Tampere University

RENNE PESONEN

Casual Reasoning

A Social Ecological Look at Human Cognition and Common Sense

Tampere University Dissertations 96

Tampere University Dissertations 96

RENNE PESONEN

Casual Reasoning

A Social Ecological Look at Human Cognition and Common Sense

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Social Sciences
of Tampere University,
for public discussion in the auditorium D11
of the Main building, Kalevantie 4, Tampere,
on 17 August 2019, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Social Sciences
Finland

Responsible supervisor	Professor Leila Haaparanta Tampere University Finland	
Supervisors	Dr. Miles MacLeod University of Twente The Netherlands	Professor Petri Ylikoski University of Helsinki Finland
Pre-examiners	Professor Albert Newen Ruhr-Universität Bochum Germany	Docent Otto Lappi University of Helsinki Finland
Opponent	Dr. Wayne Christensen University of Warwick United Kingdom	
Custos	Professor Arto Laitinen Tampere University Finland	

Cover design: Roihu Inc.

ISBN 978-952-03-1171-1 (print)
ISBN 978-952-03-1172-8 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1172-8

PunaMusta Oy – Yliopistopaino
Tampere 2019

Acknowledgements

This dissertation project begun somewhere in December 2010 when Petri Ylikoski suggested that after my MA graduation I should familiarize myself with dual-process theories of cognition. Petri was finishing his temporary post in Tampere (before moving to University of Helsinki, where he is currently located) and Leila Haaparanta was the head of our department at the time until her retirement at the end of 2018. Along with Petri, Leila helped me to get my research started and commented on my work in detail over the years. She has an excellent background in analytic philosophy and her comments on my early work made me to focus on issues concerning the philosophy of language, hence transforming my research from purely an exercise in the philosophy of cognitive science to incorporate issues related to the mind, language, and meaning more broadly.

In 2014 I attended to Kazimierz naturalism workshop in Kazimierz Dolny, Poland (now relocated to Warsaw, then organized by Łukasz Afeltowicz, Marcin Milkowski, and Konrad Talmont-Kaminski) titled "Cognitive science of science," which had a marked impact on my subsequent research. I was doing rather standard philosophy of science related to the dual-process theories; however, as a side project I was also developing ideas on how theoretical reasoning might be more related to intuitive than explicit reasoning, contrary to what seemed to be the received view among dual-process theories. At the workshop I presented my first sketch of these ideas and learned about social reasoning and ritual learning from presentations by Hugo Mercier and Christine Legare who were attending as invited speakers. At the time I did not fully realize the significance of these topics for my later research, but most importantly there I met Miles MacLeod (then working at Helsinki, now at University of Twente) who was interested in supervising my thesis. Under the guidance of Miles, that side project turned into my main research effort and remained in close touch with empirical cognitive sciences. I am indebted to Miles for several invaluable discussions and directions that shaped this work; and apologies for all the constant delays!

In our department's weekly research seminar, some of our analytically minded philosophers raised suspicions about whether my work was actually about concepts or about psychological theories of concept representation. As I started writing in 2012–2013, I saw myself as a heir to analytic tradition but these doubts gradually forced me to shift my alliance toward the phenomenological tradition. I thought that I *was* talking about concepts, language, and

meaning but eventually I realized that my ideas did not resonate with the universalist logical program but rather with ones emphasizing idiosyncratic subjective experience. I absorbed the idea that ”meaning” is not a homonym referring to linguistic content in addition to significance associated with the human condition but semantics is inseparable from everyday pragmatics. Hence the title ”Casual reasoning;” and no, it is not a misspelling of ”causal.”

I thank all my friends and colleagues from our department for valuable criticism and all kinds support and help along the years. I have not been around at the campus much recently but many of our staff were already teaching when I started my bachelor studies in 2002, and hence many of my current colleagues have had a significant impact on my conduct as a philosopher and my conception of philosophy.

Many of my colleagues are also long time friends I studied together with. As I argue in this work, theoretical understanding develops not by absorbing information but by learning how to put that information to use through discussions and debates. Persistent engagements in mutual argumentation and critical exchange of ideas are necessary for philosophical competence to develop and as for myself this is something I eventually owe to my friends—including many old ones outside the academia. Because all the relevant people are far too numerous to list here, I specifically want to thank the few who have directly contributed to the content of this work: Jaakko Belt, Miika Haverinen, Aatu and Risto Koskensilta, several contributors of the häkkikärrymies autonomous think-tank along the years, and members of our small discussion group of empirically oriented philosophers of mind in Finland, especially the founding posse of Valtteri Arstila, Heidi Haanila, Pii Telakivi, and Jaana Virta.

