus it turns out that intentions, the technical solution to a reinforcement learning exploration problem, relate to many important considerations in the context of creative achievement and creativity research including computational creativity. Intentions give a basis for attributing to an agent the credit for its creative achievement: depending on whether or not the effects of an action are consistent with the agent's intention, we can attribute merit to the agent for its creative product.

7.3.4 Society and creativity

In recent years, creativity research has increasingly focused on the social aspect of creativity: the manner in which creative achievement and creative processes are embedded in social interaction and in socio-cultural practices, expectations, and judgements (Csikszentmihalyi, 2014; Glăveanu, 2018); recently, a manifesto promoting socio-cultural perspectives on creativity was authored by several of the most prominent creativity researchers (Glăveanu et al., 2019). This thesis has focused on insight and creativity from the perspective of an isolated individual; here I will briefly discuss its relevance in a social context.

[7] It is not clear whether or not similar standards are also applied to artists. In literary criticism, the intention of the author has been called irrelevant for the value and interpretation of the piece (Wimsatt and Beardsley, 1946) (though this is increasingly questioned, for instance by Farrell (2017)). But that may be because there is comparatively less doubt about whether artists intend the aesthetic aspects of their work, compared to computers. By contrast, much of conceptual art (for instance Duchamp's famous "Fountain") derives its value mostly from the creative process and intentions of the artist (Colton, 2008)

The agents discussed in this thesis are a-cultural: they evolve in a world which has physical dynamics (push the box to make it move), but is devoid of social dynamics. However, this is a result of their simulated environments, more so than of their own architectures. The reinforcement learning mechanisms used are general enough to allow for learning even social behavior. That is, if a reinforcement learning agent is placed within a "society" of such agents, emergent social phenomena occur - a setting explored within a rich literature, e.g. Foerster et al. (2016); Littman (1994); Tan (1993). The work presented in this thesis can be seen as the foundation for an extension to the multi-agent setting, in which group dynamics akin to a simplified "culture" may emerge; see e.g. Mordatch and Abbeel (2018); Noble and Franks (2012).

Which kinds of group dynamics, or emergent processes, can a multi-agent reinforcement learning setting say something about? At least three: the emergence of cooperation; the importance of communication, including manipulation; and the role of education.

Cooperation: In chapter 1, I have argued that what distinguishes the concept of creativity in its social sense, as opposed to its individual sense, is the source of the context and norms. The lonely agent behaves in an original manner by contrast to its own past; the social agent is original by contrast to its peers. The socially celebrated "creative" is praised for valuable contributions to the group, rather than for their creative tax reports. Thus "creative machines", in the socio-cultural view, ought to combine aligned values (as enforced by a reward function) with original behavior. In a multi-agent setting, this "alignment of values" is dependent on the emergence of cooperative behavior, which has been studied e.g. by Lowe et al. (2017); Tan (1993).

Communication: if creativity is to be measured by recognition within a field, as suggested notably by Csikszentmihalyi (2009), and as implemented in practice in fields such as academic research and artistic achievement via systems of awards, or citation counts, then the communicative behavior accompanying (or constituting part of) the creative product is essential. There is a risk, here, of allowing skills in influence or manipulation to replace individual creativity as the cognitive ability underlying "socio-cultural" creativity: *"impact, fame, eminence, reputation, and accomplishment are often present when creativity is absent"* (Runco, 2014a). It is in part due to such concerns that critical theorists have sought to exclude considerations of intention from literary criticism (Wimsatt and Beardsley, 1946). Notwithstanding such worries, the ability to communicate about and promote a creative product is undoubtedly of major societal value, and ought to be recognized and encouraged alongside the ability to make that product. Thus research aiming at promoting socially beneficial forms of creativity cannot ignore communication. In this respect, intentions are promising concept. Indeed, they are an abstraction for the features of future behavior that are valuable to the agent: the answer to "what are you doing that for" is whatever is encoded by an agent's intention: an agent with intentions is thus well-positioned to communicate about and promote its creation.

Education: Recall that the 7 C's framework of Lubart (2017) included a category called "curricula". The artificial pigeons of this thesis owed their creativity not only to their own architecture, but also to a training curriculum designed to teach them relevant skills. Large swathes of the creativity research literature have sought to find ways to improve the creativity of schoolchildren or employees by giving them relevant training. Whereas the present work cannot claim without ridicule to model the learning abilities or environment of human beings, it can contribute to a better understanding of which types of learning increase or hinder creativity. For instance, in chapter 4, we found that overfitting led to reduced performance on the test task, resembling the psychological phenomenon of negative transfer/the einstellung effect (Luchins, 1942); whereas the intention framework suggests that the ability to formulate long-term "plans" is essential to acquiring creative fluency. Such results may be inspiring to researchers investigating and testing new teaching methods designed to encourage creativity.

Although this thesis has focused on establishing a foundation for individual creativity, the results have relevance to work conducted within a socio-cultural perspective. The philosophical groundwork of chapter 1 allows us to relate the individual perspective adopted in chapters 2 to 6 with the socio-cultural perspective of other creativity researchers, such as Glăveanu et al. (2019). In particular, we find that intentions (in general, rather than specifically as they are found in the ACI architecture) are likely relevant in insuring cooperation, communication, and in improving our understanding of educative challenges.

7.4 Conclusion

7.4.1 Summary

This chapter has sought to summarize the main findings of this thesis, putting them in context with the relevant literature in the multiple fields concerned. I first discussed the ACI architecture in the context of reinforcement learning, then the combined evidence for an intention-based theory of insight, and finally I covered the place of the ACI architecture, intentions, and insight, in creativity research writ large.

In a reinforcement learning context, the ACI architecture is an original approach to temporal abstraction. Although additional work is required to test whether it can be competitive against existing techniques, it offers original theoretical perspectives. Future implementations of intention-based architectures may be combined with recent algorithms for learning to generate structured output, potentially increasing their efficiency.

Multiple strands of evidence support the intention theory of insight. Intentions offer an account of insight based on a reinforcement learning paradigm which is placed at the crossroads of psychological paradigms (cognitive and associationist). Biological reinforcement learning offers a plausible explanation for the pattern of activation in the prefrontal cortex during insight. Furthermore, the meta-cognitive components of insight are both explained by the intention framework: a temporally extended change in representations (restructuring) is a consequence of the change in intention, whereas the "Aha!" moment corresponds to the evaluation of a new positive intention by the critic. Finally, evidence from sleep studies corroborates the essential importance of temporal sequence in learning intentions and thereby acquiring the capacity for insight on a given problem. Taken together, these separate pieces of evidence make up the case for the ACI model, and for an intention theory of insight.

In chapter 1, I justified this investigation of insight for its promise to explain aspects of creativity. How did this turn out? Insight is an individual cognitive phenomenon, and is therefore most directly relevant to creativity relative to an individual - the creativity of the learning child and of the practical problem-solver. If the intention theory is correct, it suggests that everyday creativity and professional creativity are rooted in the acquisition of relevant skills, best achieved by repeated practice; but also that excessive practice without variation leads to overfitting (in machine learning terms) or the "einstellung effect" (in psychological terms). There is most likely a "sweet spot" in between excessive and insufficient training, such that the creative student acquires a relevant range of skills without becoming set in rigid ways.

But everyday creativity and genius are likely related. Research on the latter typically takes into account the social recognition of creative feats. A common factor in this recognition is the attribution, or not, of merit for the creative contributions – is it due to luck or to talent? Intentions may allow for answers to such questions.

7.4.2 Contributions

This chapter extended the discussion of the RL model of insight presented gradually from chapters 2 to 6. It also discussed the relationship of intentions with some of the existing temporal abstraction and exploration techniques in RL, and finally presented intentions as a component of creativity.

7.4.3 Bibliographical remarks

- **Temporal abstraction in RL:** The main framework for temporal abstraction in RL continues to be the Option framework of Sutton et al. (1999), although work based on earlier proposals for HRL, such as feudal RL (Dayan and Hinton, 1993) continues to be inspiring (Vezhnevets et al., 2017). Much of this work (HAM, Parr (1998); MAXQ, Dietterich (2000a)) focuses on creating a hierarchical structure, similar (sometimes quite explicitly as in the "programmable RL" of Andre (2003)) to the routines and subroutines of a programming language. There is very little work investigating temporal abstraction in any form other than hierarchical (see e.g. Jong et al. (2008), which largely equates the two). For a formal discussion of state abstraction in the discrete MDP case, with some relevance to temporal abstraction, see Li et al. (2006). For an older but still relevant review, see Barto and Mahadevan (2003).

- **Alternative models of insight:** Appendix C presents brief summaries of some alternative models of insight, including the "switch between options" model of Ohlsson (2011), the "implicit vs. explicit cognition" model of Hélie and Sun (2010), and a few comments on the active inference approach of Friston et al. (2017), among others. This chapter has focused on listing the evidence for the proposed model of insight, without proposing an extended comparison with alternative models. This is in part because the different models are not necessarily incompatible; but because they make use of different theoretical frameworks, it is complicated to link any two of them together. For instance, the behavior of a neural network undergoing insight via a radical change in intentions might be aptly described as a dynamical system undergoing a phase change, as in the work of Stephen et al. (2009); but spelling out exactly how a reinforcement learning, "engineering" approach of insight maps into a descriptive dynamical systems account is beyond the scope of this thesis.

- **Intentions and creativity:** the work connecting creativity with intentions or at least intentional behavior occurred primarily in the field of computational creativity, where establishing, or at least suggesting the validity of the authorship of a work is more pressing than in other domains. Much of this work was conducted by Colton and colleagues (Colton et al., 2011; Guckelsberger et al., 2017; Pease and Colton, 2011b, ...). Within psychology and neuroscience, however, I am not aware of any substantial connections between creativity and intentions, or even between creativity and prospective memory. In philosophy, the philosophy of creativity, the philosophy of art, and the philosophy of actions and intentions have remained largely isolated from one another (but see the recent book by Farrell (2017)).

Conclusion

General summary

This investigation began as an attempt to understand creativity, or some aspect of it, through the combined lenses of psychology and machine learning. In bringing this project to fruition, I have analyzed the plurality of the concept of creativity (chapter 1). I have narrowed the focus to a study of insight problem-solving (chapters 2), which I decided to model using reinforcement learning (chapter 3). This was because of promising, unexplored analogies between them: notably the sense of progress characteristic of insight and the value function in reinforcement learning; the combination of representation learning and problem solving in reinforcement learning, and the phenomenon of restructuring in insight.

Seeking to produce new experimental evidence for these analogies, in chapter 4 I have reproduced in simulation a classic insight experiment, demonstrating the extent to which animal learning resembles the behavior of reinforcement learning agents. In particular, the first experiment demonstrated the efficiency of shaping techniques in RL, showed transfer, and revealed an "einstellung"-like effect in which excessive training misled the agent. However, this experiment also revealed the gap in adaptability between an intelligent animal (a pigeon) and the artificial pigeons: the standard (flat) reinforcement learning agent did not achieve sudden insights, but only accelerated learning which still required considerable amounts of trial and error. I also investigated whether options (obtained by training separate networks on different aspects of the task) could produce behavior resembling insight; however, this approach was less efficient than a "flat" (non-hierarchical) reinforcement learning agent.

In chapter 5, I proposed a novel model of insight based on self-prediction called the Actor-Critic-Intention (ACI) architecture, which aims to discover the sort of generalizable temporally extended behavior needed to explore efficiently. The model is capable in principle of the main features of insight: "Aha!" moments and restructuring; the latter being achieved by exploring not only at the level of actions, but also at the level of representations.

An important contribution of this thesis, then, is the introduction of a new approach towards temporal abstraction, the ACI architecture, which seeks to reduce the *redundancy* *of decision-making* across time. The relationship of the ACI architecture with the study of intention in philosophy, psychology, and artificial intelligence was discussed in chapter 6. The proposed architecture resembles closely these other approaches, and has capabilities that are promising for the field of reinforcement learning.

The ACI architecture, like insight itself, relates most directly to creativity at the level of an individual. But it is also relevant, albeit more loosely, to the field of creativity research in general. In particular, intentions may be relevant for assessing the relationship between an author and its creation, which is an important question in computational creativity. These and other considerations of the relationship between intentions, Reinforcement Learning, insight, and creativity, are discussed in chapter 7.

I have thus explored the nature of creativity through one of its manifestations: the phenomenon of insight. Following several different threads – in psychology, neuroscience, artificial intelligence, even philosophy – I have proposed a theory of insight based on *intentions* which weaves them all together.

Directions for future work

If this thesis offers some tentative answers, it also opens up many questions and directions for future work.

In reinforcement learning, experiments are needed to test the intention architecture, both on small and large scale problems. The latter especially might first require theoretical progress in structured prediction, which at the time of writing is an active area of research. A related, promising direction is categorical learning - made possible by techniques such as the Gumbel-Softmax (Jang et al., 2016). Finally, this thesis only presented informal justifications for the manner in which the predictive network could reduce redundancy. I believe a formal theoretical justification of the approach is possible using information theory or Bayesian approaches.

This thesis claims that insight is made possible by self-prediction; this is probably testable in a psychological experiment, though I have not looked into experimental set-ups susceptible to verify this. Likewise, the theorized relationship between insight and reinforcement learning ought to be confirmed, to validate the results obtained by Tik et al. (2018), and to confirm the presence of a large temporal difference error, affecting dopaminergic structures, at the moment of insight.

The ACI architecture, based on self-prediction, aims to reproduce creative insight, but functions as a model of problem solving based on experience, with several interesting properties noted in chapter 6. Could it serve as a component of a larger model of creative thinking, experimenting, reasoning? Is some implementation of *intentions* one of the secret ingredients in the recipe for intelligence?

Main contributions

- (Building upon the existing literature,) an analysis of the conceptual space of creativity along two dimensions: a context (for originality) and a norm (for effectiveness).
- A flat reinforcement learning model of insight, based on transfer and overfitting effects.
- The ACI architecture, which is both a model of insight and a novel approach to temporal abstraction in reinforcement learning.

Appendix A

Some insight problems

This appendix is a compendium of the problems used to study insight in psychological experiments on humans and animals. Most problems which have been used extensively, or with which important results were obtained, are presented. Solutions are provided for most problems, but readers with time to spare are encouraged to solve the problem themselves, thereby experiencing "Aha!" moments.

A.1 Problems for humans

A.1.1 Remote Associates Test and Compound Remote Associates

The Remote Associates Test (RAT), first proposed by Mednick (1962), consists of presenting 3 words, and asking the subject to find a fourth that is related to all 3. For instance the solution to *falling/actor/dust* is *star*.

In the neuro-imaging literature, a slightly more specific version of the test is often used, called the Compound Remote Associates (CRA), and proposed by Bowden and Jung-Beeman (2003b). This test is similar to the RAT, but focuses on compound words and expressions (this excludes associations such as "actor/movie star"). Below are a few examples of increasing difficulty:

1. nuclear / feud / album

2. fish / mine / rush

3. dust / cereal / fish

4. note / chain / master

5. over / plant / horse
6. land / hand / house

The proportions of US university students solving these items in 30 seconds are, respectively: 1: Family (85%); 2: Gold (75%); 3: Bowl (49%); 4: Key (26%); 5: Power (10%); 6: Farm (0%) (Bowden and Jung-Beeman, 2003b).

A.1.2 Logogriphs, riddles, ambiguous sentences

Consider this "riddle" (imagined for the purpose of this example): "It has gold, but it is not rich": a goldfish. This works because the word "goldfish" has "gold" in it, and a goldfish probably cannot be rich. The chinese language is written using logograms which can be combined, allowing for similar riddles, in which one may refer to both the thing and the corresponding written word. Such riddles were used starting with Qiu et al. (2006) to investigate insight.

In a related set-up, Luo et al. (2004b) used japanese riddles such as: "What is the animal that can win three times?" In Japanese, "salamander" is read as "san-shou-uo", with "san" having the same pronunciation as "three" and "shou" the same pronunciation as "win."

In another set-up by Luo et al. (2004c), one of the very first fMRI studies of insight, sentences from Auble et al. (1979) are presented to subjects; then a cue is presented. Subjects were considered to have achieved "insight" if the cue caused them to suddenly understand the sentence. Below are two examples sentence/cue pairs:

- The betting was halted because the wheel was spun/roulette
- The audience cheered because the five balls rotated/juggler

In a study by Mai et al. (2004), participants must solve riddles:

- "The thing that is very old, but very valuable" "antique"
- "Though they veil your eyes, you see clearer" "glasses"

A.1.3 Prototypes

In this design used in fMRI studies, participants read about a "prototype" solution (e.g. the structure of shark skin) and then are presented with a related or unrelated technical problem (e.g. how to make an efficient submarine hull). Participants are expected to experience insight if the prototype and the technical problem are related. Variations of this design were used by e.g. Dandan et al. (2013) and Luo et al. (2013).

A.1.4 Brainteasers

In one study of Sheth et al. (2009), participants were given brainteasers such as this:

> **Problem:** There are three on-off light switches on the wall of the first floor of a building. One of the switches controls an incandescent bulb in a lamp on the third floor of the building. The bulb is initially off. The other two switches do not control the bulb or anything else (they are disconnected). You are allowed to toggle the switches as many times as you want and for as long as you want. How can you find out which one of the three switches turns the light bulb on and off? The only constraint is that you can walk only once to the third floor to check on the light bulb.
>
> **Hint:** Keep one of the switches on for an hour and then turn it off.
>
> **Solution:** You turn the first switch on and leave it on for an hour. Then you turn it off and turn the second switch on, leave the third switch in the off position, and you go upstairs. If the bulb is on, then it's switch number two, which is the one that's on. If the bulb is off and it's cold, then it is switch number three (the switch you never touched) that controls that light. If the bulb is off but it's hot, then it is switch number one.

A.1.5 The number reduction task

This set-up was used primarily in investigations of the role of sleep in facilitating insight (Wagner et al., 2004), as well as in one EEG study (Lang et al., 2006).

Participants are instructed to apply a certain methodology to "reduce" a number according to certain rules. They are presented with a stimulus-number consisting of only the digits 1, 4, and 9, such as:

s=11449494

The first digit of the response-number depends on the first 2 digits of the stimulus number, according to these rules:

1. If these digits are identical (e.g. 1 and 1), then the response is the same digit (also 1).
2. If these digits are different (e.g. 1 and 4), then the response is the remaining digit (9 in this case).

Therefore:

s=11449494

r=**1**...

Then the participants must continue producing response digits by using the next digit in the stimulus-number (in this case 4) together with the last found response-digit (in this case 1). According to rule 2, 4 and 1 make 9:

s=11449494

r=1**9**...

And so on:

s=11449494

r=19**1**...

Continuing this procedure one eventually gets to the end of the sequence, producing the full response-number:

s=11449494

r=191441**9**

However, unbeknownst to the participants, the stimuli are rigged to have a certain property: the last three response-digits 5 to 7 are the mirror image of the previous three. In the example “419” is the “mirror image” of “914”.

Given many trials with different stimulus-numbers, participants may come to have the insight of this regularity; this realization comes much more easily when the second block of trials is separated by a night of sleep, compared to waking time.

A.1.6 The mutilated checkerboard

This problem was used in a classic article on insight by Kaplan and Simon (1990). It is fairly difficult. The problem is:

> Suppose a standard 8×8 chessboard has the top right and bottom left corners removed, leaving 62 squares. Is it possible to place 31 dominoes of size 2 by 1, so as to cover all of these squares?

One should recognize that because the chessboard squares are alternatively black or white, a domino by necessity covers one white and one black square. Therefore 31 dominoes must cover 31 white squares and 31 black squares; but the mutilated chessboard contains instead 32 white squares and 30 black squares, so this is impossible.

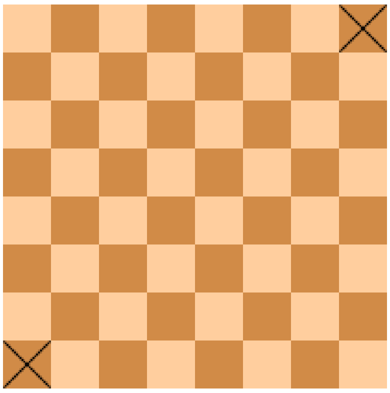


Fig. A.1 The mutilated chessboard

A.1.7 Duncker's candle problem

The original formulation is found in the monograph "On problem-solving" by Duncker and Lees (1945):

> *The "box problem":* On the door, at the height of the eyes, three small candles are to be put side by side ("for visual experiments"). On the table lie, among many other objects, a few tacks and the crucial objects: three little pasteboard boxes (about the size of an ordinary matchbox, differing somewhat in form and color and put in different places). *Solution:* with a tack apiece, the three boxes are fastened to the door, each to serve as platform for a candle (...).

Depending on whether the boxes were initially full of other objects or empty, Duncker observed, in preliminary experiments, respectively 100% and 42.9% of solving success.

A.1.8 The altar window

This geometrical problem is from "Productive Thinking" by Wertheimer and Wertheimer (1959).

> Painters are at work, painting and decorating the inner walls of a church. Somewhere above the altar there is a circular window. For decoration, the painters have been asked to draw two vertical lines tangent to the circle, and of the same height as the circular window; they were then to add half circles above and below, closing the figure. This area between the lines and the window is to be covered with gold. For every square inch, so and so much gold is needed.

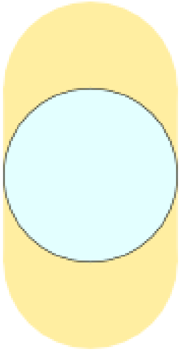


Fig. A.2 The altar window

How much gold is needed to cover the yellow area in figure A.2 - or in other terms, what is the area of the yellow area, assuming a window diameter of 1?

Wertheimer recounts the experiences of several young children with the problem, one of which exclaims: “How blind I was! How simple this is! (...) Excellent problem!”

A.1.9 The 8-coin problem

8 coins are positioned in some initial configuration.

Problem: Can you move two coins such that each coin touches exactly three coins?

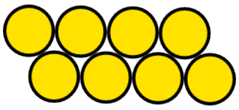


Fig. A.3 8-coin problem: initial configuration

Hint: Experimenters may try different initial configurations to make the problem more or less difficult. Figure A.3 presents one of the more difficult initial positions. the problem is easier if the coins are presented, for instance, in two groups of four coins.

Solution: the subject must form two pyramids of four coins, thus exploiting the vertical dimension.

A.1.10 Binarized images

A few experiments, such as that by Minami et al. (2014), have investigated insight using binarized images. The objective for the participant is to identify the contents of a picture that was reduced to two tones, usually black and white. Figure A.4 is an example of this.

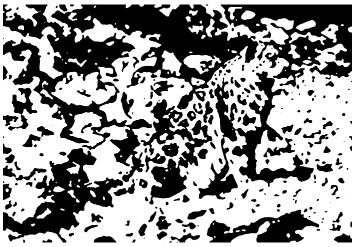


Fig. A.4 Binarized image

A.1.11 Rebus puzzles

MacGregor and Cunningham (2008) have proposed the use of rebus puzzles as insight problems. For instance:

- PAINS
- poPPd
- $\frac{iiii}{ooo}$

Solutions: (1) growing pains; (2) two peas in a pod; (3) circles under the eyes.

A.1.12 Magic tricks

Danek et al. (2013) investigated insight by showing participants videos of magic tricks and asking them to discover the trick. Because magic tricks are designed to induce people in error, distracting them from a simple solution and encouraging a supernatural interpretation, they constitute a reservoir of "insight problems". Repeated showings of the video at a slowed-down speed can serve as hints.

A.1.13 Other problems

Problems already discussed extensively in chapter 2 are not discussed here. This includes the 9-dot problem and matchstick arithmetic problems.

A.2 Problems for animals

A.2.1 Chimpanzees, pigeons, elephants, and bananas

Experiments by Köhler (1921) and their replications by Birch (1945) and Epstein et al. (1984) were extensively discussed in chapters 2 and 4. Below is the description of Köhler's experiment by Epstein:

> Köhler placed a banana out of reach in one corner of a room and a small wooden crate about 2.5 m from the position on the floor beneath it. After a number of fruitless attempts by all six chimpanzees in the room to jump for the banana, one of them paced for several minutes, then suddenly moved the box half a metre from the position of the banana "and springing upwards with all his force, tore down the banana". Both research and theory suggest that chimpanzees will not solve problems of this sort if they have not first had certain experiences. We speculated that two behaviours had to have been acquired: pushing objects towards targets and climbing on objects to reach other objects. Since a pigeon normally does neither, it seemed an ideal candidate to test the contribution that previous learning might make to success in this problem.

Since then, the setting has been re-used successfully on pigeons by Cook and Fowler (2014), and even elephants (Gillian Hill, personal communication).

A.2.2 Betty: bird genius

In chapter 2 I briefly evoked Betty the crow. This is her story (Weir et al., 2002):

> a captive female [Betty] spontaneously bent a piece of straight wire into a hook and successfully used it to lift a bucket containing food from a vertical pipe (Fig. 1A). This occurred on the fifth trial of an experiment in which the crows had to choose between a hooked and a straight wire and only after the hooked wire had been removed by the other subject (a male). The animals had prior experience with the apparatus, but their only previous experience with pliant material was 1 hour of free manipulation with flexible pipe-cleaners a year before this experiment, and they were not familiar with wire (6).

Following this, the ability of that particular New Caledonian crow to adapt its behavior to novel problems was verified (Weir and Kacelnik, 2006):

> Betty quickly developed novel techniques to bend the material, and appropriately modified it on four of five trials when unbending was required. She did not mechanically apply a previously learned set of movements to the new situations, and instead sought new solutions to each problem.

A.2.3 The crow and the pitcher

The fable of the Crow and the Pitcher (traditionally attributed to Aesop) goes as follows (Townsend, 1871):

> A Crow perishing with thirst saw a pitcher, and hoping to find water, flew to it with delight. When he reached it, he discovered to his grief that it contained so little water that he could not possibly get at it. He tried everything he could think of to reach the water, but all his efforts were in vain. At last he collected as many stones as he could carry and dropped them one by one with his beak into the pitcher, until he brought the water within his reach and thus saved his life.
>
> Necessity is the mother of invention.

The problem of accessing water (or some floating reward) from a recipient, by dropping pebbles or other sinking objects in order to raise the level of water, has been used as an experimental setup to test animal intelligence, interpreted as a proof of either “creative problem solving” or “causal reasoning”. The test has been successfully passed by corvids such as rooks and crows (Bird and Emery, 2009; Jelbert et al., 2014), apes such as orangutans, chimpanzees, gorillas, and human children (Hanus et al., 2011; Mendes et al., 2007), and most recently in raccoons (Stanton et al., 2017). Depending on the exact behavior displayed by the animal in solving the problem, this has been taken as evidence of “creative problem-solving”.

Appendix B

Mythical and historical insights

This appendix covers legendary accounts of major discoveries, some of them arguably by means of a single "insight". These accounts are sometimes of dubious authenticity. However, they have had a considerable influence on our cultural perception of insight; and they constitute part of the cultural background which makes toy "insight problems" so interesting to psychologists. By making them explicit, we can perhaps better understand the field.

Note that some of these stories are not solely about insight, but also about restructuring, "productive thinking", or even creativity writ large.

B.1 Archimedes or the first "Eurêka!"

The earliest known account of Archimedes' famed "Eureka!" moment is from the Roman Marcus Vitruvius Pollio in book IX of his "De Architetura libri decem", written circa the 1st century B.C., some two centuries after the purported events. I quote below the translation by Morris H. Morgan (Pollio et al., 1914):

> In the case of Archimedes, although he made many wonderful discoveries of diverse kinds, yet of them all, the following, which I shall relate, seems to have been the result of a boundless ingenuity. Hiero, after gaining the royal power in Syracuse, resolved, as a consequence of his successful exploits, to place in a certain temple a golden crown which he had vowed to the immortal gods. He contracted for its making at a fixed price, and weighed out a precise amount of gold to the contractor. At the appointed time the latter delivered to the king's satisfaction an exquisitely finished piece of handiwork, and it appeared that in weight the crown corresponded precisely to what the gold had weighed.

> But afterwards a charge was made that gold had been abstracted and an equivalent weight of silver had been added in the manufacture of the crown. Hiero, thinking it an outrage that he had been tricked, and yet not knowing how to detect the theft, requested Archimedes to consider the matter. The latter, while the case was still on his mind, happened to go to the bath, and on getting into a tub observed that the more his body sank into it the more water ran out over the tub. As this pointed out the way to explain the case in question, without a moment's delay, and transported with joy, he jumped out of the tub and rushed home naked, crying with a loud voice that he had found what he was seeking; for as he ran he shouted repeatedly in Greek, "Eurêka, eurêka".
>
> Taking this as the beginning of his discovery, it is said that he made two masses of the same weight as the crown, one of gold and the other of silver. After making them, he filled a large vessel with water to the very brim, and dropped the mass of silver into it. As much water ran out as was equal in bulk to that of the silver sunk in the vessel. Then, taking out the mass, he poured back the lost quantity of water, using a pint measure, until it was level with the brim as it had been before. Thus he found the weight of silver corresponding to a definite quantity of water.
>
> After this experiment, he likewise dropped the mass of gold into the full vessel and, on taking it out and measuring as before, found that not so much water was lost, but a smaller quantity: namely, as much less as a mass of gold lacks in bulk compared to a mass of silver of the same weight. Finally, filling the vessel again and dropping the crown itself into the same quantity of water, he found that more water ran over for the crown than for the mass of gold of the same weight. Hence, reasoning from the fact that more water was lost in the case of the crown than in that of the mass, he detected the mixing of silver with the gold, and made the theft of the contractor perfectly clear.

The veracity of the story is very doubtful.

B.2 Newton's unfolding of the universe

Nowadays, the embellished story has the apple falling directly upon Newton's head. Of course, this never happened; and it is doubtful whether the original story, presented below, occurred either, or was merely a pleasant manner in which Newton romanticized his discoveries. However, it can hardly be doubted that Newton did tell this story as an account of how he came to the theory of gravitation. The account below is from Stukeley (1752), a contemporary scholar to whom Newton, then an old man, recounted the apple story:

> After dinner, the weather being warm, we went into the garden & drank thea under the shade of some appletrees; only he & my self. Amidst other discourse, he told me, he was just in the same situation, as when formerly the notion of gravitation came into his mind. Why sh'd that apple always descend perpendicularly to the ground, thought he to himself; occasion'd by the fall of an apple, as he sat in contemplative mood. Why shd it not go sideways, or upwards? But constantly to the Earths centre? Assuredly, the reason is, that the Earth draws it. There must be a drawing power in matter. & the sum of the drawing power in the matter of the Earth must be in the Earth's centre, not in any side of the Earth. Therefore dos this apple fall perpendicularly or towards the centre. If matter thus draws matter; it must be in proportion of its quantity. Therefore the apple draws the Earth, as well as the Earth draws the apple.
>
> [in the margin] that there is a power like that we here call gravity wh. extends its self thro' the universe
>
> & thus by degrees, he began to apply this property of gravitation to the motion of the earth, & of the heavenly bodys: to consider thir distances, their magnitudes, their periodical revolutions: to find out, that this property, conjointly with a progressive motion impressed on them in the beginning, perfectly solv'd thir circular courses; kept the planets from falling upon one another, or dropping alltogether into one center. & thus he unfolded the universe. This was the birth of those amazing discoverys, whereby he built philosophy on a solid foundation, to the astonishment of all Europe.

B.3 The fluttering atoms of Kekulé

Kekulé gave the following account of his discovery of the structure of the Benzene molecule, which paved the way for the entire field of organic chemistry (Rothenberg, 1995):

> During my stay in London I lived for some time in Clapham Road near the Commons. I often spent evenings with my friend Hugo Muller in Islington at the opposite end of the huge city. There we talked about many things, but mostly about our beloved chemistry. Once, on a lovely summer day, I rode the last bus through the empty streets of the otherwise so busy metropolis, "outside"

> as always, on the top deck of the bus. I sank in reverie. There atoms fluttered before my eyes. I had always seen those tiny particles in motion, but I had never succeeded in fathoming the manner of their motion. Today I saw how frequently two smaller ones merged into a pair; how larger ones engulfed two small ones, still larger ones bonded three and even four of the small ones, and how everything turned in a whirling dance. I saw how the larger ones formed a string and dragged along still smaller ones only at the ends of the chain. I saw what Old Master Kopp, my revered teacher and friend, in his “The World of the Molecule” described to us in such a charming way; but I saw it long before he did. The cry of the conductor, “Clapham Road,” roused me from my reveries, but I spent a part of the night putting at least sketches of those musings down on paper. This is how the structure theory came into being.
>
> The same thing happened with the benzene theory. During my sojourn in Ghent in Belgium I occupied an elegant bachelor apartment on the main street. My study, however, was located in a narrow side lane and during the day had no light. For a chemist, who spends daylight hours in the laboratory, this was no disadvantage. There I sat, writing on my textbook; but it wasn’t going right; my mind was on other things. I turned the chair to face the fireplace and slipped into a languorous state. Again atoms fluttered before my eyes. Smaller groups stayed mostly in the background this time. My mind’s eye, sharpened by repeated visions of this sort, now distinguished larger figures in manifold shapes. Long rows, frequently linked more densely; everything in motion, winding and turning like snakes. And lo, what was that? One of the snakes grabbed its own tail and the image whirled mockingly before my eyes. I came to my senses as though struck by lightning; this time, too, I spent the rest of the night working out the results of my hypothesis.
>
> Let us learn to muse, gentlemen, then perhaps we will discover the truth:
>
> “A man not lost in thought
>
> Is given what he’s sought,
>
> He’ll have it with no effort.”
>
> but let us guard against publishing our musings before they have been tested by a vigilant mind.

Interestingly, Kekulé’s account of his own discovery was challenged by Wotiz and Rudofsky (1984), a chemistry professor, who incited a substantial controversy in part due to frequent appearances in the popular press. Wotiz was motivated not only by historical accuracy (claiming that Kekulé might have been motivated by an attempt to assert a doubtful paternity over the discovery), but perhaps also by certain deep-set beliefs about the nature of scientific discovery (Seltzer, 1985):

> Chemists don't operate by dreaming up things. We do experimental work and get hard facts first.

Indeed, Wotiz, together with Rudofsky and Wotiz (1988), explicitly voiced opposition to the use of the anecdote in the psychological literature, though they conclude that, ultimately, psychologists viewed the story as anecdotal rather than as providing substantial support for any theory.

So, did Kekulé have that "dream"? Wotiz was not taken seriously by professional historians of science (Seltzer, 1985), so it is permitted to believe Kekulé's own retelling.

B.4 Poincaré, mathematician and psychologist

Henri Poincaré remains famous today for his many contributions to mathematics, including much of the mathematical basis for the theory of relativity. But for the project undertaken in this thesis, he stands out primarily for his essay *Mathematical invention* (Poincaré, 1909), in which he relates his thoughts on the psychology of mathematical discovery as well as his own experience as a mathematician prone to insight.

I translate the passage below directly from the French:

> A mathematical demonstration is not a mere juxtaposition of syllogisms, it is syllogisms *placed in a certain order*, and the order of these elements is much more important than the elements themselves. If I have the feeling, the intuition so to speak, of this order, such that I perceive at a glance the reasoning as a whole, I will not fear forgetting any of these elements, each one will take by itself its rightful place in the frame prepared for it, without any effort of memory on my part.
>
> (...)
>
> One understands that this feeling, this intuition of mathematical order, which lets us guess harmonies and hidden connections, cannot be possessed by everyone. Some will have neither this delicate feeling, so difficult to define, nor powers of memory and attention above the ordinary, and then they will be absolutely incapable of understanding somewhat higher mathematics; such are the majority.

> Others will have that feeling only to a slight degree, but they will be gifted with an uncommon memory and a great capacity of attention. They will learn by heart the details one after the other, they will understand mathematics and sometimes apply them, but they cannot create. Others finally will have to some extent the special intuition which I just discussed and then they will be capable not only of understanding mathematics, even though their memory is nothing out of the ordinary, but they may become creators and seek to invent with more or less success, depending on whether this intuition is more or less developed in them.
>
> In fact, what is mathematical invention? It is not to make new combinations with previously known mathematical beings. This, anyone can do, but the possible combinations are in infinite number, and the greatest amount is entirely devoid of interest. To invent consists precisely in not constructing useless combinations, and in constructing those that are useful and that are only a minuscule minority.
>
> (...)
>
> The mathematical facts deserving study are those that, by their analogy with other facts, are susceptible to conduct to the knowledge of a mathematical law in the same manner that experimental facts lead to knowledge of a physical law. They are those that reveal unsuspected kinships between other facts, known for a long time, but which we wrongly believed were strangers to one another.
>
> (...)
>
> Sterile combinations will not even present themselves to the mind of the inventor. In the scope of consciousness are only shown genuinely useful combinations, and some others which he will reject, but which resemble the useful ones.

There follow several accounts of insights, each related in precise detail. For instance, at one point Poincaré left his work for a geological trip organized by the École des Mines, and forgot all about mathematics. Then during one leg of the trip,

> (...) at the moment when I stepped on the omnibus, the idea came to me, without anything in my previous thoughts seeming to have paved the way for it, that the transformations which I had used to define the Fuchsian functions were identical to those of non-Euclidian geometry. I didn't check; I wouldn't have had the time since, as soon as I was seated, we resumed the group conversation; but I immediately had total certainty.

Poincaré then theorizes as to what may be happening. He finds a pattern, according to which invention requires a period of conscious reflection on the problem, followed by a period of unconscious thought, followed by an intuition, followed by additional conscious work which verifies and spells out, so to say, the consequences of the intuition. What happens during the subconscious part of this process? For Poincaré, it is guided by experience, which selects which combinations are worth trying; and by a "special aesthetic sensibility" which favors the most promising ones.

B.5 Van Gogh's real intention and purpose

This appendix is for the most part a list of scientific insights; indeed, because this thesis focuses on the more practical personal creativity, it is easier to focus on scientists or engineers than on artists. But in chapter 1, we saw that Csikszentmihalyi (2009) chose the painter Vincent van Gogh as the paradigmatic example of eminent creativity. Van Gogh (1884), as it happens, has described his own creative process as follows in a letter to a friend:

> If I hit it off with a model so that it's calm and quiet and I'm already familiar with it, if I draw that model repeatedly, then among the studies one will eventually come through that's something different from an ordinary study, more true to type, that's to say: more felt.
>
> Yet, that one's made under the same conditions as more wooden, less felt studies that preceded it. This is a way of working like any other – just as understandable, to my mind.
>
> For instance, these little winter gardens – you say it yourself, they're felt – very well, but that's not a fluke, I drew them repeatedly before these and the feeling wasn't in them. After that – after those iron-like ones – came these; so too the clumsy and awkward ones. How it comes about that I express something with them is: because the thing has already formed itself in my mind when I begin. The first ones are utterly unpalatable to other people. I say this so that you should know that if there's something in it, this isn't a fluke but in fact properly reasoned and sought.

B.6 Einstein does not play dice

Gestalt psychologist Max Wertheimer had the opportunity to personally discuss relativity and creativity with Albert Einstein: "those were wonderful days, beginning in 1916, when for hours and hours I was fortunate enough to sit with Einstein, alone in his study, and hear from him the story of the dramatic developments which culminated in the theory of relativity"

(Wertheimer and Wertheimer, 1959, chapter 10). In his book "Productive Thinking", he devotes a chapter to Einstein, and lays out the following "briefest" formulation of Einstein's thought process. In reading this account, one should keep in mind Wertheimer's privileged access to Einstein, but also that Wertheimer was a leading proponent of the Gestalt theory and that his interpretation might have colored Einstein's own rendition of the events:

> In a passionate desire for clearness, Einstein squarely faced the relation between the velocity of light and the movement of a system, and confronted the theoretical structure of classical physics and the Michelson result[1].
>
> A part-region in this field became crucial and was subjected to a radical examination.
>
> Under this scrutiny a great gap was discovered (in the classical treatment of time).
>
> The necessary steps for dealing with this difficulty were realized.
>
> As a result, the meaning of all the items involved underwent change.
>
> When a last arbitrariness in the situation had been eliminated, a new structure of physics crystallized.
>
> Plans were made to subject the new system to experimental test.

Thus Einstein's discovery, as recounted by Wertheimer, was caused first by taking seriously a surprising result, and, dealing "squarely" with the difficulty, challenging assumption after assumption, until finally the entire "gestalt" of modern physics had been restructured.

It is noteworthy that the account seems to follow an almost analytical structure. Indeed, Wertheimer adds:

> One could imagine that some of the necessary changes occurred to Einstein by chance, in a procedure of trial-and-error. Scrutiny of Einstein's thought always showed that when a step was taken it happened because it was required. Quite generally, if one knows how Einstein thought, one knows that any blind and fortuitous procedure was foreign to his mind.

In a 1926 letter to Max Born, Einstein famously wrote: "I, in any case, am convinced that He does not play dice". Seemingly Einstein did not want to rely on coincidences for his own work either. Wertheimer also quotes Einstein as using the expression "the miracle of

[1] The results of the Michelson-Morley experiment, conducted in 1887, puzzled the scientific community with regards to the speed of light.

thinking"; "I rarely think in words at all. A thought comes, and I may try to express it in words afterwards"; and (about the accounts of the theory of relativity) "no really productive man thinks in such a paper fashion [from axioms] (...) in this process [the essentials of the theory of relativity] did not grow out of any manipulation of axioms".

Appendix C

Some contemporary theories of insight

This appendix covers some of the theories aiming to explain insight. These theories adopt different perspectives. The approach of Ohlsson (2011) is essentially that of a cognitive psychologist's; the approach Shen et al. (2017) is that of a neuroscientist; Hélie and Sun (2010) propose a computer model, and Friston et al. (2017) a theoretical/mathematical view.