At a critical stage of this research I came to think that an important domain any cognitive theory of commonsense reasoning must address is political thought. Political thinking consist of an interesting constellation of practical and moral values, theoretical ideas and ideals in concert with very concrete and mundane aims pertaining to daily life, influenced by cultural history, science, media, social affiliations, personal experience, and so on. In short, it is a mess with high contextual and idiosyncratic variability where most of the interesting aspects of commonsense reasoning become visible, such as the dynamics of explicit and implicit inference and different ways in conceptualizing shared concepts. This work should not be read as a treatise on the psychology of politics, but for helping me to grasp the convoluted nature of the matter at the right time I want to thank especially Heikki Sirviö, Lasse Poser, and Lotta Tenhunen for many of our equally convoluted discussions. Lengthy conversa-

tions with Olli Herranen have clarified my understanding on how sociology and cognitive psychology could be mutually informative.

Otto Lappi (University of Helsinki) and Albert Newen (University of Bochum) pre-examined this work. I am grateful for their favorable statements and accurate criticism, some of which I hopefully managed to address in the final version of this thesis. I express my gratitude also to Wayne Christensen from University of Warwick, who agreed to act as my opponent. Arto Laitinen, the custos and the current head of our department, has helped me with many practicalities involved with finishing this dissertation. It would be unreasonable to miss the opportunity to thank Arto also for all the professional help over the years as well as all the fun and enlightening conversations in- and outside the conference room.

Last but not least, studying and doing philosophy is peculiar work with its own quirks and eventualities. From all the people who have helped me to navigate the academia and made my life easier along the way I especially want to acknowledge Tommi Vehkavaara and Ari Virtanen. Both have also helped me intellectually. Ari is to thank for keeping me studying mathematics intensively some years back. Without that effort I would not have been able to understand many of the research addressed in this work.

This research has been funded by The University of Tampere, The Finnish Cultural Foundation, Kela, The Finnish Doctoral Programme of Philosophy, and The Academy of Finland through a project belonging to the ERA-NET Neuron consortium "The integration of cross-disciplinary research in neuroscience and social sciences" (INSOSCI).

Abstract

This thesis promotes a pragmatist and ecological approach to human cognition and concepts. Namely, that our conceptual system primarily tracks affordances and other causal properties that have pragmatic relevance to us as embodied and active agents. The bulk of the work aims to show how various research programs in cognitive psychology naturally intersect and complement each other under this theoretical standpoint, which is influenced by enactivist and embodied approaches to cognitive science as well as linguistic pragmatism. The term "ecological" in the title refers to an approach that emphasizes the interaction of agents and their environment and "social ecology" means that this includes social interaction between agents and that our material environment is extensively a cultural product, constantly reproduced and altered through cultural behavior.

The extent of the argument is not supposed to be confined to the theoretical psychology and philosophy of science but has somewhat wider motives pertaining to philosophy of language and knowledge. I do not attempt to reform extant theories of cognitive processing and representation (unless arguments against logical computationalism are still considered reformist these days) but to explain the nature of conceptual understanding. I take it that having a concept is principally not having a particular information structure in one's brain but rather a set of interlocking capacities that support intentional action. In effect, I claim that conceptual understanding should be understood as a cognitive skill and psychological research on concepts should not identify concepts as static information structures but as capacities which are integral parts of procedural knowledge that support skillful know-how in situated action.

Conceptual mental representations deal with information but such information structures are active constructs that cannot be understood without pragmatic and ecological perspective on human cognition. I argue for the claim, earlier proposed for instance by Eleanor Rosch, that contexts or situations are the proper unit that categorization research needs to concentrate on. In accordance with Edouard Machery's well-known claim, I conceive classical category theories of cognitive science, namely prototype, exemplar, and knowledge accounts, to tap real cognitive phenomena; however, *pace* Machery I aim to show that they do not form distinct conceptual representations but rather participate in human conceptual capacities as interlocking component processes.

The main problem with theories that emphasize situated direct interaction with the environment is to explain abstract and symbolic reasoning. One theoretically promising way to resolve the issue is to invoke some version of the dual-process theories of cognition; that is, to explain rule-based, theoretical, and symbolic reasoning by resorting to a distinct cognitive system, which is more or less dedicated to those kind of tasks. While dual-process theories seem to license such a move, they can work only as a partial solution because expert scientific reasoning , for example, necessitates implicit skills just like any area of expertise. Second, commonsense reasoning is partly schematic and utilizes theoretical concepts. As an alternative explanation, I offer a hypothesis influenced by philosophical linguistic pragmatism which posits that discursive reasoning is incrementally learned tacit know-how in cultural praxis, which determines how we understand linguistic concepts. This interactive know-how exploits mostly the same cognitive mechanisms as situated and pragmatic procedural knowledge. The explanation has immediate implications for the analytic philosophy of language. When we interpret a text or engage in conceptual analysis, our conscious conceptual interpretation of the associated contents is a product of implicit processes intimately tied with procedural knowledge; in short, explicit know-that is rooted in implicit know-how.

Contents

1 Introduction

Gideon Keren (2013) opens his critique of dual-process theories of cognitive science "A Tale of Two Systems: A Scientific Advance or a Theoretical Stone Soup?" by reciting a folk tale on how to prepare a delicious soup that requires nothing but a soup stone. All you need is to add a carefully selected soup stone to boiling water. But to make it taste right you might also want to add a little salt and pepper. Naturally, every decent soup calls for some vegetables and to get an excellent result, you are advised to add a little meat too, and so on. After you are done, you can take the stone out and wonder what the moral of the story is.

According to Keren, the recent efforts to understand the workings of human mind as consisting of two principal cognitive systems—one automatic, fast, rigid, and subconscious; and the other controlled, slow, flexible, and conscious—looks pretty much like making a stone soup: You throw a simple and intriguing theoretical idea in and hope it unlocks many mysteries of the human mind. However, you need to adjust details before long and reassess some of the key presumptions. This is how research generally works, of course, but eventually you have to give up on what were supposed to be the core ideas of the theory, and in the end, it may be advisable to throw it away altogether. What you are left with is a somewhat random mixture of elements that may or may not make an agreeable whole. In any case, it is the other ingredients that make the outcome palatable.

The story above may be a somewhat accurate description of this work. It started as a philosophy of science project comparing various dual-process theories with the relevant data gathered in the literature. My aim was to figure out what would be the most theoretically illuminating and empirically motivated way to conceptualize the two systems, given the numerous different characterizations in circulation. The outcome was something quite different: basically a theory of conceptual understanding and tacit reasoning based on a sort of cognitive psychological version of pragmatism.

The result built mainly on ideas developed in enactivist and embodied cognition research, in that cognitive processes and representations are highly contextualized, pragmatic, action oriented, and shaped by our human practices, goals, and needs. The notions of *situation* and *action* rather than *symbol* and *referent* are the keys to understanding human cognition. Concepts are principally devices for action rather than building blocks of thoughts. Psychologically, they are structured as feature clusters gathered around causal prop-

erties. In effect, *the human conceptual system primarily tracks affordances and other causal properties that have pragmatic relevance to us as embodied and active organisms.*

Cognitive theories preceding these embodied trends aimed principally to explain planning, logical reasoning, and general abstract problem-solving. In short, they were classic examples of the *artificial* intelligence approach. They fell short of explaining many interesting aspects of *natural* cognition, such as perceptually guided action. That was not principally their intent, though, but crucially they also always had problems in modeling distinctively human commonsense reasoning. The issue quite likely is attributable to how human conceptual cognition is organized by how we engage with our environment in the ways embraced by enactive approaches.

On the other hand, enactive theories emphasize dynamic sensorimotor loops as the fundamental constituents of intentional action and content. It is hard to turn this idea into working cognitive science because it is difficult how you manage to model theoretical and other symbolic reasoning by taking sensorimotor interaction as a starting point. An obvious and tried solution is bootstrapping—to try to show how increasingly abstract and "intellectual" representations can be built from sensorimotor processes. Another option is to discard the idea of continuity of concrete action and reflective thought and simply postulate a dedicated cognitive system for symbolic reasoning. However *ad hoc* this may sound, dual-process theories apparently license you to do that.