C.1 Ohlsson: option switches and re-distribution of activation

From a cognitive psychology perspective, the most elaborate description of the restructuring process is given by Ohlsson (2011) (p. 108-109), illustrated in figure C.1. Below, I summarize Ohlsson's account:

- **Architecture:** Problem-solving is based on perception of a current state (e.g. the "state" units in figure C.1). Output from this state perception is propagated in a series of selective layers of processing units – where a processing unit could itself be a neural network. Each of these processing units receives weighted inputs from units in previous layers, and forwards weighted outputs to units in further layers. Each unit connects to units on the next layer, potentially activating them, while also providing relevant problem-related information. All initial weights are learned prior to problem-solving based on past successes and other factors, providing the problem-solver with preferences based on experience.

- **Within-layer dynamics:** In the course of problem-solving, units are activated (when the sum of inputs is above a certain threshold) and deactivated (when the sum of inputs is below threshold). When an activated unit proves unsuccessful (based on interaction with the environment, planning, or heuristic estimation), it receives negative feedback reducing its level of activation. On top of exciting forward connections and inhibiting feedback, these layers have internal dynamics implementing Winner-Take-All (WTA) behavior (Feldman and Ballard, 1982). Thus when a unit is de-activated, within-layer dynamics (in layers WTA#1, ... WTA#n in figure C.1) cause an alternative unit to activate instead.

- **Between-layers dynamics:** Negative feedback is also propagated to previous layers. When the feedback propagated to previous layers leads to deactivating one unit, all of the dependent subsequent activation must be redistributed accordingly – this is restructuring understood as redistribution of activation. Depending on the amount of affected subsequent activation, such restructuring can have a large or small impact. When the entire network is affected, this can be interpreted as the cognitive basis of a strong "Aha!"-experience.

Because processing units can also manage representations, this theory accounts for representational change – including sudden, large-scale changes. Further, as the weights are learned from experience, each unit comes with bias corresponding to a search heuristic. Finally, the theory also describes instances of systematic search (since units of a given layer are activated sequentially unless a change occurs in a previous layer or a solution is found), and, via state changes, it accounts for cases of analytic thinking (Weisberg, 2015) in which restructuring is triggered by the discovery of new information during failed attempts.

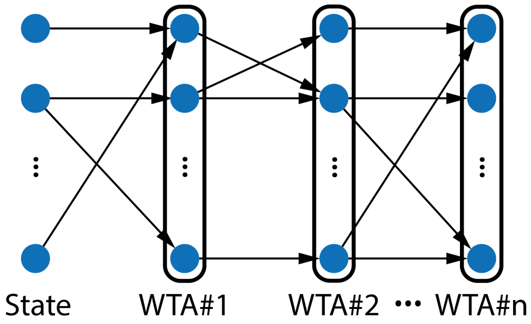


Fig. C.1 A possible visualization of Ohlsson's theory (Ohlsson, 2011). A series of Winner-Take-All (WTA) layers of processing units serves to make decisions based on a perceived state and forward connections. Connection weights are determined by past experience. Unsuccessful options cause negative feedback to backpropagate through the layers, reducing activation. This can push a WTA layer to switch activation to a different option, triggering a chain reaction of redistribution of activation in the units of subsequent layers.

C.2 Shen et al.: brain anatomy of insight

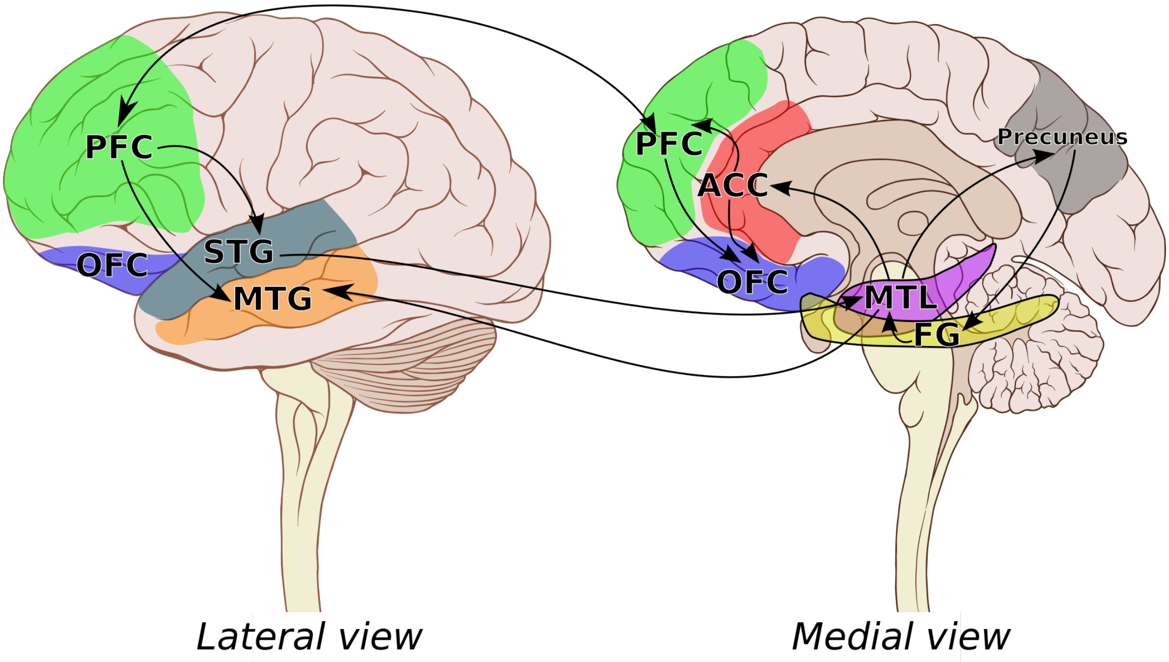


Fig. C.2 Summary of the neural model of insight by Shen et al. Key: OFC - Orbitofrontal cortex and amygdala; PFC - prefrontal cortex; ACC - Anterior Cingulate Cortex; FG - Fusiform Gyrus; MTL - Medial Temporal Lobe (Hippocampus and Parahippocampal Gyrus), STG - Superior Temporal Gyrus, MTG - Medium Temporal Gyrus. The MTL and FG are shown "floating" in front of the medial view; they are respectively on the inside and below the temporal lobe (behind the STG and MTG). Hemispheric differences and inter-hemispheric connections are not shown. Brain region borders are only rough indications.
In this model, the MTL has a central role of novelty recognition and routing of further processing accordingly, for instance to the precuneus for redistribution of attention, or to the MTG for further semantic search. The STG and Fusiform Gyrus are both involved in (re)structuring/integrating information, and in turn connect towards the MTL. Whereas the temporal lobe is primarily involved in search for the solution, the frontal lobe has a monitoring role, with the ACC processing conflicts, and PFC providing executive control over the ongoing search process. If there is an insight, it is processed by the PFC and the Amygdala, generating the emotional "Aha!"-moment that concludes the search.
Background brain drawings by Patrick J. Lynch, medical illustrator; C. Carl Jaffe, MD, cardiologist).

This model by Shen et al. (2017) focuses on spelling out the role of the temporal lobes. Although the model is in many ways speculative - almost every part of it is subject to doubt - it is to our knowledge the only model to have achieved this level of specificity in terms of neural anatomy, and as such it can help to make more concrete the various hypotheses that make up the current state-of-the-art. Figure C.2 is a whole-brain graphical interpretation of that model.

Of particular relevance to this thesis is the critical role of the ACC in processing conflicts, and of the PFC more generally in monitoring the search process. The amygdala, which is considered in this model to be the cause of the emotional aspect of insight, is part of the dopamine pathways. However (perhaps because they explicitly focus their attention on the temporal lobe), Shen et al. (2017) do not discuss a potential role for the basal ganglia or for Reinforcement Learning processes (other than what may be implied by the roles of the PFC and amygdala); for instance there is no mention of the striatum. This is primarily a model of insight interpreted as search and retrieval of potential solutions from memory, in the temporal lobe.

C.3 Helie and Sun: implicit processing

Extensive work by Hélie and Sun led to the influential 2010 article "Incubation, insight, and creative problem solving: a unified theory and a connectionist model" (Hélie and Sun, 2010). The goal of this work was to integrate various views and models of creative problem-solving, insightful or otherwise; in particular, stage-based models (e.g. "preparation, incubation, illumination, and verification", cf. section 1.2 in chapter 1) and process theories such as theories of insight.

The fundamental idea is the distinction between an implicit, associative and subconscious system, and a rule-based, explicit, and conscious system. For Hélie and Sun, insight occurs when implicit processing (incubation) is successful, rises to consciousness, and leads to a sudden and radical change in the explicit system. In this theory, candidate solutions are implicitly generated and explicitly tested.

C.4 Insight as compression or simplification

Several theorists of insights from artificial intelligence backgrounds have re-invented, in sophisticated and contemporary mathematical formulations, the Gestalt idea that insight consists in discovering order and unity in a previously "messy", inefficient or unsuitable representation.

Schmidhuber has not focused on insight but on "creativity, fun, and intrinsic motivation". Work on this theory, ranging from 1991 to 2019 (Schmidhuber, 1991, 2010), focuses on the notion of compression: the reduction of diverse experience to a simpler representation (or world model) should be pleasurable and "fun" to the agent. Friston et al. (2017) express similar views using the framework of Bayesian inference, and argue that insight is an instance of model reduction.[1]

Another notable and influential work following this direction of research is the investigation of small-world networks by Schilling (2005); once again, insight is theorized as the restructuring of experience in a compressed form.

A major weakness of these approaches is the lack of experimental results: how comes insight is so radically sudden? Nevertheless, they appear to capture something essential about insight; the conceptual similarity with Gestalt psychology[2] is suggestive that both have zoned in on an important property of the phenomenon.

C.5 Conclusion

The theories presented above are only a sample of the literature. For instance, *the progress monitoring theory* (according to which agents monitor their progress in solving a problem, and switch strategies when no progress is made) was not discussed, nor was the dynamical systems approach of Stephen et al. (2009). The theories presented here were selected because they each brought an important issue to the foreground: impasse and restructuring for Ohlsson, a neurological model for Shen, subconscious processing for Helie and Sun, and information-theoretic measures for Schmidhuber, Friston, and Schilling. The intention architecture presented in this thesis was inspired by all of these models.

[1] In a broader sense, Friston's work is part of a framework that aims to compete with Reinforcement Learning approaches (Friston et al., 2009); his thesis being that the minimization of expected surprise is, by itself, a sufficient principle to explain cognition.

[2] Neither Schmidhuber nor Friston appear to be aware of the connection.

Appendix D

Reinforcement Learning and insight in the brain

Albeit this thesis is primarily concerned with experimental psychology of insight, it would be an oversight to omit a discussion of the evidence obtained using neuroscientific methods, including brain imaging and stimulation. In order to interpret the neuroscientific evidence on insight, it is necessary to also discuss the Reinforcement Learning paradigm in a psychological and biological context. This appendix is a brief review of the neuroscience of executive control, from a Reinforcement Learning perspective, followed by an equally short summary of the neuroscientific findings about insight.

Reinforcement learning has its roots in behavioural work in animal psychology (Sutton and Barto, 1998, p16), as well as in artificial intelligence and operations research. Interestingly, developments within RL as a paradigm in AI have since found their way back into the psychological and neuroscientific literature, with the discovery that the phasic activity of dopaminergic neurons resembles the implementation of the temporal difference learning signal (Montague et al., 1996; Schultz et al., 1997). This has led to considerable work in the following decades, such that various RL concepts serve as inspiration for neuroscientific theories of habit-formation and executive control (O'Doherty et al., 2015).

Indeed, the influence of RL in psychology and neuroscience has led to certain psychological and neuroscientific approaches being described as "Reinforcement Learning theories" (Ullsperger et al., 2014). However, the psychological and neuroscientific paradigms are not identical with the machine learning one: despite cross-fertilization efforts, disciplinary boundaries cause divergences. Psychological and neuroscientific theories of learning and decision-making often do not brand themselves as RL theories despite displaying many features of RL, e.g. in terms of estimating action-values. See Ullsperger et al. (2014) or

Kable and Glimcher (2009) for reviews of the neuroscience of adaptive behavior and the place of psychological RL in it, along with other influences.

These findings have practical uses, for instance in understanding and combating drug addictions: cocaine, amphetamine and methamphetamine, nicotine, and alcohol overstimulate dopamine mechanisms (Schultz, 2016). They are also relevant to the study of various mental illnesses representing crucially important public health issues, such as schizophrenia (Brisch et al., 2014), depression (Tye et al., 2013), or Parkinson's disease (Damier et al., 1999). Finally, they can help understand economic behavior (Montague and Berns, 2002). I leave those applications aside to focus on the theoretical core of biological RL: decision making in a healthy brain.

In the first two sections of this appendix, I provide a brief, tutorial introduction to the functional anatomy of executive control and reinforcement learning in the brain.

D.1 Psychology

RL and associative learning theories explain animal learning based on the gradual acquisition of rewarding behaviors and the gradual extinction of unrewarding ones; temporal-difference based theories within RL extend this explanation to delayed rewards necessitating sequences of actions which do not have any *immediate* rewarding effect.

RL is related to the earliest studies in animal behavior: namely, conditioning dogs to salivate on cue (classical conditioning; Pavlov (1927/2010)), and making animals learn by rewarded trial-and-error (operant conditioning, Thorndike (1898a)). A major milestone (Miller et al., 1995) for this thread of experimental psychological research was the Rescorla-Wagner model (Rescorla et al., 1972), which predicts the changes in the strength of association V (e.g. between a stimulus and a response action) given a compound stimulus AX according to the following equation:

$$\delta V_A = \alpha_A \beta (\lambda - V_{AX})$$

where λ indicates the asymptote of learning (or, as we would say in machine learning, the observed value), V_A indicates the associative strength based on A and V_{AX} indicates the associative strength based on the compound AX. The variable α denotes the salience of the stimulus, and β is a learning rate. From a machine learning perspective, one would interpret this model as a type of linear perceptron that learns by minimizing the observed error $\lambda - V_{AX}$ between observation and prediction.

The Rescorla-Wagner model is one of many models of animal behavior, valuable for its simplicity. One may interpret TD learning as an extension of Rescorla-Wagner to reward-based behavior extending over multiple time-steps.

Time-Difference RL models of conditioning are nowadays popular, but they are not a panacea: to take one example, they do not readily explain the fast renewal of an "unlearned" association. However, more support for them is found in neuroscientific studies.

D.2 Neuroscience

D.2.1 Dopamine

Dopamine (see figure D.1) is thought to have a crucial role in cognitive control by implementing a time-difference algorithm (Montague et al. (1996); Schultz et al. (1997); see Schultz (2016) for a recent tutorial review)[1]. Indeed, the dopamine phasic response has been found to encode temporal difference errors:

- Surprising rewards elicit activation of dopamine neurons;
- Surprising absence of rewards cause dopamine neurons to be inhibited;
- Predicted rewards have no effect;
- A surprising stimulus, predicting a future reward, elicits activation;
- Delayed rewards are temporally discounted.

Dopamine neurons are found principally in the midbrain (VTA and SN). However, learning from reward requires a broad range of abilities: reward evaluation, associative learning, choice between incompatible strategies, and planning (Haber and Knutson, 2010). Thus many brain areas beyond dopaminergic neurons themselves are recruited in the networks that lead to dopaminergic activity.

The key structures indirectly involved in dopaminergic activity are (see figure D.1 for their localization in the brain and figure D.2 for an overview of their connectivity; see Haber and Knutson (2010) for a more in-depth discussion):

- The prefrontal cortex;
- The striatum;
- The thalamus;
- Midbrain nuclei.

The next section discusses the relationships between these structures.

[1] Dopamine transients have also been found to correlate with e.g. non-rewarding prediction errorsSharpe et al. (2017). This appendix focuses specifically on time-difference errors, but there is evidence that this does not exhaust the functions of dopamine in the brain.

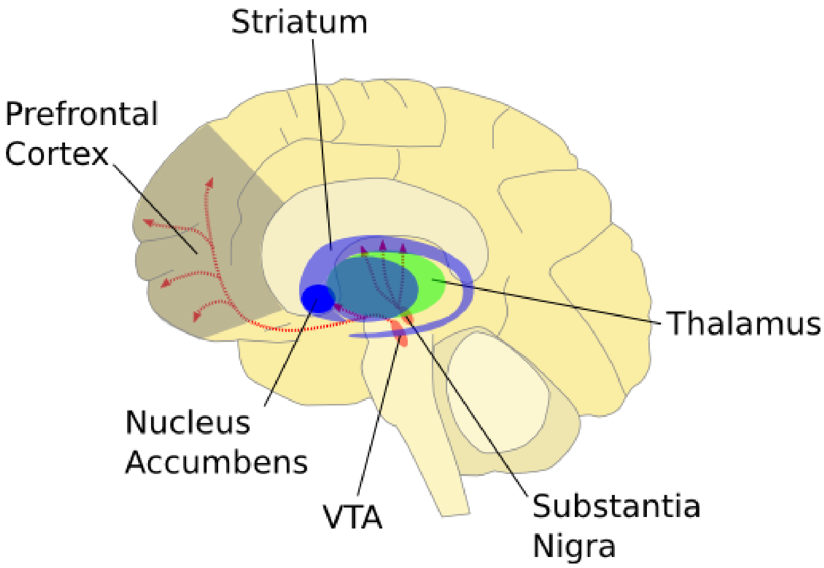


Fig. D.1 Dopamine in the brain. The red dotted arrows indicate dopaminergic connections. Note that many of the structures involved (VTA, Substantia Nigra, Thalamus, Striatum) are centrally located in the head. This makes them difficult to study using non-invasive methods, due to interferences caused by other brain structures, heart beat, anatomical differences, and so on.

D.2.2 Subcortical structures

Figure D.2 provides an overview of the brain structures involved in reinforcement learning and decision making.

The basal ganglia are a collection of nuclei forming a large network underneath the cortex (cf. figure D.1), including the striatum[2], dopaminergic areas (the substantia nigra and ventral tegmental area), and the subthalamic nucleus (STN). The basal ganglia receive inputs from virtually the entire cortex, as well as from certain mid-brain areas (such as the Raphe and Pedunculopontine nuclei). These inputs have been theorized to encode, or serve in the computation of time difference errors. For instance, the Raphe nucleus has been associated with primary rewards (Liu and Ikemoto, 2007); the lateral habenula, which serves as a relay in this network (not shown in figure D.2) has been associated in negative reward (Matsumoto and Hikosaka, 2007). The outputs of the basal ganglia, besides dopaminergic outputs in the PFC, are principally located in the thalamus, and from there in the frontal lobe (but also in temporal and parietal areas), as well as to several subcortical structures involved in modulating behavior (Haber and Knutson, 2010).

[2] The striatum is further divided into ventral and dorsal striatum; caudate nucleus and putamen; including the nucleus accumbens, the olfactory tubercle, the globus pallidus (internal and external parts); and sometimes the amygdala.

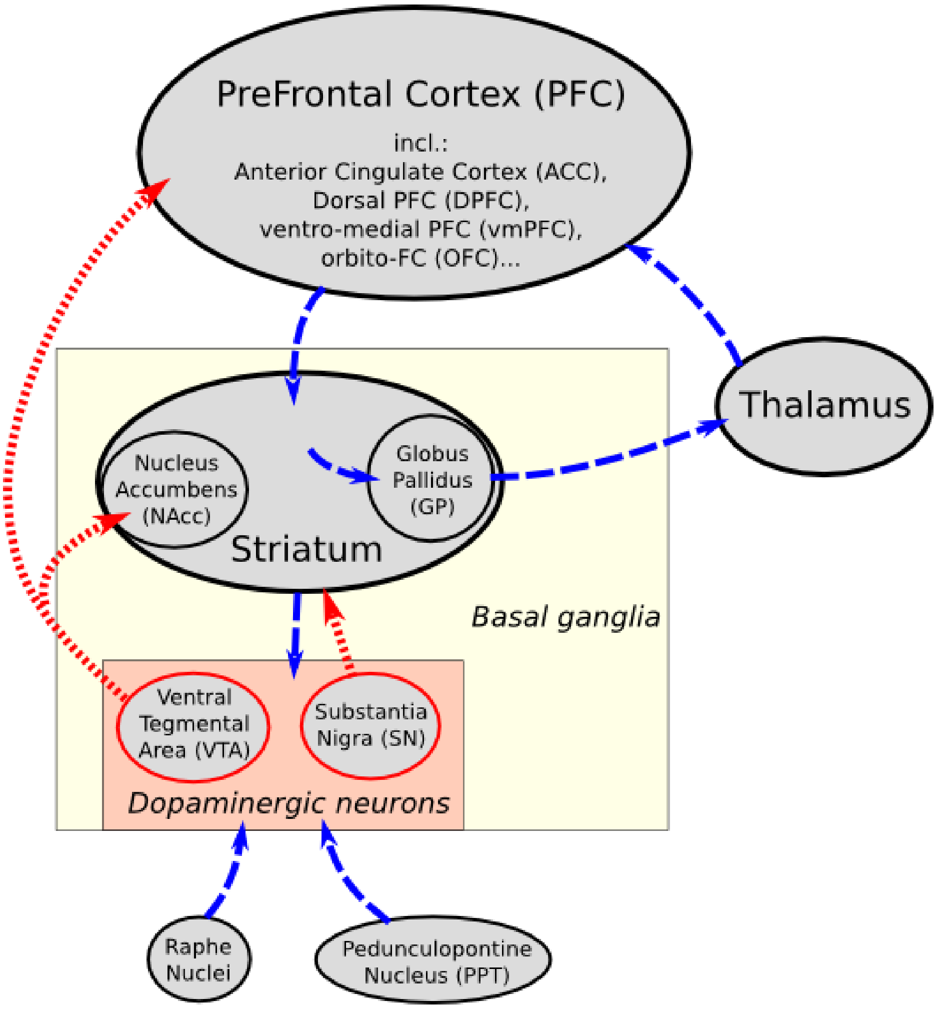


Fig. D.2 Network of decision-making and reinforcement learning. Dopaminergic neurons in the midbrain (red box) project onto the striatum and prefrontal cortex (short red dashes), dopamine encoding the time-difference error signal of reinforcement learning. Non-dopamine connectivity is shown in long blue dashes, and shows the structures likely involved in decision-making, computation of time-difference errors, planning, and so on. Afferent connections (inputs) to dopaminergic neurons originate in the Raphe nuclei, the pedunculopontine nucleus, and Striatum; the circuit also involves the PFC and Thalamus. Some of the core structures of this network (striatum, dopaminergic neurons) are called the basal ganglia (yellow box).

The precise function of each element in this network is beyond the scope of this chapter, and, to a large extent, beyond the understanding of contemporary neuroscience. For instance, from a reinforcement learning perspective it is surprising that a separate structure (the lateral habenula) should specialize in negative time-difference errors: does this organization serve a useful function, or is it an evolutionary vestige?

D.2.3 Prefrontal Cortex

Much of the neuroimaging results with respect to insight find cortical, rather than subcortical activity. From a Reinforcement Learning and Executive Control Network perspective, this suggests taking a closer look at the prefrontal cortex (PFC).

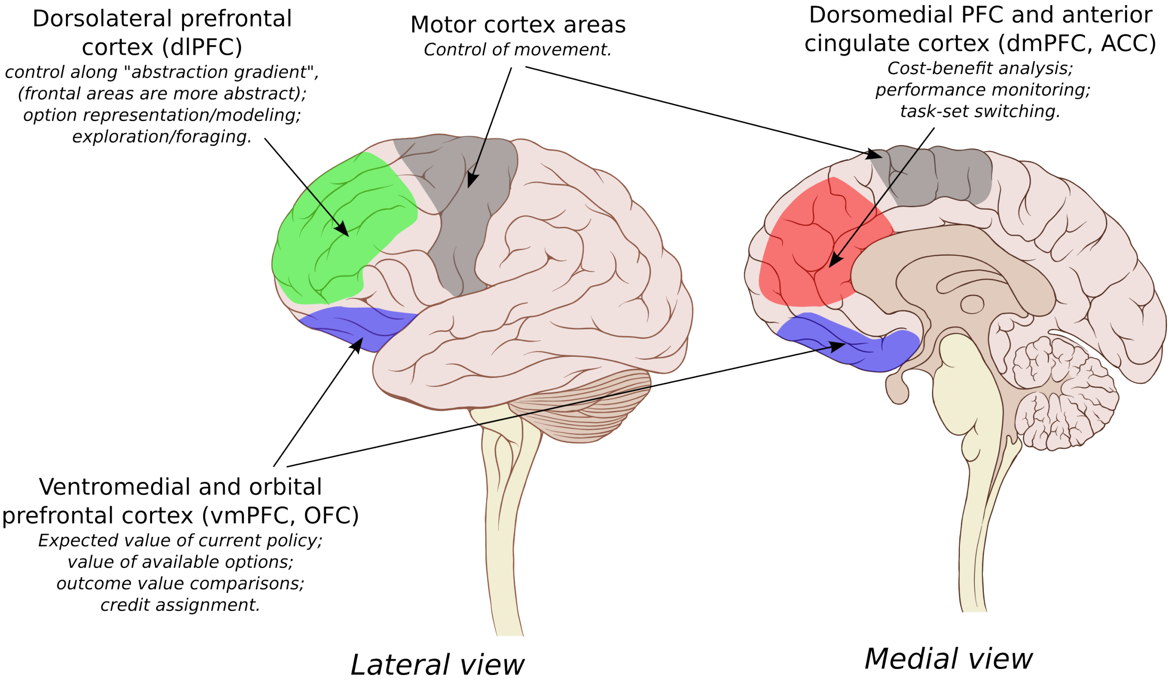

Dorsolateral prefrontal cortex (dlPFC)
control along "abstraction gradient", (frontal areas are more abstract); option representation/modeling; exploration/foraging.
Motor cortex areas
Control of movement.
Dorsomedial PFC and anterior cingulate cortex (dmPFC, ACC)
Cost-benefit analysis; performance monitoring; task-set switching.
Ventromedial and orbital prefrontal cortex (vmPFC, OFC)
Expected value of current policy; value of available options; outcome value comparisons; credit assignment.
Lateral view
Medial view

Fig. D.3 Hypothesized functions of prefrontal cortex areas. *(Background brain drawings by Patrick J. Lynch, medical illustrator; C. Carl Jaffe, MD, cardiologist.)*

The considerable amount of experimental work has yet to converge on consensual cortical architectonics[3]. Nonetheless, some general observations have been confirmed by a combination of fMRI, EEG, lesion-studies, and single-neuron studies in both humans and animals, and are thus exceptionally robust. The discussion below is primarily based on reviews of the neuroscientific evidence (Kable and Glimcher, 2009; Rushworth et al., 2011; Ullsperger et al., 2014). These results are summarized on figure D.3; I qualify them below area by area:

- **Ventromedial and orbital PFC:** considerable evidence links this area to the estimation of value (Rushworth et al., 2011); however, exactly what type of value it represents is not entirely clear, with evidence supporting multiple candidates (not necessarily

[3] "Architectonics" is the art and science of identifying brain region. Identifying precise regions in the prefrontal cortex remains a challenge and is the subject of ongoing debate (Passingham and Wise, 2012), hence the intentionally approximate region borders shown in figure D.3.

incompatible). A popular account associates the lateral Orbital Frontal Cortex (OFC) with the expected outcome value (Schoenbaum et al., 2009) (in RL terms, the expected return), whereas the ventromedial PFC could estimate differences between the value of distinct options (Ullsperger et al., 2014) (in RL terms, the advantage). But see for instance (Wilson et al., 2014) for an alternative account according to which OFC encodes task-relevant states.

- **Dorsomedial PFC, esp. Anterior Cingulate Cortex:** the understanding of the role of the Anterior Cingulate Cortex (ACC) has changed in recent years. Previously seen as monitoring conflict, the ACC is nowadays more commonly understood as integrating information related to performance monitoring (incl. new task-relevant information, motivational information...) to evaluate and trigger task-set switching, i.e. allocating control at an abstract, high level (Shenhav et al., 2013).
- **Dorsolateral PFC:** the lateral PFC is primarily associated with working memory and control (Venkatraman and Huettel, 2012). Furthermore, it seems to be organized along a rostro-caudal abstraction gradient, such that more rostral regions are more abstract (Badre, 2008), with the gradient extending to the less abstract premotor and motor areas, suggesting a hierarchical organization. At the most rostral and abstract level, the frontopolar cortex has been implicated in choices to explore and in foraging (Daw et al., 2006). Along with control, the lateral PFC appears to help generate and maintain goal-related representations in working memory Badre (2008).

The specific role of each structure, as well as the global function of their cooperative behavior, remains elusive: anatomical descriptions and experimental observations remain difficult to reformulate as functional explanations. Are these circuits implementing an actor-critic system (Joel et al., 2002), with separate structures or networks for valuation and decision? Is part of the circuit concerned with model-based and/or model-free RL (Dolan and Dayan, 2013)? What about temporally extended action and hierarchical RL (Botvinick, 2012)? While there has been considerable attention, speculation, and even though considerable evidence was obtained for various theoretical constructs, no clear picture of decision making in the brain as a whole has emerged; and so it is only possible to provide controversial accounts of the role of several structures. Besides the already discussed role of dopamine, the rest of the algorithm has yet to reveal itself unambiguously.

Nonetheless, the rough functional anatomy discussed here is sufficient to make certain predictions about how a executive control might be involved in insight. For instance, one would expect strategy switches to involve the dorsomedial and dorsolateral PFC; and successful insights to activate dopamine neurons.

D.3 Neuroscience of insight

We have seen in chapter 2 that the main characteristics of insight are an incorrect initial representation, a sudden representational change, and a sense of instantaneous progress accompanying this change; we have also seen that it appears to depend on the availability of relevant previous experience (Birch, 1945; Wiley, 1998, ...), and that its probability of occurrence is increased by sleep (Stickgold and Walker, 2004, ...). This short description sets up the enigma I wish to solve: what kind of "mechanical" cognitive process could possibly have these properties? Before trying to implement such a mechanism on a computing machine, it is worth looking at one "machine" that already achieves this: the human brain. In this section I take a closer look at the human brain during insight.

As of 2019, there are about 35 published neuroscientific studies of insight, including around 20 imaging studies using Electro-Encephalogram (EEG) or functional Magnetic Resonance Imaging (fMRI) and about 10 transcranial stimulation studies using Transcranial Magnetic Stimulation (TMS) or direct Current Transcranial Stimulation (dCTS) (Sprugnoli et al., 2017). These studies used a great variety of different tasks, including variations on the remote associates test Bowden and Jung-Beeman (2003b); Mednick (1962), anagrams, and various types of riddles, with most problems being linguistic in nature (Sprugnoli et al., 2017); see appendix A for a non-exhaustive review of insight tasks.

This review of the neuroscientific literature relies in part on several existing reviews which readers are invited to consult for a more detailed account; these are by Dietrich and Kanso (2010), Kounios and Beeman (2014), Sprugnoli et al. (2017), and Shen et al. (2017).

D.3.1 Correlates of insight in the brain

Figure D.4 shows the brain regions for which correlations with the "Aha!" moment have consistently been found in the literature, with an overlap for the two main imaging methods: fMRI and EEG[4], based on the recent meta-analysis by Sprugnoli et al. (2017). Of these three regions, those accumulating the most evidence are the Anterior Cingulate Cortex and the Superior and Middle Temporal Gyri (STG and MTG), all of which display activity consistent with initiating an insight.

The fMRI/EEG overlap meta-analysis (figure D.4), is a much simplified picture, focusing only on the strongest and most reproduced effects. There is in fact a great variety of findings. With respect to the "Aha!"-moment itself, other findings includes activity across the prefrontal cortex, including lateral PFC (Luo et al., 2004c) and dorsolateral PFC (Goel and Vartanian,

[4] EEG measures electrical activity with good temporal resolution but poor spatial resolution; whereas fMRI measures activity by means of blood flow, with good spatial resolution but poor temporal resolution

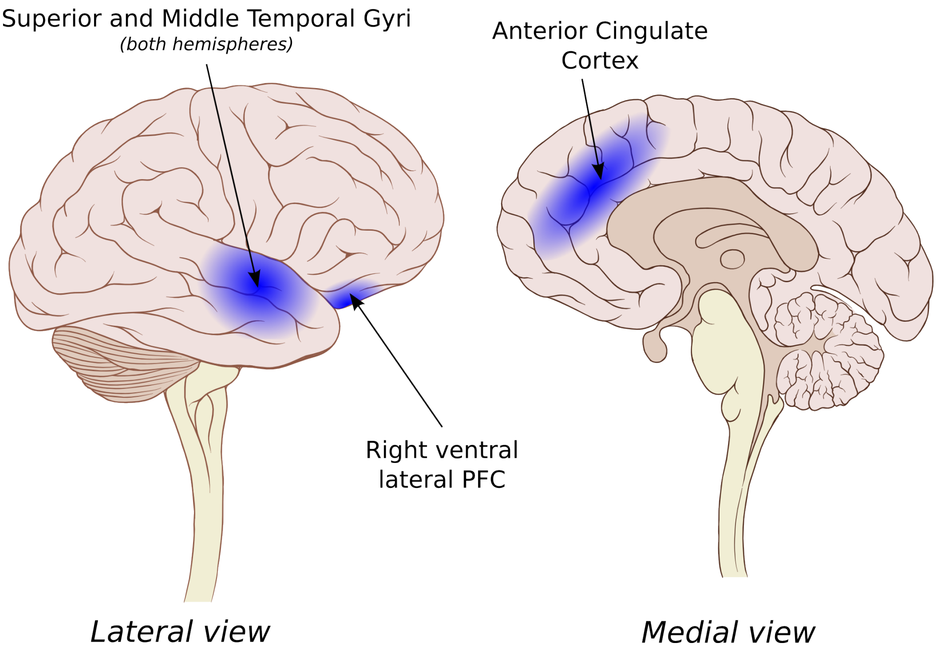

Superior and Middle Temporal Gyri
(both hemispheres)
Anterior Cingulate Cortex
Right ventral lateral PFC
Lateral view
Medial view

Fig. D.4 Overlap between EEG and fMRI sources *(figure based on the meta-analysis by Sprugnoli et al Sprugnoli et al. (2017); background brain drawings by Patrick J. Lynch, medical illustrator; C. Carl Jaffe, MD, cardiologist.).*

2004), insula (Aziz-Zadeh et al., 2009)... as well as the hippocampus (Luo et al., 2004b; Zhao et al., 2013) or even the cerebellum (Goel and Vartanian, 2004).

Besides activity occurring during the “Aha!”-moment, one may also be interested in activity that precedes insight: if insight is the result of subconscious processing, then one might expect to detect characteristic activity in the moments preceding the “Aha!”. The main correlate of mental preparation for insight (both prior to problem presentation and during active problem-solving) consist in changes in oscillation patterns in the PFC and temporal lobes, and a decrease of activity in the occipital lobe (Kounios et al., 2006, 2008).

Finally, one may observe the effects on insight problem-solving of interventions that disrupt the brain’s normal functioning. Existing studies have often used Transcranial Magnetic Stimulation (TMS) to affect neural activity in a localized area of cortex. For instance, Chi and Snyder (2011, 2012) used TMS on the temporal lobes, demonstrating higher insight ability when the right temporal lobe was stimulated. Cerruti and Schlaug (2009) and Metuki et al. (2012) also found performance improvements for interventions on dorsolateral PFC . A study on patients with lesions by Reverberi et al. (2005) found that lesions to lateral PFC *increased* the ability to solve matchstick arithmetic problems; this validated the experimenters’ prediction that lateral PFC serves in “biasing response space”, such that patients with lesions revert to “trial and error”. The same study found similar but weaker effects for lesions to the medial PFC.

D.3.2 Temporal lobe

One of the leading theories in the neuroscientific literature of insight is based on hemispheric asymmetry[5]. Indeed, there are connectivity differences between the two hemispheres: certain right-hemisphere neurons in association cortices have larger input fields than do their left-hemisphere counterparts, meaning more synapses, and synapses further away from the cell body Jacobs et al. (1993). This suggests, according to Jung-Beeman et al. (2004); Kounios and Beeman (2014), that the right hemisphere may be capable of remote linguistic associations, whereas the left hemisphere would be specialized in making close associations (Jung-Beeman et al., 2004; Kounios and Beeman, 2014). However this view is disputed by Dietrich and Kanso (2010); Sprugnoli et al. (2017), in particular because several studies found bilateral activations of the temporal gyri (e.g. Luo et al., 2004b), or even the opposite lateralization (e.g. Luo et al., 2004c; Tian et al., 2011).

Beyond the hemispheric specialization theory, the involvement of temporal regions (MTG, STG, but also sometimes fusiform gyrus (Shen et al., 2017)) in insight is a robust finding. Although some have remarked that the predominance of linguistic tasks might at least partly explain the activation of these regions (Dietrich and Kanso, 2010), most specialists agree that the region is important for insight in general (Kounios and Beeman, 2014; Shen et al., 2017; Sprugnoli et al., 2017).

D.3.3 Frontal lobe

Frontal lobe activations correlating with insight, especially the ACC (Kounios and Beeman, 2014; Sprugnoli et al., 2017), have for the most part been discussed in the literature in terms of attention, conflict detection and monitoring of competing or incompatible strategies, in relationship with mental set[6] and mood.

The role of attention has been discussed extensively in the more general case of creativity (Beaty et al., 2015; Kasof, 1997; Zabelina and Robinson, 2010), often specifically in terms of its modulation by dopamine (Boot et al., 2017; Zabelina et al., 2016). It has been suggested that creative thinking, and specifically insightful thinking, depends on the ability to

[5] This view echoes beliefs (usually considered myths) that people can be “left-brained” (analytical) or “right-brained” (creative).

[6] Mental set is the tendency to approach a problem in a certain way based on experience. The Einstellung effect (German for “set” effect, but referring typically to misleading rather than helpful set effects) is one of the potential sources of impasse (Bilalić et al., 2008) during problem-solving.

achieve “flexible” cognition by avoiding both rigid thinking and distraction (Boot et al., 2017; Zabelina and Robinson, 2010). For Kounios and Beeman, “it appears that the tendency to solve problems insightfully is associated with broad perceptual intake as the default mode of resting-state attention deployment, coupled with the tendency to focus inwardly in preparation for, and during, solving” Kounios and Beeman (2014). Discussions of insight in terms of attention often adopt a metaphorical view in terms of which a “broad attention” allows a diverse set of weakly activated associations to enter a competition for further processing.

A candidate for the arbitration of such a competition is the anterior cingulate cortex (ACC), whose function was understood as detecting conflict between competing hypotheses (Botvinick et al., 2001); but which is now more often considered as allocating control (Shenhav et al., 2013) (as discussed in annex D.2.3). This has suggested a role for the ACC in detecting new strategies and allocating control to them[7], thereby breaking mental set (Luo et al., 2011; Öllinger et al., 2008).

Finally, Subramaniam et al. (2009) considered the effect of mood on insight following up on earlier findings in the psychological literature (Isen et al., 1987; Suzanne K. Vosburg, 1997). The results suggest that mood affected preparatory activity in the ACC, which in turn affected the attention and executive control mechanisms discussed above.

D.3.4 Summary of neuroscientific findings

Dietrich and Kanso (2010) remarks that “research on the basis of insight reflects greater consistency than research on either divergent thinking or artistic creativity”[8]. However, the neuroscientific discussion has often been dominated by the controversy over the hemispheric interpretation proposed by Kounios and Beeman (2014), and rejected or doubted by Dietrich and Kanso (2010) and Sprugnoli et al. (2017). Apart from the role of the hippocampus in memorizing insight solutions[9], much of the congruent data remains in wait of a consensus interpretation. The reinforcement learning theory proposed in this thesis can be, I believe, a new contender in explaining the prefrontal activations observed during insight, although it has less to say about temporal lobe activations.

[7] However, an alternative account sees the role of the ACC not in breaking the mental set to enable insight, but merely in processing an already retrieved solution (Anderson et al., 2009; Shen et al., 2017).

[8] Perhaps owing to ambiguities in conceptualizing creativity, as discussed in chapter 1?

[9] The stronger activation of the hippocampus for insight compared to non-insight solutions, first reported by Luo et al. (2004b), is likely to correspond to the improved memorization of insight solutions. Danek et al. (2013) verified this prediction, observing that, two weeks after problem-solving, recall was facilitated for solutions found with insight compared to solution without insight.

D.4 Conclusion

D.4.1 Summary

- RL is in large part inspired by psychological models of animal learning.
- RL has been an influential framework in neuroscience, providing useful interpretations of executive control networks and of dopamine-driven learning in the brain.
- The Anterior Cingulate Cortex (ACC) is understood to have a key role in choosing strategies, arbitrating between competing options, or in hierarchical reinforcement learning.
- Neuro-imaging studies show that the ACC is also implicated in insight-problem solving, along other structures including areas in the temporal lobes and hippocampus; however, theoretical efforts seeking to explain insight, e.g. by Kounios and Beeman (2015); Shenhav et al. (2013), have focused on these other areas, rather than the ACC.

The evidence from neuroscience, both with regards to Reinforcement Learning and Insight, is therefore at least superficially compatible with the model presented in this thesis; albeit a more thorough, experimental investigation, using imaging techniques, would be required to confirm that the model has neuroscientific validity.

D.4.2 Bibliographical remarks

This appendix tackles a very large field of research; a thorough review from the original experimental literature would be a daunting task, well beyond the scope of this thesis. The psychology section about reinforcement learning is largely based on chapter 14 from Sutton and Barto (2018). The neuroscience sections about reinforcement learning are based on recent review articles, e.g. Badre (2008); Botvinick (2012); Dolan and Dayan (2013); Joel et al. (2002); Schultz (2016); Shenhav et al. (2013); Ullsperger et al. (2014); Wilson et al. (2014); the neuroscience sections about insights are based on Dietrich and Kanso (2010), Kounios and Beeman (2014), Sprugnoli et al. (2017), and Shen et al. (2017).

Appendix E

Experiment details

This appendix contains descriptions of algorithms or environments used in various experiments discussed or mentioned in the main text of the thesis.