Eventually, I grew dissatisfied with both solutions. Instead, I opted to employ linguistic pragmatism for the task. It is a form of inferentialism where linguistic and other symbolic expressions ultimately receive their contents from how they are used in actual social practices. The public and derivatively private reasoning is taken to be founded on pragmatic discursive skills, which do not track intellectual essences but the causal fabric of material social praxis of giving and asking for reasons. These practices, in turn, may be ultimately constrained by non-social factors relevant to their intrinsic rationale. For example, in science, the discursive acts are not purely linguistic but also involve gathering and analyzing data. Although these practices are fundamentally social innovations, their specific applications are still constrained by the underlying reality of the subject matter of the research.

Taking this route to theoretical and symbolic reasoning tends to blur the lines between thinking and doing and between the abstract and the practical. I also do not maintain strict demarcation between enactivism and linguistic pragmatism. Both theories claim that mental content is produced by how we

use our conceptual resources in our practical undertakings, or perhaps more accurately, how concepts participate in the organism–environment interactions. For my purposes, I consider linguistic pragmatism as an extension of enactivist ideas to cover discursive interactions, and, for the reasons explained in Chapter 2, I find it difficult to understand how linguistic pragmatism can do without enactivism or other complementary theory of intentional content.

For cognitive science, the proposal effectively means that mainly the same cognitive faculties execute concrete skills, commonsense judgment, and abstract theoretical reasoning. These faculties, in turn, are tacit cognitive capacities for immediate coping with the environment, which shape our conceptual system and furnish our explicit thought with meaning. My research problem then transformed into finding an empirically sound account of tacit cognition that can accommodate all this while maintaining reasonable fidelity to key ideas of enactivism and linguistic pragmatism. To that, end I employ fairly standard cognitive psychology. No particularly new proposals about mental representations or processes are introduced but only perhaps a novel way to put together a stock of empirical research on conceptual cognition. Chapters 4 and 5 are dedicated to the empirical subject matter relevant to the working hypothesis of this work. However, in what follows, the philosophical theories of intentional content and empirical theories of cognitive processing are tightly entangled, and hence their treatments spill from one chapter to another. In Chapter 3, I discuss some suggestive reasons why concept research in philosophy and psychology should be entangled. There I will also offer a brief and slightly critical review of classical triad of category theories in cognitive psychology.

Any use theory of meaning holds that use is constitutive of content and, therefore, content and process cannot be strictly separated. In the naturalistic version of pragmatist inferentialism that I am selling, this translates to the claim that you cannot understand conceptual representation without understanding the ecological nature of cognitive processing: You need to account for how, why, and what the cognitive system tracks in its environment to understand what sort of representations it exploits. I hasten to add that I still do not advocate psychological reductionism about conceptual content. My account is interactionalist rather than internalist. Our cognitive and biological constitution with our physical and cultural environments are all genuine determinants

of meaning, and often it is far more illuminating to look at the environment rather than the brain.[1]

Only some basic ideas of dual-process theories remained in the outcome: the distinction between consciously controlled and intuitive modes of cognitive processing, with the assumption of the relative impotence of the former. However, these ideas, especially the distinction, were rediscovered many times in the past and are not unique to modern dual-process theories. Perhaps Keren was right, in that if you dig deep enough, the theory seems to lose its initial allure. Nonetheless, working with it helped me to put lots of previous research together in the way that would be impossible (for me, that is) without taking that theoretical framework seriously. Some of the empirical research discussed in the following chapters has been previously fairly unconnected to dual-process theories.

I often find it often in understanding the point of long arguments if the structure somewhat follows the narrative structure of the discovery of the main ideas. Hence, to illuminate the reader as to why I ended up with the questions and conclusions I did, the introduction first portrays how I started and what went wrong, and then what course the research subsequently took. The discussion of this juncture also provides background information about the expertise and dual-process research relevant to other chapters. The introduction concludes with a brief description of my final research hypothesis whose reasoning is laid out in the rest of the chapters.

1.1 Getting started

Since antiquity, the notions of intuitive (or habitual) and reflective operations of the mind have been around. During the long history of contemplation of the human mind and affairs, somewhat similar observations about this duality have resurfaced. Nonetheless, at least in the Western tradition, deliberative reason has usually been considered paramount either in quantity or quality. Automatic acts have been seen as physiological reflexes with only a supplementary role in human behavior or representing the animal side of us, which the reason needs to keep in check. For example, Schopenhauer (1818/1966)

[1] See Bruner (1990, Chapter 1) for an illuminating discussion about this point, especially on how culture is constitutive to meaning, and hence psychology cannot be strictly individualistic but needs to account for cultural history and other meaning-producing phenomena; at least to the extent that psychology is a science of meaning-making and meaning-using processes (pp. 11–15). This point presumes that there is no strict demarcation of cognitive content and processes; a controversial but defensible claim (e.g. Nisbett et al., 2001).

thought that human conduct is mostly dictated by non-conscious will, which was a pessimistic conviction. The will was a bestial thing—a strong blind man carrying a lame man that can see. The grand rationalist and dualist Descartes was aware that there are forces in the human body that shape our behavior and which are neither conscious volitions nor simple reflexes (Hatfield, 2007). However, for Descartes, these impulses were entirely physiological while for Schopenhauer somewhat obscure metaphysical.[2] The birth of the idea of the psychological unconscious had to wait until the latter half of the 19th century.