This includes descriptions of experiments, or of their results, that are too detailed or insufficiently related to the main findings to feature in the main text; they are included to ensure all experimental details necessary for replication are adequately described.

This also includes exploratory experiments whose scientific interest is limited though non-zero, and which are therefore only described briefly here.

E.1 Details from chapter 4

E.1.1 Preliminary test of RL algorithms

The poor performance of A2C (Wu et al., 2017) (cf. figure 4.5 was a major surprise. To verify the implementation, I compared my implementation of A2C with OpenAI's implementation (Wu et al., 2017), and found similar performance when using the same hyperparameters on the pendulum balancing task. The explanation thus presumably lies in the task itself. Indeed, A3C (Mnih et al., 2016) and its variant A2C were tested on tasks which allow for frequent rewards, whereas this task has sparse rewards. A2C's Monte-Carlo batch updates presumably allow information to travel faster from the rewarding transitions, hence the initial advantage over other algorithms. However, in this task each episode is different and internally correlated (the states observed within an episode are similar to each other, i.e. the spot is always in the same position within an episode). This is perhaps the source of instability which made it necessary to use very small learning rates, as shown in table 4.1.

E.1.2 Hyperparameters for the pigeon models

Table E.1 shows the hyperparameters used for the pigeon models.

	Learning rate (actor)	Learning rate (critic)	Parallel MDPs	Switching bias
Condition 1 and 2	0.006	0.0006	16	n/a
Condition 3	0.01	0.001	16	n/a
Condition 4	0.006	0.0006	16	0.0
Condition 5	0.006	0.0006	16	0.1

Table E.1 Hyperparameters used in the final pigeon models. Recall that conditions 1 and 2 use the flat RL model, condition 3 uses the flat RL model and does not undergo shaping, and conditions 4 and 5 use the HRL model.

Note that a higher learning rate could be used for condition 3, the model which learned from scratch. This is presumably because these agents did not encounter abrupt changes in the task difficulty, by contrast with the agents who underwent shaping. For conditions 4 and 5, learning rate was slowed down to $0,0006$ for the option-actors post-shaping to preserve stability. The top-level Sarsa-critic had a learning rate of $0,0005$ and an exploration rate ε of 0.1.

E.1.3 Shaping

In the main text, only one individual graph of shaping, and a graph of the average shaping trajectory, are included. Figure E.1 below shows several individual graphs of agents undergoing shaping. The first one is the one shown in the main text. These 4 graphs correspond, in order, to the first four runs from the twenty runs that served to produce figure 4.14; they are representative of the agents' learning behavior.

E.2 Individual "insights"

Figure E.2 shows how each of the 20 individual agents in condition 2 (flat agents undergoing additional shaping training) performed on the first 20,000 timesteps of the test. (All agents eventually solved the task, but the great variability between agents made it difficult to find a timescale at which the full curves could be shown for all agents, while preserving a good resolution.)

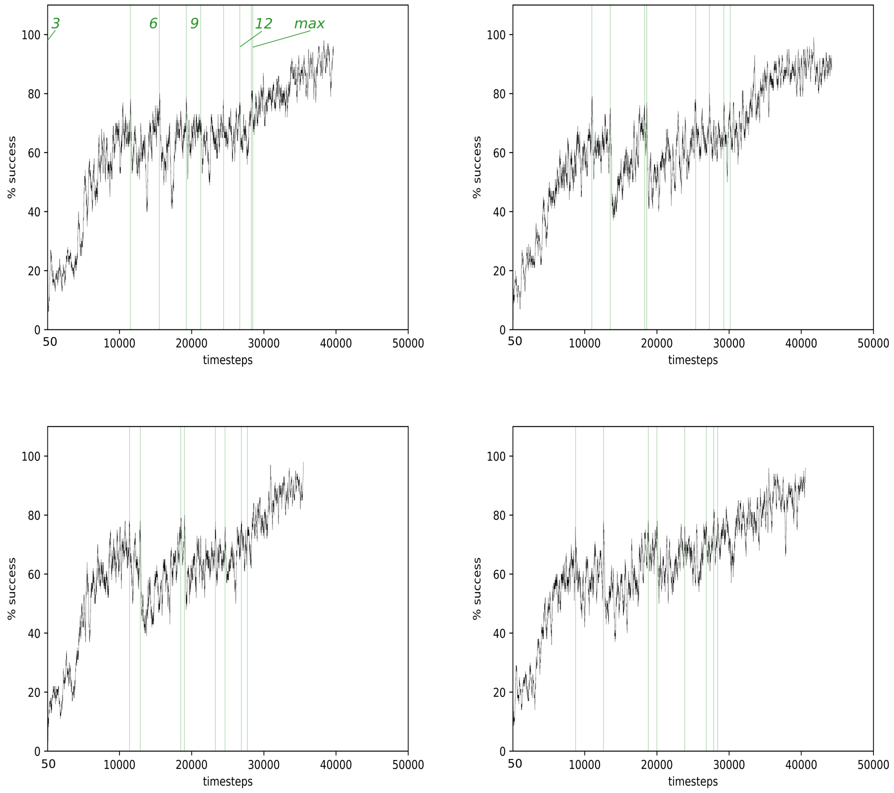


Fig. E.1 The course of learning for individual agents undergoing shaping. The vertical lines indicate transitions to a more difficult version of the task.

E.3 Option-critic and general options

The four-rooms environment is a classic test-bed for hierarchical reinforcement learning. In this experiment, which was preparatory work for the experiment in chapter 4, I tested an option-critic architecture on a variation of the four-rooms environment in which episodes alternated randomly between four possible objectives (see figure E.3) in an explorative experiment. This particular four-rooms environment is continuous. In any episode, the agent perceived its state (encoded using tile-coding) and a 1-hot encoding of which objective was active. The objective of the experiment was to find out whether an option-critic algorithm could find re-usable options - which, in the four-rooms environment, are understood to correspond to the gates (which bottleneck states).

In this implementation of the option-critic architecture, I used 6 separate actor-network-options, one actor-over-options, and a critic. Because all objectives are in different locations of the bottom right room, one may expect the agent to learn general options which lead to the gates of that room, then give back control for the specific final step.

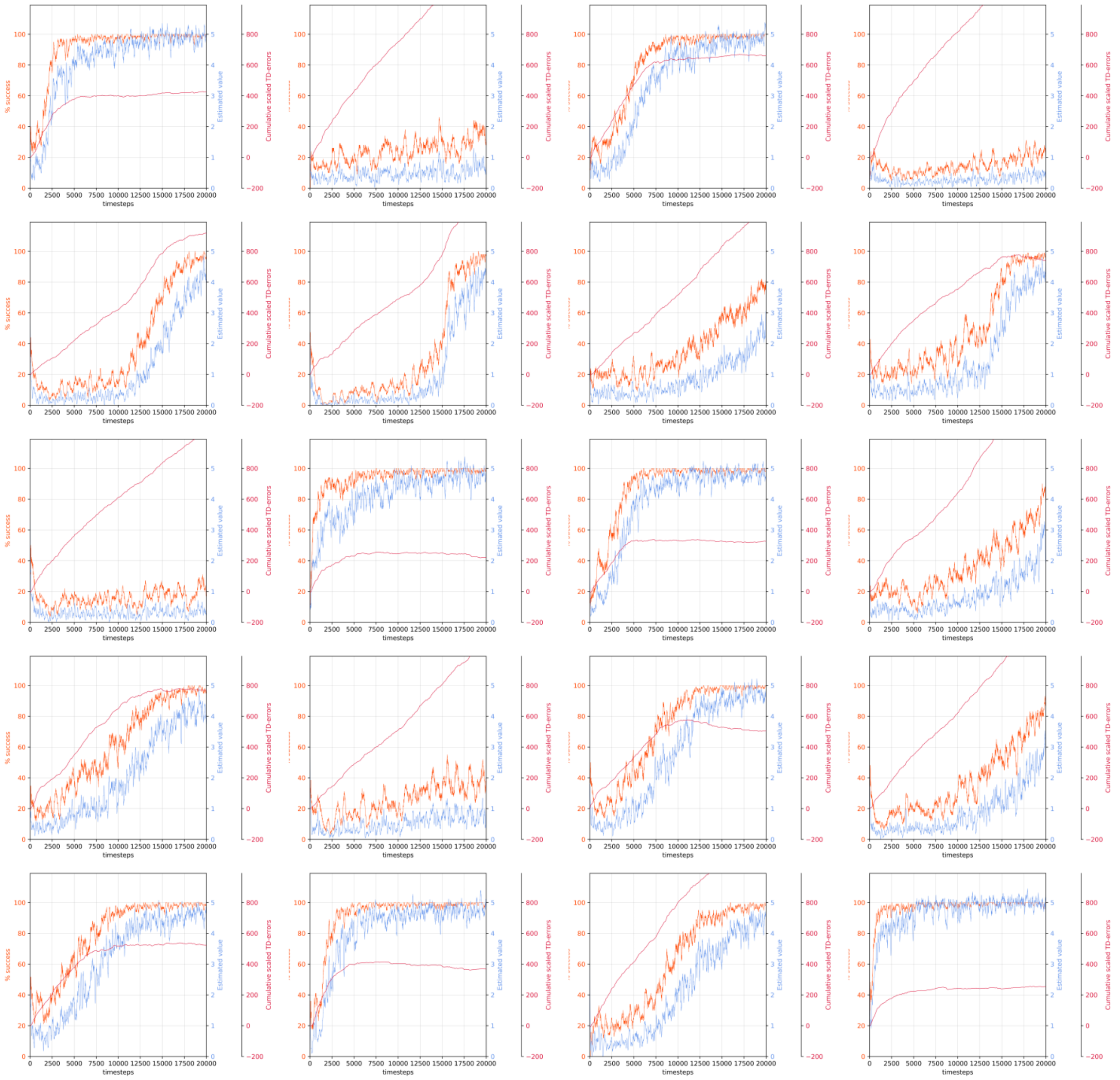


Fig. E.2 Test-phase behaviors for all 20 runs. This shows the high variability of behavior following over-training in the shaping tasks. All instances (some not shown) eventually achieved high performance; the lowest-performing one needing about 70,000 time-steps to achieve 95% accuracy, and the best-performing one about 1,500 time-steps. All were considerably faster than inexperienced agents.

The results were difficult to interpret, but overall suggested that the agent did not learn options representing the structure of the task. Figure E.4 shows data for what was perhaps the most “meaningful” set of options, for one of the four objectives. (Only three of the six options are shown - the other three were not significantly used in this configuration.)

Overall, although some cases seemed to show some specialization of options as in figure E.4, in most cases one option ended up being used all the time, or options were used in an

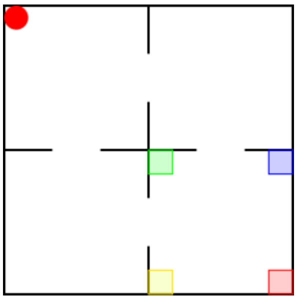


Fig. E.3 A variation on the "four rooms" MDP. There are four possible objectives, in the four corners of the bottom-right room. The agent always starts in the top-left corner.

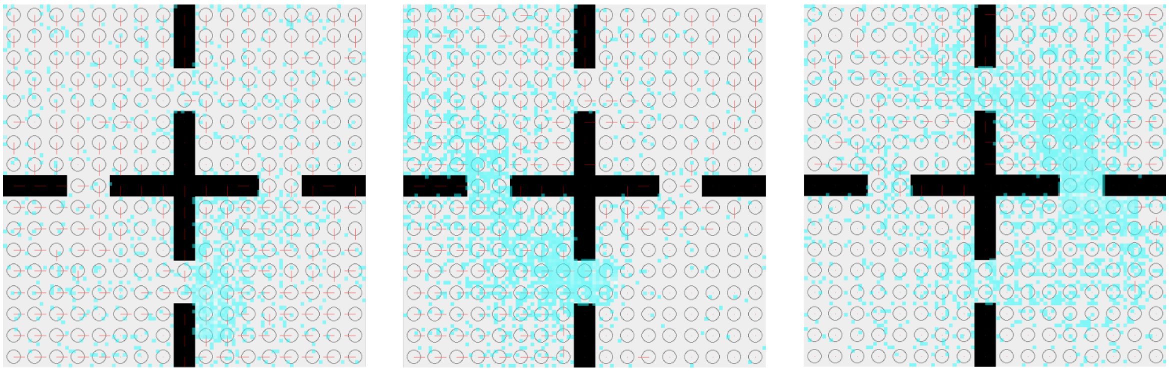


Fig. E.4 Screenshots of an option-critic architecture operating on the four rooms environment. The three images correspond to the three options that are used. The blue overlay represents the state distribution for that option (places where the option was in control of the agent's behavior). The red overlay represent the preferred action at some states: a red line indicates the direction of movement indicated by the policy of the option for that state, whereas a dot indicates termination of the option. In this case, option 1 led to the north gate of the bottom-right room, option 2 led to the south gate of the bottom-right room, and option 3 was used mostly in the final room, leading to the bottom-left corner.

apparently non-meaningful way, starting and terminating in arbitrary states rather than in gateway states such as after going through a "door". These results were a factor in designing the HRL pigeon models in such a way that meaningful "options" could be learned during shaping (one for "jump-and-peck" and one for "push-to-spot").

E.4 Details about experiments from Chapter 5

E.4.1 Deep exploration

Section 5.1.1 of chapter 5 presented a technique for deep exploration and some results obtained using it. This technique was tested on a toy MDP, shown in figure E.5, reminiscent of the East-West metaphor developed in chapter 5.

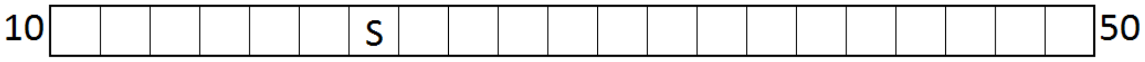


Fig. E.5 A toy MDP for deep exploration. A reward of 10 is easily accessible on the left, but a higher reward is available on the right. Agents find the reward of 10 easily, but this biases them against exploring in the other direction and finding the reward of 50.

Figure E.6 shows some results on this toy MDP. The new algorithm discovers the far reward quickly, thanks to exploring by “bursts”.

Note that the objective was not to solve the MDP efficiently, but to show that exploration bursts can improve performance, all other things being equal. Because of this, I set the initial value function to 0 everywhere. Setting the initial value function *optimistically* (to a high value everywhere, e.g. 100) would have encouraged exploration and resulted in a much smaller difference between the two algorithms. Note however that optimistic initialization is only possible for tabular reinforcement learning.

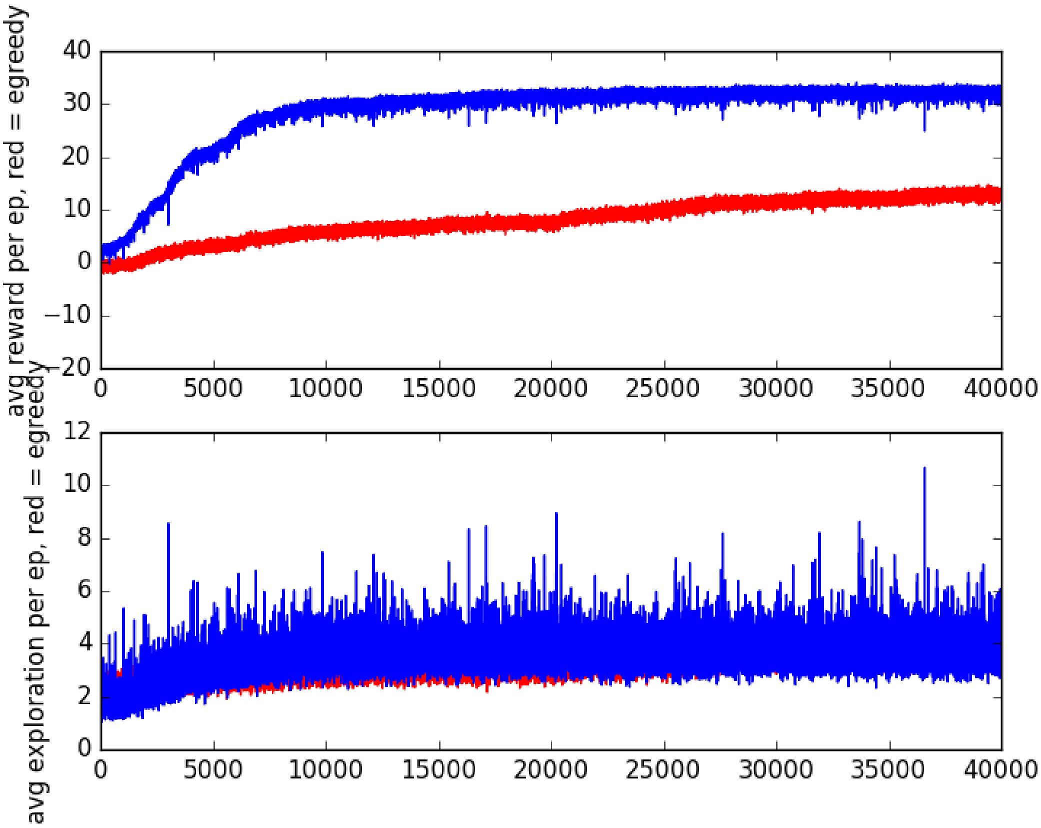


Fig. E.6 Improvements using upsilon-greedy exploration. Top: average reward. Bottom: total exploration per episode. Note that upsilon-greedy explores in prolonged bursts.

References

Artificial intelligence

D. Andre. *Programmable reinforcement learning agents*. PhD thesis, University of California, Berkeley, 2003.

A. Augello, I. Infantino, A. Manfré, G. Pilato, and F. Vella. Analyzing and discussing primary creative traits of a robotic artist. *Biologically Inspired Cognitive Architectures*, 17:22–31, 2016. doi: 10.1016/j.bica.2016.07.006.

P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 1726–1734, 2017.

A. G. Barto. Intrinsic motivation and reinforcement learning. In *Intrinsically Motivated Learning in Natural and Artificial Systems*, pages 17–47. Springer, 2013.

A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(1-2):41–77, 2003. doi: 10.1023/A:1022140919877.

A. G. Barto, G. Konidaris, and C. Vigorito. Behavioral hierarchy: exploration and representation. In *Computational and Robotic Models of the Hierarchical Organization of Behavior*, pages 13–46. Springer, 2013.

R. E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 41–48. ACM, 2009.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828, 2013a. doi: 10.1109/TPAMI.2013.50.

Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013b.

T. R. Besold and K.-U. Kühnberger. Towards integrated neural–symbolic systems for human-level AI: Two research programs helping to bridge the gaps. *Biologically Inspired Cognitive Architectures*, 14:97–110, 2015. doi: 10.1016/j.bica.2015.09.003.

J. Bird and D. Stokes. Evolving minimally creative robots. In *Proceedings of the 3rd International Joint Workshop on Computational Creativity*, pages 1–5, 2006.

S. J. Bradtke and M. O. Duff. Reinforcement learning methods for continuous-time Markov decision problems. In *Advances in Neural Information Processing Systems*, pages 393–400, 1995.

R. I. Brafman and M. Tennenholtz. R-MAX – a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3: 213–231, 2003. doi: 10.1162/153244303765208377.

A. Bundy, G. Luger, M. Stone, and R. Welham. MECHO: year one. In *Proceedings of the 2nd Summer Conference on Artificial Intelligence and Simulation of Behaviour*, pages 94–103, 1976.

E. Cambria and B. White. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57, 2014. doi: 10.1109/MCI.2014.2307227.

J. W. Charnley, A. Pease, and S. Colton. On the notion of framing in computational creativity. In *Proceedings of the 2nd International Conference on Computational Creativity*, pages 77–81, 2012.

T. R. Colin, T. Belpaeme, A. Cangelosi, and N. Hemion. Hierarchical reinforcement learning as creative problem solving. *Robotics and Autonomous Systems*, 86:196–206, 2016. doi: 10.1016/j.robot.2016.08.021.

S. Colton. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems*, 2008.

S. Colton, J. W. Charnley, and A. Pease. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the Second International Conference on Computational Creativity*, pages 90–95, 2011.

M. Cook and S. Colton. Automated collage generation-with more intent. In *Proceedings of the 2nd International Conference on Computational Creativity*, pages 1–3. Citeseer, 2011.

C. Daniel, G. Neumann, O. Kroemer, and J. Peters. Hierarchical relative entropy policy search. *The Journal of Machine Learning Research*, 17(1):3190–3239, 2016. doi: 10.5555/2946645.3007046.

P. Dayan and G. E. Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 271–271, 1993.

D. C. Dennett. *Cognitive wheels: The frame problem of AI*. Routledge/Taylor & Francis Group, 2006.

T. G. Dietterich. State abstraction in MAXQ hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 994–1000, 2000a.

T. G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *The International Journal of Artificial Intelligence Research*, 13:227–303, 2000b.

T. Erez and W. D. Smart. What does shaping mean for computational reinforcement learning? In *7th IEEE International Conference on Development and Learning*, pages 215–219. IEEE, 2008. doi: 10.1109/DEVLRN.2008.4640832.