The notion of the unconscious mind is most famously associated with Sigmund Freud. He actually can be considered as a sort of predecessor of dual-process theories for giving substantial weight to unconscious processing and characterizing it as associative while describing the conscious mind as analytic. His notion of the *dynamic unconscious* is, however, very different from the idea employed in dual-process theories, which can be termed *adaptive unconscious* (Wilson, 2002). For Freud, the unconscious was the place of repressed emotions and memories that influenced behavior through neuroses and other ailments; however, the adaptive unconscious consists of habituated actions and other implicitly learned associations that automatize by repetition and subside from conscious awareness. The psychological significance of habitual action was already appreciated by Freud's contemporaries, perhaps most famously by William James 1890, a pioneer of empirical psychology and American pragmatism.

The notion of habituation soon became popular and forged the central theoretical doctrines of behaviorism (and associationist psychology in general), which was the dominant form of empirical psychology in North America in the early 20th century, until the 1950s when its importance waned under the pressure of the emerging computational cognitive science. The new computational paradigm maintained that the mind was an information-processing system, executing logical computations over symbolic structures rather than learned stimulus-response associations. While it was clear that a vast amount of information processing that occurred in the brain was beyond our awareness, it was supposed to happen in low-level perceptual analyzers and other subpersonal supplementary systems that supported the explicit central cognition and connected it to the world via sensory and motor organs. The central (or higher) cognition, in turn, was thought to be responsible for personal level

[2] See Frankish & Evans (2009) for a historical overview of themes and developments related to dual-process theories.

information processing—i.e. conceptually structured thought, like judgment, reasoning, and decision making (See e.g. Fodor, 1983).

Although not necessarily anti-computationalistic at heart, modern dual-process theories depart from the above image because many of the automatic processes are assumed to be learned and not limited to lower cognition. Learning happens automatically, and the activation of learned associations become mandatory. The process is adaptive because the resultant capacity acts rapidly and frees limited cognitive resources—such as attention and working memory—from routine tasks and makes cognitive performance more effective in familiar environments. Moreover, automatic processes are considered the default cognitive mechanisms underlying behavior while explicit reasoning is considered far more flexible, although it is too slow and too limited in capacity to steer much of our real-time behavior. Automatized capacities do not execute flawlessly and they often produce predictable errors, especially with unfamiliar tasks. Methodologically, the research has focused mainly on these defects because the error profiles give information about the workings of the intuitive mind. Regardless, the theory effectively switches the conceptions about implicit and explicit mind according to what comes to the primary nature, quality, and quantity, traditionally associated with these notions.

Dual-process theories emerged slowly from the late 1970s, gathered momentum in the 1990s, and hit the mainstream after the turn of the millennium; perhaps this was fueled partly by concurrent, albeit mostly unrelated, theoretical developments. For the reader acquainted with theoretical cognitive science, dual-process theories can be approximately described as an amalgamation of classical computationalism with connectionism. Connectionism is a species of associationist tradition that resurfaced in a new form in the 1980s with the invention of powerful learning algorithms for complex neural networks (Rumelhart et al., 1986; McClelland et al., 1986). After the emergence of these new network models, the 1980s and 1990s witnessed heated debates between the avant-garde connectionists and traditional computationalists. Connectionist networks were well suited to model pattern recognition, fuzzy categorization, associative semantic networks, and in general many ill-structured and vague low-level but important cognitive capacities, which are hard to capture using logic-based systems. The older computationalist models always had an edge on associationism in modeling cognitive competencies that ostensibly require systematic symbolic processing, such as sentence parsing and logical inference; this is especially true for competencies requiring hierarchical sequential operations like complex inference and planning. However, soon network models emerged

that could implement rule-based symbolic processing (Bechtel & Abrahamsen, 2002), and mathematical results established the computational equivalence of neural nets and symbolic models (Siegelmann & Sontag, 1991, 1995).