C. Florensa, D. Held, M. Wulfmeier, M. Zhang, and P. Abbeel. Reverse curriculum generation for reinforcement learning. *arXiv preprint arXiv:1707.05300*, 2017.

J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145, 2016.

M. Garnelo, K. Arulkumaran, and M. Shanahan. Towards deep symbolic reinforcement learning. *arXiv preprint arXiv:1609.05518*, 2016.

M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 8(5-6):359–483, 2015. doi: 10.1561/2200000049.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning. The MIT Press, 2016. ISBN 9780262035613.

S. Gu, S. Levine, I. Sutskever, and A. Mnih. Muprop: Unbiased backpropagation for stochastic neural networks. *arXiv preprint arXiv:1511.05176*, 2015.

C. Guckelsberger, C. Salge, and S. Colton. Addressing the “why?” in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the 8th International Conference on Computational Creativity*, 2017.

P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2): 100–107, 1968. doi: 10.1109/TSSC.1968.300136.

H. V. Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, pages 2613–2621, 2010.

P. J. Hayes. The frame problem and related problems in artificial intelligence. In *Readings in Artificial Intelligence*, pages 223–230. Elsevier, 1981.

T. Hester, M. Quinlan, and P. Stone. Generalized model learning for reinforcement learning on a humanoid robot. In *2010 IEEE International Conference on Robotics and Automation*, pages 2369–2374, 2010.

G. E. Hinton. Neural networks for machine learning. Video lectures, 2012.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647.

G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The “wake-sleep” algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995. doi: 10.1126/science.7761831.

X. Huang and J. Weng. Novelty and reinforcement learning in the value system of developmental robots. In *Lund University Cognitive Studies*, 2002.

A. J. Ijspeert, J. Nakanishi, and S. Schaal. Learning attractor landscapes for learning motor primitives. In *Advances in Neural Information Processing Systems*, pages 1547–1554, 2003.

H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000. doi: 10.1162/089976600300015411.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with Gumbel-Softmax. *arXiv preprint arXiv:1611.01144*, 2016.

N. K. Jong. *Structured Exploration for Reinforcement Learning*. PhD thesis, The University of Texas at Austin, 2010.

N. K. Jong and P. Stone. State abstraction discovery from irrelevant state variables. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 752–757, 2005.

N. K. Jong, T. Hester, and P. Stone. The utility of temporal abstraction in reinforcement learning. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, volume 1, pages 299–306, 2008.

A. Jonsson and A. G. Barto. Automated state abstraction for options using the U-tree algorithm. pages 1054–1060, 2001.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, pages 237–285, 1996.

V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

G. Konidaris and A. G. Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. In *Advances in Neural Information Processing Systems*, pages 1015–1023, 2009.

G. Konidaris, S. Kuindersma, R. A. Grupen, and A. G. Barto. Autonomous skill acquisition on a mobile manipulator. In *Proceedings of the 2011 AAAI Conference on Artificial Intelligence*, 2011.

G. Konidaris, S. Kuindersma, R. A. Grupen, and A. Barto. Robot learning from demonstration by constructing skill trees. *The International Journal of Robotics Research*, 31(3):360–375, 2012. doi: 10.1177/0278364911428653.

A. Koop. *Investigating experience: Temporal coherence and empirical knowledge representation*. PhD thesis, University of Alberta, 2008.

J. R. Koza, M. A. Keane, M. J. Streeter, T. P. Adams, and L. W. Jones. Invention and creativity in automated design by means of genetic programming. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 18(03):245–269, 2004. doi: 10.1017/S089006040404017X.

A. S. Lakshminarayanan, S. Sharma, and B. Ravindran. Dynamic action repetition for deep reinforcement learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

A. Lazaric. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer, 2012. doi: 10.1007/978-3-642-27645-3_5.

K. Lee, S. Choi, and S. Oh. Sparse Markov decision processes with causal sparse Tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018. doi: 10.1109/LRA.2018.2800085.

L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the 2006 International Symposium on Artificial Intelligence and Mathematics*, pages 531—-539, 2006.

L.-J. Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3-4):293–321, 1992. doi: 10.1007/BF00992699.

M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, pages 157–163. Elsevier, 1994.

P. Lopes, A. Liapis, and G. N. Yannakakis. Framing tension for game generation. In *Proceedings of the 7th International Conference on Computational Creativity*, 2016.

W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.

M. C. Machado, M. G. Bellemare, and M. Bowling. A Laplacian framework for option discovery in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2295–2304, 2017.

A. R. Mahmood and R. S. Sutton. Representation search through generate and test. In *Workshops at the 27th AAAI Conference on Artificial Intelligence*, 2013.

S. Mannor, I. Menache, A. Hoze, and U. Klein. Dynamic abstraction in reinforcement learning via clustering. In *Proceedings of the 21st International Conference on Machine Learning*, page 71, 2004. doi: 10.1145/1015330.1015355.

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier, 1989.

A. McGovern and A. G. Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. In *Proceedings of the 18th International Conference on Machine Learning*, pages 361—-368, 2001.

A. McGovern and R. S. Sutton. Macro-actions in reinforcement learning: An empirical analysis. *Computer Science Department Faculty Publication Series*, page 15, 1998.

L. R. Medsker. *Hybrid intelligent systems*. Springer Science & Business Media, 2012.

J. H. Metzen. Online skill discovery using graph-based clustering. In *European Workshop on Reinforcement Learning*, pages 77–88, 2013.

M. Minsky. Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, 49(1):8–30, 1961. doi: 10.1109/JRPROC.1961.287775.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015. doi: 10.1038/nature14236.

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1928–1937, 2016.

I. Mordatch and P. Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, volume 99, pages 278–287, 1999.

J. Noble and D. W. Franks. Social learning in a multi-agent system. *Computing and Informatics*, 22(6):561–574, 2012.

I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, pages 4026–4034, 2016.

I. Osband, B. Van Roy, D. Russo, and Z. Wen. Deep exploration via randomized value functions. *arXiv preprint arXiv:1703.07608*, 2017.

P.-Y. Oudeyer and F. Kaplan. How can we define intrinsic motivation? In *Proceedings of the 8th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 93–101, 2008.

P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007. doi: 10.1109/TEVC.2006.890271.

R. Parr and S. Russell. Reinforcement learning with hierarchies of machines. In *Advances in Neural Information Processing Systems*, pages 1043–1049, 1998.

R. E. Parr. *Hierarchical control and learning for Markov decision processes*. PhD thesis, University of California at Berkeley, 1998.

A. Pease and S. Colton. On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In *Proceedings of the AISB symposium on AI and Philosophy*, page 39, 2011b. ISBN 9781908187031.

T. J. Perkins and D. Precup. Using options for knowledge transfer in reinforcement learning. CMPSCI Technical Report 99-34, University of Massachusetts, Amherst, MA, USA, 1999.

J. Peters, K. Mulling, and Y. Altun. Relative entropy policy search. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010.

A. S. Rao and M. P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1991.

R. Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990. doi: 10.1037/0033-295x.97.2.285.

G. A. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, 1994.

C. Salge, C. Glackin, and D. Polani. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer, 2014. doi: 10.1007/978-3-642-53734-9_4.

V. G. Santucci, G. Baldassarre, and M. Mirolli. Intrinsic motivation mechanisms for competence acquisition. In *IEEE international conference on Development and learning and epigenetic robotics (ICDL)*, pages 1–6. IEEE, 2012.

R. Saunders. Multi-agent-based models of social creativity. In C. F. Veale T., editor, *Computational Creativity*, pages 305–326. Springer, Cham, 2019. doi: 10.1007/978-3-319-43610-4_14.

Y. Savas, M. Ornik, M. Cubuktepe, and U. Topcu. Entropy maximization for Markov decision processes under temporal logic constraints. *arXiv preprint arXiv:1807.03223*, 2018.

J. Schmidhuber. Curious model-building control systems. In *IEEE International Joint Conference on Neural Networks*, pages 1458–1463. IEEE, 1991.

J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi: 10.1109/TAMD.2010.2056368.

J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz. Trust region policy optimization. In *Proceeding of the 37th International Conference on Machine Learning*, volume 37, pages 1889–1897, 2015.

D. Silver. Reinforcement learning. UCL course, 2015. Videos available at https://www.youtube.com/watch?v=2pWv7GOvuf0.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. doi: 10.1038/nature16961.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Ö. Şimşek and A. G. Barto. Skill characterization based on betweenness. In *Advances in Neural Information Processing Systems*, pages 1497–1504, 2009.

Ö. Şimşek, A. P. Wolfe, and A. G. Barto. Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the 22nd International Conference on Machine learning*, pages 816–823. ACM, 2005.

S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38(3):287–308, 2000. doi: 10.1023/A:1007678930559.

S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2 (2):70–82, 2010. doi: 10.1109/TAMD.2010.2051031.

S. Singh, M. James, and M. Rudary. Predictive state representations: A new theory for modeling dynamical systems. *arXiv preprint arXiv:1207.4167*, 2012.

B. D. Smith and G. E. Garnett. Reinforcement learning and the creative, automated music improviser. In *International Conference on Evolutionary and Biologically Inspired Music and Art*, pages 223–234. Springer, 2012b. doi: 10.1007/978-3-642-29142-5_20.

R. Sosa and J. S. Gero. Creative social systems. In *AAAI spring symposium: Creative intelligent systems*, pages 90–94, 2008.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. doi: 10.5555/2627435.2670313.

P. Stone and D. McAllester. An architecture for action selection in robotic soccer. In *Proceedings of the 5th International Conference on Autonomous agents*, pages 316–323, 2001.

R. Sun and F. Alexandre. *Connectionist-symbolic integration: From unified to hybrid approaches*. Psychology Press, 2013. ISBN 9780805823493.

Y. Sun, F. Gomez, and J. Schmidhuber. Planning to be surprised: Optimal Bayesian exploration in dynamic environments. In *Artificial General Intelligence*, pages 41–51. Springer, 2011. doi: 10.1007/978-3-642-22887-2_5.

R. S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the 7th International Conference on Machine Learning*, pages 216–224, 1990.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press Cambridge, 1998. ISBN 9780262193986.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 2 edition, 2018. ISBN 9780262039246.

R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, 1999. doi: 10.1016/S0004-3702(99)00052-1.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

A. Tamar, Y. Wu, G. Thomas, S. Levine, and P. Abbeel. Value iteration networks. In *Advances in Neural Information Processing Systems*, pages 2154–2162, 2016.

M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the 10th International Conference on Machine Learning*, pages 330–337, 1993.

F. Tanaka and M. Yamamura. Multitask reinforcement learning on the distribution of MDPs. In *Proceedings of the 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, volume 3, pages 1108–1113. IEEE, 2003. doi: 10.1109/CIRA.2003.1222152.

M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009. doi: 10.5555/1577069.1755839.

G. Tesauro. TD-gammon: A self-teaching backgammon program, achieves master-level play. Technical report, AAAI Technical Report FS-93-02, 1993.

S. B. Thrun. Efficient exploration in reinforcement learning. Technical Report CMU-CS-92-102, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PE, 1992.

S. B. Thrun. Lifelong learning algorithms. In *Learning to Learn*, pages 181–209. Springer, 1998. doi: 10.1007/978-1-4615-5529-2_8.

N. Tishby and D. Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011. doi: 10.1007/978-1-4419-1452-1_19.

D. S. Touretzky and L. M. Saksida. Operant conditioning in skinnerbots. *Adaptive Behavior*, 5(3-4):219–247, 1997. doi: 10.1177/105971239700500302.

H. Van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double Q-learning. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3540–3549, 2017.

C. M. Vigorito and A. G. Barto. Hierarchical representations of behavior for efficient creative search. In *AAAI Spring Symposium: Creative Intelligent Systems*, pages 135–141, 2008.

O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, J. Quan, S. Gaffney, S. Petersen, K. Simonyan, T. Schaul, H. van Hasselt, D. Silver, T. P. Lillicrap, K. Calderone, P. Keet, A. Brunasso, D. Lawrence, A. Ekermo, J. Repp, and R. Tsing. Starcraft II: A new challenge for reinforcement learning. *CoRR*, abs/1708.04782, 2017. URL http://arxiv.org/abs/1708.04782.

O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, T. Ewalds, D. Horgan, M. Kroiss, I. Danihelka, J. Agapiou, J. Oh, V. Dalibard, D. Choi, L. Sifre, Y. Sulsky, S. Vezhnevets, J. Molloy, T. Cai, D. Budden, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, T. Pohlen, Y. Wu, D. Yogatama, J. Cohen, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, C. Apps, K. Kavukcuoglu, D. Hassabis, and D. Silver. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/, 2019.

K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27 (12):2591–2600, 2017. doi: 10.1109/TCSVT.2016.2589879.

Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*, 2015.

C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992. doi: 10.1007/BF00992698.

E. Wiewiora. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, 2003.

G. A. Wiggins. Searching for computational creativity. *New Generation Computing*, 24(3): 209–222, 2006b. doi: 10.1007/BF03037332.

G. A. Wiggins. Creativity, information, and consciousness: The information dynamics of thinking. *Physics of Life Reviews*, 2018. doi: https://doi.org/10.1016/j.plrev.2018.05.001.

G. A. Wiggins and J. Forth. Idyot: a computational theory of creativity as everyday reasoning from learned information. In *Computational Creativity Research: Towards Creative Machines*, pages 127–148. Springer, 2015.

R. J. Williams. On the use of backpropagation in associative reinforcement learning. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 1, pages 263–270, 1988. doi: 10.1109/ICNN.1988.23856.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. doi: 10.1007/BF00992696.

Y. Wu, E. Mansimov, R. B. Grosse, S. Liao, and J. Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. In *Advances in Neural Information Processing Systems*, pages 5279–5288, 2017.

Psychology and Neuroscience

R. A. Adams, S. Shipp, and K. J. Friston. Predictions not commands: active inference in the motor system. *Brain Structure and Function*, 218(3):611–643, 2013. doi: 10.1126/science.275.5306.1593.

J. R. Anderson, J. F. Anderson, J. L. Ferris, J. M. Fincham, and K.-J. Jung. Lateral inferior prefrontal cortex and anterior cingulate cortex are engaged at different stages in the solution of insight problems. *Proceedings of the National Academy of Sciences*, 106(26): 10799–10804, 2009. doi: 10.1073/pnas.0903953106.

I. K. Ash, P. J. Cushen, and J. Wiley. Obstacles in investigating the role of restructuring in insightful problem solving. *The Journal of Problem Solving*, 2(2):3, 2009. doi: 10.7771/1932-6246.1056.

P. M. Auble, J. J. Franks, and S. A. Soraci. Effort toward comprehension: Elaboration or "Aha"? *Memory & Cognition*, 7(6):426–434, 1979. doi: 10.3758/BF03198259.

L. Aziz-Zadeh, J. T. Kaplan, and M. Iacoboni. "Aha!": The neural correlates of verbal insight solutions. *Human Brain Mapping*, 30(3):908–916, 2009. doi: 10.1002/hbm.20554.

D. Badre. Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5):193–200, 2008. doi: 10.1016/j.tics.2008.02.004.

M. Batey and A. Furnham. Creativity, intelligence, and personality: A critical review of the scattered literature. *Genetic, Social, and General Psychology Monographs*, 132(4): 355–429, 2006. doi: 10.3200/MONO.132.4.355-430.

R. E. Beaty, M. Benedek, S. B. Kaufman, and P. J. Silvia. Default and executive network coupling supports creative idea production. *Scientific reports*, 5:10964, 2015. doi: 10.1038/srep10964.

M. Bilalić, P. McLeod, and F. Gobet. Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition*, 108(3):652–661, 2008. doi: 10.1016/j.cognition.2008.05.005.

H. G. Birch. The relation of previous experience to insightful problem-solving. *Journal of Comparative Psychology*, 38(6):367, 1945.

C. D. Bird and N. J. Emery. Rooks use stones to raise the water level to reach a floating worm. *Current Biology*, 19(16):1410–1414, 2009. doi: 10.1016/j.cub.2009.07.033.

N. Boot, M. Baas, S. van Gaal, R. Cools, and C. K. De Dreu. Creative cognition and dopaminergic modulation of fronto-striatal networks: Integrative review and research agenda. *Neuroscience & Biobehavioral Reviews*, 78:13–23, 2017. doi: 10.1016/j.neubiorev.2017.04.007.

M. M. Botvinick. Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 22(6):956–962, 2012. doi: 10.1016/j.conb.2012.05.008.

M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen. Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624, 2001. doi: 10.1037/0033-295X.108.3.624.

E. M. Bowden and M. Jung-Beeman. Aha! insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin & Review*, 10(3):730–737, 2003a. doi: 10.3758/BF03196539.

E. M. Bowden and M. Jung-Beeman. Normative data for 144 compound remote associate problems. *Behavior Research Methods*, 35(4):634–639, 2003b. doi: 10.3758/BF03195543.

E. M. Bowden, M. Jung-Beeman, J. Fleck, and J. Kounios. New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9(7):322–328, 2005. doi: 10.1016/j.tics.2005.05.012.

M. Brandimonte, G. Einstein, and M. McDaniel. *Prospective Memory: Theory and Applications*. L. Erlbaum, 1996. ISBN 9780805815368.

R. Brisch, A. Saniotis, R. Wolf, H. Bielau, H.-G. Bernstein, J. Steiner, B. Bogerts, K. Braun, Z. Jankowski, J. Kumaratilake, et al. The role of dopamine in schizophrenia from a neurobiological and evolutionary perspective: old fashioned, but still in vogue. *Frontiers in Psychiatry*, 5:47, 2014. doi: 10.3389/fpsyt.2014.00047.

P. W. Burgess, G. Gonen-Yaacovi, and E. Volle. Functional neuroimaging studies of prospective memory: What have we learnt so far? *Neuropsychologia*, 49(8):2246 – 2257, 2011. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2011.02.014. Neuropsychology of Prospective Memory.

D. T. Campbell. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review*, 67(6):380, 1960b. doi: 10.1037/h0040373.

C. Cerruti and G. Schlaug. Anodal transcranial direct current stimulation of the prefrontal cortex enhances complex verbal associative thought. *Journal of cognitive neuroscience*, 21 (10):1980–1987, 2009. doi: 10.1162/jocn.2008.21143.

R. P. Chi and A. W. Snyder. Facilitate insight by non-invasive brain stimulation. *PloS One*, 6 (2):e16655, 2011. doi: 10.1371/journal.pone.0016655.

R. P. Chi and A. W. Snyder. Brain stimulation enables the solution of an inherently difficult problem. *Neuroscience Letters*, 515(2):121–124, 2012. doi: 10.1016/j.neulet.2012.03.012.

T. R. Colin and T. Belpaeme. Reinforcement learning and insight in the artificial pigeon. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 1533–1539, 2019.

G. Cona, C. Scarpazza, G. Sartori, M. Moscovitch, and P. S. Bisiacchi. Neural bases of prospective memory: A meta-analysis and the “Attention to Delayed Intention” (AtoDI) model. *Neuroscience & Biobehavioral Reviews*, 52:21 – 37, 2015. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2015.02.007.

R. G. Cook and C. Fowler. “insight” in pigeons: absence of means–end processing in displacement tests. *Animal cognition*, 17(2):207–220, 2014. doi: 10.1007/s10071-013-0653-8.

D. H. Cropley, J. C. Kaufman, and A. J. Cropley. Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal*, 20(2):105–115, 2008. doi: 10.1080/10400410802059424.

M. Csikszentmihalyi. *Creativity: Flow and the Psychology of Discovery and Invention*. Harper Perennial Modern Classics. HarperCollins, 2009. ISBN 9780061844034.

M. Csikszentmihalyi. Society, culture, and person: A systems view of creativity. In *The Systems Model of Creativity*, pages 47–61. Springer, 2014. doi: 10.1007/978-94-017-9085-7_4.

P. Damier, E. Hirsch, Y. Agid, and A. Graybiel. The substantia nigra of the human brain: II. patterns of loss of dopamine-containing neurons in Parkinson's disease. *Brain*, 122(8): 1437–1448, 1999. doi: 10.1093/brain/122.8.1437.

T. Dandan, Z. Haixue, L. Wenfu, Y. Wenjing, Q. Jiang, and Z. Qinglin. Brain activity in using heuristic prototype to solve insightful problems. *Behavioural Brain Research*, 253: 139–144, 2013. doi: 10.1016/j.bbr.2013.07.017.

A. H. Danek and J. Wiley. What about false insights? Deconstructing the Aha! Experience along its multiple dimensions for correct and incorrect solutions separately. *Frontiers in Psychology*, 7:2077, 2017. doi: 10.3389/fpsyg.2016.02077.

A. H. Danek, T. Fraps, A. von Müller, B. Grothe, and M. Öllinger. Aha! experiences leave a mark: facilitated recall of insight solutions. *Psychological Research*, 77(5):659–669, 2013. doi: 10.1007/s00426-012-0454-8.

A. H. Danek, T. Fraps, A. von Müller, B. Grothe, and M. Öllinger. It's a kind of magic – what self-reports can reveal about the phenomenology of insight problem solving. *Frontiers in Psychology*, 5:1408, 2014. doi: 10.3389/fpsyg.2014.01408.

N. D. Daw, J. P. O'doherty, P. Dayan, B. Seymour, and R. J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876, 2006. doi: 10.1038/nature04766.

A. Dietrich and R. Kanso. A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological Bulletin*, 136(5):822, 2010. doi: 10.1037/a0019749.

R. J. Dolan and P. Dayan. Goals and habits in the brain. *Neuron*, 80(2):312–325, 2013. doi: 10.1016/j.neuron.2013.09.007.

K. Duncker and L. S. Lees. On problem-solving. *Psychological Monographs*, 58(5):i, 1945. doi: 10.1037/h0093599.

R. Epstein. Resurgence of previously reinforced behavior during extinction. *Behaviour Analysis Letters*, 3(6):391–397, 1983.

R. Epstein. The spontaneous interconnection of three repertoires. *The Psychological Record*, 35:131–141, 04 1985. doi: 10.1007/BF03394917.

R. Epstein. The spontaneous interconnection of four repertoires of behavior in a pigeon (Columba livia). *Journal of Comparative Psychology*, 101(2):197, 1987.

R. Epstein. Skinner, creativity, and the problem of spontaneous behavior. *Psychological Science*, 2(6):362–370, 1991. doi: 10.1111/j.1467-9280.1991.tb00168.x.

R. Epstein. On the orderliness of behavioral variability: Insights from generativity theory. *Journal of Contextual Behavioral Science*, 3(4):279–290, 2014. doi: 10.1016/j.jcbs.2014.08.004.

R. Epstein and B. F. Skinner. Resurgence of responding after the cessation of response-independent reinforcement. *Proceedings of the National Academy of Sciences*, 77(10): 6251–6253, 1980. doi: 10.1073/pnas.77.10.6251.

R. Epstein, C. E. Kirshnit, R. P. Lanza, and L. C. Rubins. "insight" in the pigeon: antecedents and determinants of an intelligent performance. *Nature*, 308:61–62, 1984. doi: 10.1038/308061a0.

R. Epstein, C. E. Kirshnit, R. P. Lanza, and L. C. Rubins. A pigeon solves the classic box-and-banana problem. Video uploaded on Youtube, 2007. http://www.youtube.com/watch?v=mDntbGRPeEU.

K. A. Ericsson and H. A. Simon. Verbal reports as data. *Psychological Review*, 87(3):215, 1980. doi: 10.1037/0033-295X.87.3.215.

J. A. Feldman and D. H. Ballard. Connectionist models and their properties. *Cognitive Science*, 6(3):205–254, 1982. doi: 10.1016/S0364-0213(82)80001-3.

S. Fischer, M. Hallschmid, A. L. Elsner, and J. Born. Sleep forms memory for finger skills. *Proceedings of the National Academy of Sciences*, 99(18):11987–11991, 2002. doi: 10.1073/pnas.182178199.

W. W. Fisher, C. C. Piazza, and H. S. Roane. *Handbook of Applied Behavior Analysis*. Guilford Press, 2011. ISBN 9781609185039.

J. I. Fleck and R. W. Weisberg. Insight versus analysis: Evidence for diverse methods in problem solving. *Journal of Cognitive Psychology*, 25(4):436–463, 2013. doi: 10.1080/20445911.2013.779248.

P. Foerder, M. Galloway, T. Barthel, D. E. Moore III, and D. Reiss. Insightful problem solving in an asian elephant. *PloS One*, 6(8):e23251, 2011. doi: 10.1371/journal.pone.0023251.

K. J. Friston and S. Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221, 2009. doi: 10.1098/rstb.2008.0300.

K. J. Friston, J. Daunizeau, and S. J. Kiebel. Reinforcement learning or active inference? *PloS One*, 4(7):e6421, 2009. doi: 10.1371/journal.pone.0006421.

K. J. Friston, M. Lin, C. D. Frith, G. Pezzulo, J. A. Hobson, and S. Ondobaka. Active inference, curiosity and insight. *Neural Computation*, 29(10):2633–2683, 2017. doi: 10.1162/neco_a_00999.

M. L. Gick and R. S. Lockhart. Cognitive and affective components of insight. In R. J. Sternberg and J. E. Davidson, editors, *The Nature of Insight*, pages 197–228. The MIT Press, 1995.

K. J. Gilhooly, L. J. Ball, and L. Macchi. Insight and creative thinking processes: Routine and special. *Thinking & Reasoning*, 21(1):1–4, 2015. doi: 10.1080/13546783.2014.966758.

V. P. Glăveanu. Rewriting the language of creativity: The five A's framework. *Review of General Psychology*, 17(1):69, 2013. doi: 10.1037/a0029528.

V. Goel and O. Vartanian. Dissociating the roles of right ventral lateral and dorsal lateral prefrontal cortex in generation and maintenance of hypotheses in set-shift problems. *Cerebral Cortex*, 15(8):1170–1177, 2004. doi: 10.1093/cercor/bhh217.

J. P. Guilford. Creativity. *American Psychologist*, 5:444, 1950. doi: 10.1037/h0063487.

J. P. Guilford. The structure of intellect. *Psychological Bulletin*, 53(4):267, 1956b. doi: 10.1037/h0040755.

S. N. Haber and B. Knutson. The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, 35(1):4–26, 2010. doi: 10.1038/npp.2009.129.

P. Haggard and B. Libet. Conscious intention and brain activity. *Journal of Consciousness Studies*, 8(11):47–64, 2001.

D. Hanus, N. Mendes, C. Tennie, and J. Call. Comparing the performances of apes (gorilla gorilla, pan troglodytes, pongo pygmaeus) and human children (homo sapiens) in the floating peanut task. *PloS One*, 6(6):e19555, 2011. doi: 10.1371/journal.pone.0019555.

H. F. Harlow. The formation of learning sets. *Psychological Review*, 56(1):51, 1949. doi: 10.1037/h0062474.

D. Hebb. *The organization of behavior: A neuropsychological theory*. John Wiley & Sons, 1949. ISBN 9781135631901.

S. Hélie and R. Sun. Incubation, insight, and creative problem solving: a unified theory and a connectionist model. *Psychological Review*, 117(3):994, 2010. doi: 10.1037/a0019532.

B. A. Hennessey and T. M. Amabile. Creativity. *Annual Review of Psychology*, 61(1): 569–598, 2010. doi: 10.1146/annurev.psych.093008.100416.

C. Heyes. Simple minds: a qualified defence of associative learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1603):2695–2703, 2012. doi: 10.1098/rstb.2012.0217.

C. B. Holroyd and N. Yeung. Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, 16(2):122–128, 2012. doi: 10.1016/j.tics.2011.12.008.

A. M. Isen, K. A. Daubman, and G. P. Nowicki. Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*, 52(6):1122, 1987. doi: 10.1037/0022-3514.52.6.1122.

B. Jacobs, M. Schall, and A. B. Scheibel. A quantitative dendritic analysis of Wernicke's area in humans. II. gender, hemispheric, and environmental factors. *Journal of Comparative Neurology*, 327(1):97–111, 1993. doi: 10.1002/cne.903270108.

S. A. Jelbert, A. H. Taylor, L. G. Cheke, N. S. Clayton, and R. D. Gray. Using the Aesop's fable paradigm to investigate causal understanding of water displacement by New Caledonian crows. *PloS One*, 9(3):e92895, 2014. doi: 10.1371/journal.pone.0092895.

D. Joel, Y. Niv, and E. Ruppin. Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, 15(4-6):535–547, 2002. doi: 10.1016/s0893-6080(02)00047-3.

M. Jung-Beeman, E. M. Bowden, J. Haberman, J. L. Frymiare, S. Arambel-Liu, R. Greenblatt, P. J. Reber, and J. Kounios. Neural activity when people solve verbal problems with insight. *PLoS biology*, 2(4):500–510, 2004. doi: 10.1371/journal.pbio.0020097.

J. W. Kable and P. W. Glimcher. The neurobiology of decision: consensus and controversy. *Neuron*, 63(6):733–745, 2009. doi: 10.1016/j.neuron.2009.09.003.

D. Kahneman and P. Egan. *Thinking, fast and slow*. Farrar, Straus and Giroux New York, 2011. ISBN 9781429969352.

C. A. Kaplan and H. A. Simon. In search of insight. *Cognitive Psychology*, 22(3):374–419, 1990. doi: 10.1016/0010-0285(90)90008-R.

F. Kaplan and P.-Y. Oudeyer. In search of the neural circuits of intrinsic motivation. *Frontiers in Neuroscience*, 1:17, 2007. doi: 10.3389/neuro.01.1.1.017.2007.

J. Kasof. Creativity and breadth of attention. *Creativity Research Journal*, 10(4):303–315, 1997. doi: 10.1207/s15326934crj1004_2.

J. C. Kaufman and R. A. Beghetto. Beyond big and little: The four C model of creativity. *Review of General Psychology*, 13(1):1–12, 2009. doi: 10.1037/a0013688.

T. C. Kershaw and S. Ohlsson. Training for insight: The case of the nine-dot problem. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 489–493, 2001.

G. Knoblich, S. Ohlsson, and G. E. Raney. An eye movement study of insight problem solving. *Memory & Cognition*, 29(7):1000–1009, 2001. doi: 10.3758/BF03195762.

K. Koffka. *Principles of Gestalt Psychology*. Routledge, 1935.

W. Köhler. *Intelligenzprüfungen an Menschenaffen [The mentality of apes]*. Berlin: Springer-Verlag, 1921.

W. Köhler. Gestalt psychology today. *American Psychologist*, 14(12):727, 1959.

J. Kounios and M. Beeman. The cognitive neuroscience of insight. *Annual Review of Psychology*, 65:71–93, 2014. doi: 10.1146/annurev-psych-010213-115154.

J. Kounios and M. Beeman. *The Eureka Factor: Creative Insights and the Brain*. Random House, 2015. ISBN 9781446473344.

J. Kounios, J. L. Frymiare, E. M. Bowden, J. I. Fleck, K. Subramaniam, T. B. Parrish, and M. Jung-Beeman. The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. *Psychological Science*, 17(10):882–890, 2006. doi: 10.1111/j.1467-9280.2006.01798.x.

J. Kounios, J. I. Fleck, D. L. Green, L. Payne, J. L. Stevenson, E. M. Bowden, and M. Jung-Beeman. The origins of insight in resting-state brain activity. *Neuropsychologia*, 46(1): 281–291, 2008. doi: 10.1016/j.neuropsychologia.2007.07.013.

S. Kyaga, P. Lichtenstein, M. Boman, C. Hultman, N. Långström, and M. Landén. Creativity and mental disorder: family study of 300,000 people with severe mental disorder. *The British Journal of Psychiatry*, 199(5):373–379, 2011. doi: 10.1192/bjp.bp.110.085316.

S. Lang, N. Kanngieser, P. Jaśkowski, H. Haider, M. Rose, and R. Verleger. Precursors of insight in event-related brain potentials. *Journal of Cognitive Neuroscience*, 18(12): 2152–2166, 2006. doi: 10.1162/jocn.2006.18.12.2152.

E. Lhommée, A. Batir, J.-L. Quesada, C. Ardouin, V. Fraix, E. Seigneuret, S. Chabardès, A.-L. Benabid, P. Pollak, and P. Krack. Dopamine and the biology of creativity: lessons from Parkinson's disease. *Frontiers in Neurology*, 5:55, 2014. doi: 10.3389/fneur.2014.00055.

X. Liu, J. Hairston, M. Schrier, and J. Fan. Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, 35(5):1219–1236, 2011. doi: 10.1016/j.neubiorev.2010.12.012.

Z.-H. Liu and S. Ikemoto. The midbrain Raphe nuclei mediate primary reinforcement via GABA_A receptors. *European Journal of Neuroscience*, 25(3):735–743, 2007. doi: 10.1111/j.1460-9568.2007.05319.x.

F. Loesche. *Investigating the Moment when Solutions emerge in Problem Solving*. PhD thesis, University of Plymouth, 2018.

K. Louie and M. A. Wilson. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29(1):145–156, 2001. doi: 10.1016/S0896-6273(01)00186-6.

T. I. Lubart. Models of the creative process: Past, present and future. *Creativity Research Journal*, 13(3-4):295–308, 2001. doi: 10.1207/S15326934CRJ1334_07.

T. I. Lubart. The 7 C's of creativity. *The Journal of Creative Behavior*, 51(4):293–296, 2017. doi: 10.1002/jocb.190.

A. S. Luchins. Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs*, 54(6):i, 1942. doi: 10.1037/h0093502.

J. Luo, K. Niki, and S. Phillips. The function of the anterior cingulate cortex (ACC) in the insightful solving of puzzles: The ACC is activated less when the structure of the puzzle is known. *Journal of Psychology in Chinese Societies*, 5(2):195–213, 2004b.

J. Luo, K. Niki, and S. Phillips. Neural correlates of the 'Aha! reaction'. *Neuroreport*, 15 (13):2013–2017, 2004c. doi: 10.1097/00001756-200409150-00004.

J. Luo, W. Li, A. Fink, L. Jia, X. Xiao, J. Qiu, and Q. Zhang. The time course of breaking mental sets and forming novel associations in insight-like problem solving: an ERP investigation. *Experimental Brain Research*, 212(4):583–591, 2011. doi: 10.1007/s00221-011-2761-5.

J. Luo, W. Li, J. Qiu, D. Wei, Y. Liu, and Q. Zhang. Neural basis of scientific innovation induced by heuristic prototype. *PloS One*, 8(1):e49231, 2013. doi: 10.1371/journal.pone.0049231.

J. N. MacGregor and J. B. Cunningham. Rebus puzzles as insight problems. *Behavior Research Methods*, 40(1):263–268, 2008. doi: 10.3758/brm.40.1.263.

J. N. MacGregor, T. C. Ormerod, and E. P. Chronicle. Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):176, 2001. doi: 10.1037//0278-7393.27.1.176.

X.-Q. Mai, J. Luo, J.-H. Wu, and Y.-J. Luo. “aha!” effects in a guessing riddle task: An event-related potential study. *Human Brain Mapping*, 22(4):261–270, 2004. doi: 10.1002/hbm.20030.

N. R. F. Maier. Reasoning in white rats. *Comparative Psychology Monographs*, 1929.

N. R. F. Maier. Reasoning and learning. *Psychological review*, 38(4):332, 1931.

M. Matsumoto and O. Hikosaka. Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, 447(7148):1111, 2007. doi: 10.1038/nature05860.

R. E. Mayer. The search for insight: Grappling with Gestalt psychology’s unanswered questions. In R. J. Sternberg and J. E. Davidson, editors, *the nature of insight*. The MIT Press, 1995.

S. Mednick. The associative basis of the creative process. *Psychological Review*, 69(3):220, 1962. doi: 10.1037/h0048850.

N. Mendes, D. Hanus, and J. Call. Raising the level: orangutans use water as a tool. *Biology Letters*, 3(5):453–455, 2007. doi: 10.1098/rsbl.2007.0198.

P. Merrotsy. A note on big-C Creativity and little-c creativity. *Creativity Research Journal*, 25(4):474–476, 2013. doi: 10.1080/10400419.2013.843921.

J. Metcalfe and D. Wiebe. Intuition in insight and noninsight problem solving. *Memory & cognition*, 15(3):238–246, 1987.

N. Metuki, T. Sela, and M. Lavidor. Enhancing cognitive control components of insight problems solving by anodal tDCS of the left dorsolateral prefrontal cortex. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, 5(2):110–115, 2012. doi: 10.1016/j.brs.2012.03.002.

R. R. Miller, R. C. Barnet, and N. J. Grahame. Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3):363, 1995. doi: 10.1037/0033-2909.117.3.363.

T. Minami, Y. Noritake, and S. Nakauchi. Decreased beta-band activity is correlated with disambiguation of hidden figures. *Neuropsychologia*, 56:9–16, 2014. doi: 10.1016/j.neuropsychologia.2013.12.026.

P. R. Montague and G. S. Berns. Neural economics and the biological substrates of valuation. *Neuron*, 36(2):265–284, 2002. doi: 10.1016/S0896-6273(02)00974-1.

P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16(5):1936–1947, 1996. doi: 10.1523/JNEUROSCI.16-05-01936.1996.

B. R. Moore. The evolution of learning. *Biological Reviews*, 79(2):301–335, 2004. doi: 10.1017/S1464793103006225.

H. B. Neves Filho, L. D. R. Stella, R. H. F. Dicezare, and M. Garcia-Mijares. Insight in the white rat: spontaneous interconnection of two repertoires in Rattus norvegicus. *European Journal of Behavior Analysis*, 16(2):188–201, 2015. doi: 10.1080/15021149.2015.1083283.

A. Newell and H. A. Simon. *Human Problem Solving*, volume 104. Prentice-Hall Englewood Cliffs, NJ, 1972.

J. P. O'Doherty, S. W. Lee, and D. McNamee. The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, 1:94–100, 2015. doi: 10.1016/j.cobeha.2014.10.004.

S. Ohlsson. Restructuring revisited. *Scandinavian Journal of Psychology*, 25(1):65–78, 1984. doi: 10.1111/j.1467-9450.1984.tb01001.x.

S. Ohlsson. Information-processing explanations of insight and related phenomena. *Advances in the Psychology of Thinking*, 1:1–44, 1992.

S. Ohlsson. *Deep learning: How the mind overrides experience*. Cambridge University Press, 2011. ISBN 9781139496759.

M. Öllinger, G. Jones, and G. Knoblich. Investigating the effect of mental set on insight problem solving. *Experimental Psychology*, 55(4):269, 2008. doi: 10.1027/1618-3169.55.4.269.

R. E. Passingham and S. P. Wise. *The neurobiology of the prefrontal cortex: anatomy, evolution, and the origin of insight*. Number 50. Oxford University Press, 2012. ISBN 9780199552917.

P. I. Pavlov. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, 17(3):136, 1927/2010. doi: 10.5214/ans.0972-7531.1017309. Translation of lecture notes; first published in 1927.

J. A. Plucker and R. A. Beghetto. Why not be creative when we enhance creativity? In J. H. Borland, editor, *Rethinking Gifted Education*, pages 215–226. Teachers College Press New York, NY, 2003.

J. Poppenk, M. Moscovitch, A. McIntosh, E. Ozcelik, and F. Craik. Encoding the future: Successful processing of intentions engages predictive brain networks. *NeuroImage*, 49 (1):905 – 913, 2010. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2009.08.049.

J. Qiu, H. Li, Y. Luo, A. Chen, F. Zhang, J. Zhang, J. Yang, and Q. Zhang. Brain mechanism of cognitive conflict in a guessing Chinese logogriph task. *Neuroreport*, 17(6):679–682, 2006. doi: 10.1097/00001756-200604240-00025.

R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79, 1999. doi: 10.1038/4580.

R. A. Rescorla, A. R. Wagner, et al. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2:64–99, 1972.

C. Reverberi, A. Toraldo, S. D'agostini, and M. Skrap. Better without (lateral) frontal cortex? insight problems solved by frontal patients. *Brain*, 128(12):2882–2890, 2005. doi: 10.1093/brain/awh577.

M. Rhodes. An analysis of creativity. *The Phi Delta Kappan*, 42(7):305–310, 1961. ISSN 00317217. doi: 10.2307/20342603.

G. Ruiz and N. Sánchez. Wolfgang Köhler's "the mentality of apes" and the animal psychology of his time. *The Spanish Journal of Psychology*, 17:E69, 2014. doi: 10.1017/sjp.2014.70.

D. E. Rumelhart, J. L. McClelland, P. R. Group, et al. *Parallel distributed processing*. MIT press Cambridge, 1987. ISBN 9780262631129.

M. A. Runco. "Big C, little c" creativity as a false dichotomy: Reality is not categorical. *Creativity Research Journal*, 26(1):131–132, 2014a. doi: 10.1080/10400419.2014.873676.

M. A. Runco. *Creativity: Theories and Themes: Research, Development, and Practice*. Elsevier Science, 2014b. ISBN 9780124105225.

M. A. Runco and G. J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. doi: 10.1080/10400419.2012.650092.

M. F. Rushworth, M. P. Noonan, E. D. Boorman, M. E. Walton, and T. E. Behrens. Frontal cortex and reward-guided learning and decision-making. *Neuron*, 70(6):1054–1069, 2011. doi: 10.1016/j.neuron.2011.05.014.

R. Sawyer. *Explaining Creativity: The Science of Human Innovation*. Oxford University Press, 2011a. ISBN 9780199737574.

M. A. Schilling. A "small-world" network model of cognitive insight. *Creativity Research Journal*, 17(2-3):131–154, 2005. doi: 10.1080/10400419.2005.9651475.

G. Schoenbaum, M. R. Roesch, T. A. Stalnaker, and Y. K. Takahashi. A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature Reviews Neuroscience*, 10(12):885, 2009. doi: 10.1038/nrn2753.

J. W. Schooler and J. Melcher. The ineffability of insight. In . R. A. F. S. M. Smith, T. B. Ward, editor, *The Creative Cognition Approach*, pages 97–133. The MIT Press, 1995.

J. W. Schooler, S. Ohlsson, and K. Brooks. Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122(2):166, 1993. doi: 10.1037/0096-3445.122.2.166.

W. Schultz. Dopamine reward prediction error coding. *Dialogues in clinical neuroscience*, 18(1):23, 2016.

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997. doi: 10.1126/science.275.5306.1593.

M. K. Scullin, M. A. McDaniel, and J. T. Shelton. The dynamic multiprocess framework: Evidence from prospective memory with contextual variability. *Cognitive Psychology*, 67 (1-2):55–71, 2013. doi: 10.1016/j.cogpsych.2013.07.001.

A. M. Seed and N. J. Boogert. Animal cognition: an end to insight? *Current Biology*, 23(2): R67–R69, 2013. doi: 10.1016/j.cub.2012.11.043.

O. Selz. *Über die gesetze des geordneten denkverlaufs: Zur Psychologie der produktiven Denkens und des Irrtums*. Cohen, 1922.

M. J. Sharpe, C. Y. Chang, M. A. Liu, H. M. Batchelor, L. E. Mueller, J. L. Jones, Y. Niv, and G. Schoenbaum. Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, 20(5):735, 2017. doi: 10.1038/nn.4538.

W. Shen, Y. Yuan, C. Liu, and J. Luo. The roles of the temporal lobe in creative insight: an integrated review. *Thinking & Reasoning*, 23(4):321–375, 2017. doi: 10.1080/13546783.2017.1308885.

A. Shenhav, M. M. Botvinick, and J. D. Cohen. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013. doi: 10.1016/j.neuron.2013.07.007.

B. R. Sheth, S. Sandkühler, and J. Bhattacharya. Posterior beta and anterior gamma oscillations predict cognitive insight. *Journal of Cognitive Neuroscience*, 21(7):1269–1279, 2009. doi: 10.1162/jocn.2009.21069.

S. J. Shettleworth. Clever animals and killjoy explanations in comparative psychology. *Trends in Cognitive Sciences*, 14(11):477–481, 2010. doi: 10.1016/j.tics.2010.07.002.

S. J. Shettleworth. Do animals have insight, and what is insight anyway? *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 66(4):217, 2012. doi: 10.1037/a0030674.

H. A. Simon. *The MIT encyclopedia of the cognitive sciences*, chapter Problem solving. MIT Press, 2001. ISBN 9780262731447.

J. S. Simons, M. L. Schölvinck, S. J. Gilbert, C. D. Frith, and P. W. Burgess. Differential components of prospective memory?: Evidence from fMRI. *Neuropsychologia*, 44(8): 1388 – 1397, 2006. doi: 10.1016/j.neuropsychologia.2006.01.005.

D. K. Simonton. Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of Life Reviews*, 7(2):156–179, 2010. doi: 10.1016/j.plrev.2010.02.002.

D. K. Simonton. Creativity and discovery as blind variation: Campbell's (1960) BVSR model after the half-century mark. *Review of General Psychology*, 15(2):158, 2011. doi: 10.1037/a0022912.

D. K. Simonton. Foresight, insight, oversight, and hindsight in scientific discovery: How sighted were Galileo's telescopic sightings? *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):243, 2012. doi: 10.1037/a0027058.

B. F. Skinner. *Science and human behavior*. Simon and Schuster, 1953.

B. F. Skinner. The shaping of phylogenic behavior. *Journal of the Experimental Analysis of Behavior*, 24(1):117–120, 1975. doi: 10.1901/jeab.1975.24-117.

B. F. Skinner. Selection by consequences. *Science*, 213(4507):501–504, 1981. doi: 10.1126/science.7244649.

M. A. Smith, A. Ghazizadeh, and R. Shadmehr. Interacting adaptive processes with different timescales underlie short-term motor learning. *PLoS biology*, 4(6):e179, 2006. doi: 10.1371/journal.pbio.0040179.

G. Sprugnoli, S. Rossi, A. Emmerdorfer, A. Rossi, S.-L. Liew, E. Tatti, G. di Lorenzo, A. Pascual-Leone, and E. Santarnecchi. Neural correlates of eureka moment. *Intelligence*, 2017. doi: 10.1016/j.intell.2017.03.004.

W. D. Stahlman, K. J. Leising, D. Garlick, and A. P. Blaisdell. There is room for conditioning in the creative process: Associative learning and the control of behavioral variability. In . J. C. K. O. Vartanian, A. S. Bristol, editor, *The Neuroscience of Creativity*, pages 45–67. The MIT Press, 2013. doi: 10.7551/mitpress/9780262019583.003.0003.

L. Stanton, E. Davis, S. Johnson, A. Gilbert, and S. Benson-Amram. Adaptation of the Aesop's Fable paradigm for use with raccoons (Procyon lotor): considerations for future application in non-avian and non-primate species. *Animal Cognition*, 20(6):1147–1152, 2017. doi: 10.1007/s10071-017-1129-z.

D. G. Stephen, R. A. Boncoddo, J. S. Magnuson, and J. A. Dixon. The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition*, 37(8):1132–1149, 2009. doi: 10.3758/MC.37.8.1132.

R. J. Sternberg. Investing in creativity: Many happy returns. *Educational Leadership*, 53(4): 80–84, 1996.

R. Stickgold and M. Walker. To sleep, perchance to gain creative insight? *Trends in Cognitive Sciences*, 8(5):191–192, 2004. doi: 10.1016/j.tics.2004.03.003.

K. Subramaniam, J. Kounios, T. B. Parrish, and M. Jung-Beeman. A brain mechanism for facilitation of insight by positive affect. *Journal of Cognitive Neuroscience*, 21(3): 415–432, 2009. doi: 10.1162/jocn.2009.21057.

G. K. Suzanne K. Vosburg. 'Paradoxical' mood effects on creative problem-solving. *Cognition & Emotion*, 11(2):151–170, 1997. doi: 10.1080/026999397379971.

H. Takeuchi, Y. Taki, Y. Sassa, H. Hashizume, A. Sekiguchi, A. Fukushima, and R. Kawashima. Regional gray matter volume of dopaminergic system associate with creativity: Evidence from voxel-based morphometry. *NeuroImage*, 51(2):578 – 585, 2010. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.02.078.

T. Tardif and R. Sternberg. What do we know about creativity? In *The Nature of Creativity: Contemporary Psychological Perspectives*. Cambridge University Press, 1988.

A. H. Taylor and R. D. Gray. Animal cognition: Aesop's fable flies from fiction to fact. *Current Biology*, 19(17):R731–R732, 2009. doi: 10.1016/j.cub.2009.07.055.

A. H. Taylor, B. Knaebe, and R. D. Gray. An end to insight? New Caledonian crows can spontaneously solve problems without planning their actions. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1749):4977–4981, 2012. doi: 10.1098/rspb.2012.1998.

E. Thorndike. Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, 2(4):i, 1898a.

E. Thorndike. Some experiments on animal intelligence. *Science*, 7(181):818–824, 1898b. doi: 10.1126/science.7.181.818.

E. Thorndike. Do animals reason? *Popular Science Monthly*, 55:480, 1899.

F. Tian, S. Tu, J. Qiu, J. Lv, D. Wei, Y. Su, and Q. Zhang. Neural correlates of mental preparation for successful insight problem solving. *Behavioural Brain Research*, 216(2): 626–630, 2011. doi: 10.1016/j.bbr.2010.09.005.

M. Tik, R. Sladky, C. D. B. Luft, D. Willinger, A. Hoffmann, M. J. Banissy, J. Bhattacharya, and C. Windischberger. Ultra-high-field fMRI insights on insight: Neural correlates of the Aha!-moment. *Human Brain Mapping*, 2018. doi: 10.1002/hbm.24073.

E. C. Tolman and C. H. Honzik. Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*, 1930.

E. P. Torrance. The nature of creativity as manifest in its testing. In R. Sternberg, editor, *The Nature of Creativity: Contemporary Psychological Perspectives*, pages 43–75. Cambridge University Press, 1988.

K. M. Tye, J. J. Mirzabekov, M. R. Warden, E. A. Ferenczi, H.-C. Tsai, J. Finkelstein, S.-Y. Kim, A. Adhikari, K. R. Thompson, A. S. Andalman, et al. Dopamine neurons modulate neural encoding and expression of depression-related behaviour. *Nature*, 493 (7433):537–541, 2013. doi: 10.1038/nature11740.

M. Ullsperger, C. Danielmeier, and G. Jocham. Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, 94(1):35–79, 2014. doi: 10.1152/physrev.00041.2012.

J. van Horik and N. J. Emery. Evolution of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(6):621–633, 2011. doi: 10.1002/wcs.144.

V. Venkatraman and S. A. Huettel. Strategic control in decision-making under uncertainty. *European Journal of Neuroscience*, 35(7):1075–1082, 2012. doi: 10.1111/j.1460-9568.2012.08009.x.

R. Verleger, M. Rose, U. Wagner, J. Yordanova, and V. Kolev. Insights into sleep's role for insight: Studies with the number reduction task. *Advances in Cognitive Psychology*, 9(4): 160, 2013. doi: 10.2478/v10053-008-0143-8.

E. Volle, G. Gonen-Yaacovi, A. de Lacy Costello, S. J. Gilbert, and P. W. Burgess. The role of rostral prefrontal cortex in prospective memory: A voxel-based lesion study. *Neuropsychologia*, 49(8):2185 – 2198, 2011. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2011.02.045.

U. Wagner, S. Gais, H. Haider, R. Verleger, and J. Born. Sleep inspires insight. *Nature*, 427 (6972):352–355, 2004. doi: 10.1038/nature02223.

M. P. Walker, T. Brakefield, A. Morgan, J. A. Hobson, and R. Stickgold. Practice with sleep makes perfect: sleep-dependent motor skill learning. *Neuron*, 35(1):205–211, 2002. doi: 10.1016/s0896-6273(02)00746-8.

G. Wallas. *The Art of Thought*. Harcourt, Brace, 1926.

J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

J. X. Wang, Z. Kurth-Nelson, D. Kumaran, D. Tirumala, H. Soyer, J. Z. Leibo, D. Hassabis, and M. Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):860, 2018. doi: 10.1038/s41593-018-0147-8.

M. E. Webb, D. R. Little, and S. J. Cropper. Insight is not in the problem: Investigating insight in problem solving across task types. *Frontiers in Psychology*, 7:1424, 2016. doi: 10.3389/fpsyg.2016.01424.

A. A. Weir, J. Chappell, and A. Kacelnik. Shaping of hooks in New Caledonian crows. *Science*, 297(5583):981–981, 2002. doi: 10.1126/science.1073433.

A. A. S. Weir and A. Kacelnik. A New Caledonian crow (Corvus moneduloides) creatively re-designs tools by bending or unbending aluminium strips. *Animal Cognition*, 9(4):317, 2006. doi: 10.1007/s10071-006-0052-5.

R. W. Weisberg. *Creativity: Genius and Other Myths*. Psychology Series. W.H. Freeman, 1986. ISBN 9780716717683.

R. W. Weisberg. Toward an integrated theory of insight in problem solving. *Thinking & Reasoning*, 21(1):5–39, 2015. doi: 10.1080/13546783.2014.886625.

R. W. Weisberg and J. W. Alba. An examination of the alleged role of "fixation" in the solution of several "insight" problems. *Journal of Experimental Psychology: General*, 110 (2):169, 1981. doi: 10.1037/0096-3445.110.2.169.

M. Wertheimer. The syllogism and productive thinking. In W. D. Ellis, editor, *A source book of Gestalt psychology*, pages 274–282. Kegan Paul, Trench, Trubner & Company, London, 1938.

M. Wertheimer and M. Wertheimer. *Productive thinking*. Harper New York, 1959. ISBN 9783030360627.

G. A. Wiggins, P. Tyack, C. Scharff, and M. Rohrmeier. The evolutionary roots of creativity: mechanisms and motivations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664):20140099, 2015. doi: doi.org/10.1098/rstb.2014.0099.

J. Wiley. Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & cognition*, 26(4):716–730, 1998. doi: 10.3758/BF03211392.

R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, and Y. Niv. Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2):267–279, 2014. doi: 10.1016/j.neuron.2013.11.005.

H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983. doi: 10.1016/0010-0277(83)90004-5.

D. L. Zabelina and M. D. Robinson. Creativity as flexible cognitive control. *Psychology of Aesthetics, Creativity, and the Arts*, 4(3):136, 2010. doi: 10.1037/a0017379.

D. L. Zabelina, L. Colzato, M. Beeman, and B. Hommel. Dopamine and the creative mind: individual differences in creativity are predicted by interactions between dopamine genes DAT and COMT. *PloS One*, 11(1):e0146768, 2016. doi: 10.1371/journal.pone.0146768.

Q. Zhao, Z. Zhou, H. Xu, S. Chen, F. Xu, W. Fan, and L. Han. Dynamic neural network of insight: a functional magnetic resonance imaging study on solving chinese 'chengyu' riddles. *PloS One*, 8(3):e59351, 2013. doi: 10.1371/journal.pone.0059351.

Philosophy and others

G. E. M. Anscombe. *Intention*. Harvard University Press, 1957.

A. Bird and E. Tobin. Natural kinds. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

M. A. Boden. Creativity. In *Artificial Intelligence*, pages 267–291. Elsevier, 1996.

M. A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004.

M. A. Boden. The Turing test and artistic creativity. *Kybernetes*, 39(3):409–413, 2010. doi: 10.1108/03684921011036132.

M. Brand. *Intending and Acting: Toward a Naturalized Action Theory*. Mit Press, 1984. ISBN 9780262022026.

M. Bratman. *Intention, plans, and practical reason*. Harvard University Press, 1987.

M. E. Bratman. What is intention? In M. J. M. E. P. Philip R. Cohen, Jerry L. Morgan, editor, *Intentions in Communication*, pages 15–31. The MIT Press, 1990.

F. Brentano. *Psychology from An Empirical Standpoint*. Routledge Classics. Taylor & Francis, 1874/2014.

R. Carnap. *Logical foundations of probability*. University of Chicago Press, 1950.

H. Castañeda. *Thinking and Doing: The Philosophical Foundations of Institutions*. Springer, 1975.

D. J. Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford University Press, 1996. ISBN 9780195117899.

J. Chandler. Descriptive decision theory. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition, 2017.

N. Chomsky, M. Foucault, and F. Elders. The Chomsky-Foucault debate: On human nature, 1971.

A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. doi: 10.1017/S0140525X12000477.

R. Cohon. Hume's moral philosophy. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.

T. R. Colin. Analyzing ambiguity in the standard definition of creativity. *AVANT*, 8:25–34, 2017.

T. R. Colin. Creativity by any other name. Talk: UK Creativity Researchers' Conference, 2019.

D. Davidson. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963. doi: 10.2307/2023177.

D. Davidson. Intending. In *Philosophy of History and Action*, pages 41–60. Springer, 1978.

D. C. Dennett. *The intentional stance*. MIT press, 1989. ISBN 9780262040938.

European Commission. European year of creativity and innovation, 2009. URL http://www.create2009.europa.eu.

J. Farrell. *The Varieties of Authorial Intention: Literary Theory Beyond the Intentional Fallacy*. Springer, 2017. ISBN 9783319489773.

J. Fodor and S. Crawford. Issues facing contemporary philosophy of mind. Interview conducted for the module "Thoughts and Experience" at The Open University, 2005.

P. Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004. ISBN 9780262071994.

B. Gaut. The philosophy of creativity. *Philosophy Compass*, 5(12):1034–1046, 2010. doi: 10.1111/j.1747-9991.2010.00351.x.

V. P. Glăveanu. Creativity in perspective: A sociocultural and critical account. *Journal of Constructivist Psychology*, 31(2):118–129, 2018. doi: 10.1080/10720537.2016.1271376.

V. P. Glăveanu, M. Hanchett Hanson, J. Baer, B. Barbot, E. P. Clapp, G. E. Corazza, B. Hennessey, J. C. Kaufman, I. Lebuda, T. Lubart, A. Montuori, I. J. Ness, J. Plucker, R. Reiter-Palmon, Z. Sierra, D. K. Simonton, M. S. Neves-Pereira, and R. J. Sternberg. Advancing creativity theory and research: A socio-cultural manifesto. *The Journal of Creative Behavior*, n/a(n/a), 2019. doi: 10.1002/jocb.395.

Google Ngram Viewer. Google books ngram viewer, 2017. URL http://books.google.com/ngrams.

D. T. Gruner and M. Csikszentmihalyi. Engineering creativity in an age of artificial intelligence. In *The Palgrave Handbook of Social Creativity Research*, pages 447–462. Springer, 2019.

R. Holton. *Willing, Wanting, Waiting*. Oxford University Press, 2009.

D. Hume. *A treatise of human nature*. Oxford University Press, 1738/2000. ISBN 9780198751724.

T. Ingold. The creativity of undergoing. *Pragmatics & Cognition*, 22(1):124–139, 2014. doi: 10.1075/pc.22.1.07ing.

I. Kant. *Critique of the Power of Judgment*. Cambridge University Press, 1790/2000. ISBN 9780521348928.

T. S. Kuhn. *The Structure of Scientific Revolutions*. Foundations of the Unity of Science. University of Chicago Press, 1970.

S. Legg and M. Hutter. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications*, 157:17, 2007.

E. Lucas. Calculating machines. *Popular Science Monthly*, 26, February 1885. The article is translated from the Revue Scientifique (numéro 16, 1884), but the illustration is found only in Popular Science Monthly.

C. Lumer. The volitive and the executive function of intentions. *Philosophical Studies*, 166 (3):511–527, 2013. doi: 10.1007/s11098-012-0048-8.

E. Margolis and S. Laurence. Concepts. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2014 edition, 2014.

A. R. Mele and P. K. Moser. Intentional action. *Noûs*, 28(1):39–68, 1994. doi: 10.2307/2215919.

G. A. Percival. *The Nature of Intention*. PhD thesis, University College London, 2014.

Plato and C. Reeve. *Cratylus*. Hackett Classics. Hackett Publishing Company, 1998.

H. Poincaré. *Science et méthode*. Bibliothèque de philosophie scientifique. Flammarion, 1909.

V. Pollio, M. Morgan, and H. Warren. *Vitruvius: the Ten Books on Architecture*. Harvard University Press, 1914.

M. Proust. *Le côté de Guermantes*. À la recherche du temps perdu. Gallimard, 1921. Translation by C. K. Scott Moncrieff, 1925.

M. Ridge. Humean intentions. *American Philosophical Quarterly*, 35(2):157–178, 1998. ISSN 00030481. doi: 10.2307/20009928.

R. Rosaldo, S. Lavie, and K. Narayan. *Creativity/anthropology*. Cornell University Press, 1993. ISBN 9780801422553.

A. Rothenberg. Creative cognitive processes in Kekulé's discovery of the structure of the benzene molecule. *The American Journal of Psychology*, 108(3):419, 1995. doi: 10.2307/1422898.

S. F. Rudofsky and J. H. Wotiz. Psychologists and the dream accounts of August Kekulé. *Ambix*, 35(1):31–38, 1988. doi: 10.1179/amb.1988.35.1.31.

J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980.

J. R. Searle. *Intentionality: An essay in the philosophy of mind*. Cambridge university press, 1983. ISBN 9780521273022.

J. R. Searle. Intentionality and its place in nature. *Synthese*, 61(1):3–16, 1984.

R. J. Seltzer. Influence of Kekulé dream on benzene structure disputed. *Chemical & Engineering News*, 63(44):22–23, 1985.

K. Setiya. Cognitivism about instrumental reason. *Ethics*, 117(4):649–673, 2007. doi: 10.1086/518954.

K. Setiya. Intention. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2015 edition, 2015.

D. K. Simonton. Reverse engineering genius: historiometric studies of superlative talent. *Annals of the New York Academy of Sciences*, 1377(1):3–9, 2016. doi: 10.1111/nyas.13054.

K. Steele and H. O. Stefánsson. Decision theory. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

W. Stukeley. *Memoirs of Sir Isaac Newton's Life*. 1752. Original manuscript available on the Royal Society's "turning the pages" online exhibit.

M. Thompson. *Life and Action*. Harvard University Press, 2008.

G. Townsend. *Three Hundred Aesop's Fables*. May G. Quigley collection. McLoughlin, 1871.

E. N. Trifonov. Vocabulary of definitions of life suggests a definition. *Journal of Biomolecular Structure and Dynamics*, 29(2):259–266, 2011. doi: 10.1080/073911011010524992.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

V. Van Gogh. Letter 437, br. 1990: 438, cl: R44 to Anthon van Rappard. Amsterdam, Van Gogh Museum, inv. nos. b8376 a-b V/2006, 1884. accessible at http://vangoghletters.org/vg/letters/let437/letter.html.

D. Velleman. *Practical Reflection*. Princeton University Press, 1989. ISBN 9781575865348.

W. K. Wimsatt and M. C. Beardsley. The intentional fallacy. *The Sewanee Review*, 54(3): 468–488, 1946.

L. Wittgenstein and C. Diamond. *Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge, 1939*. University of Chicago Press, 2015.

L. Wittgenstein, G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte. *Philosophical Investigations*. Wiley, 1953/2010.

J. H. Wotiz and S. Rudofsky. Kekulé's dreams: fact or fiction? *Chemistry in Britain*, 20: 720–3, 1984.