To cut the long story short, it became less of an issue what these formalisms can accomplish in principle and more of a question what sort of theory and concept formation they promote in cognitive science. If you need to implement sequential symbolic processing anyway, why bother with neural networks and not model these capacities by using the theoretical apparatus of classical computationalism? If you need fast, parallel, and fuzzy learning systems to model perceptual pattern recognition, for example, why not use neural networks that are readily tailored for the task? After all, these are epistemological choices on how to best understand the workings of the mind, or its components, and you need a proper grain size for your theoretical apparatus to make cognitive processes intelligible. Too fine detail—e.g. looking at neuronal details when explaining the psychology of logical inference—may just mask critical higher-level regularities and lead you to fail to see the forest for the trees. Theoretical proposals, albeit perhaps not hugely popular, emerged for hybrid modeling, that is to use mixed systems consisting of both associative and rule-based computation.[3] There has been a somewhat vague and long-standing demarcation between higher and lower cognitive processes (e.g. reasoning and perception, respectively) and it seems reasonable, at least as a first approximation, to use different tools to model different aspects of cognition and see if you can integrate them into a working unified theory at some point.

Meanwhile, on the empirical front, these sorts of hybrid ideas were already gaining ground. Their similarity to the mentioned theoretical developments was on the observation that there are qualitatively two kinds of cognitive processes: (a) quick and dirty, associative, implicit, and automatic and (b) clean, analytic, explicit, and controlled. However, these two modes of cognitive processing were not assumed to align with the prevailing distinction between lower and higher cognition. Perhaps some lower-level processes are exclusively associative and cognitively impenetrable (e.g. perception) and some higher-level processes are sequential and consciously controlled (e.g. logical inference and planning), but the emerging picture was that actually the mind executes the very same tasks in two different ways often at the same time. An influential study by Walter Schneider and Richard Shiffrin (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977) suggested that there are generic cognitive

[3] E.g. Clark (1989); Harnad (1990); Clark & Karmiloff-Smith (1993); Sun (2002).

processes—namely detection, search, and attention—that have this sort of dual nature. Around the same time it was also discovered that this might be true of simple reasoning tasks (Wason & Evans, 1975). Two decades later more encompassing studies about the dual-nature of human reasoning were carried out (e.g. Evans & Over 1996; Sloman 1996; Stanovich 1999), and the idea proliferated to judgment and decision making (Gilovich et al., 2002; Kahneman, 2011), learning (Reber, 1993), social and moral psychology (Wilson, 2002; Smith & Collins, 2009; Greene, 2002), and neuroscience (Goel, 2005, 2007).

Given the various independently discovered but highly similar proposals, the research was clearly onto something but what exactly? The results came from various subfields of cognitive psychology, and while they all pointed out that there were roughly fast and intuitive versus slow and deliberative processes at play, different researchers tended to characterize these processes quite differently. The most intriguing hypothesis is quite evident, I suppose: That the human mind is composed of two major subsystems of *higher* cognition, and the dual nature of different competencies reflect this global arrangement (Evans & Over, 1996; Stanovich, 1999). Some even went so far to claim that we effectively have two minds (Evans, 2003).

Since the different characterizations of these two systems were rarely contradictory but often orthogonal, this proved to be a great place for some practical philosophy of science; viz. to take a careful look at the empirical results in order to (a) check if there really is any common core that supports the strong two-systems (or two-minds) hypothesis, (b) to regiment the prevailing conceptual chaos, and (c) to proceed to find out what aspects of the mind can be brought under the umbrella of dual-process theories under rigorous analysis—e.g., do emotions have this sort of dual nature, or do they exclusively belong to the automatic system or outside the scope of the two-systems theory altogether?

My principal aim was plank (b). The enthusiasm for dual-process theories seemed partly to stem from the prevailing conceptual vagueness. There are virtually perennial demarcations that seem to characterize the mind and are important in epistemology and philosophy of language: intuition versus reason, habitual versus deliberate, practical versus theoretical, linguistic versus imagistic, desire versus commitment, tacit versus explicit, and so on. It seemed to me that the existing elliptic descriptions of the two systems were at least partly drawn from these kinds of pretheoretical conceptualizations of mental faculties. If these pretheoretical intuitions could be cashed out empirically, we could then have a scientific theory for illuminating the factual nature of these dualisms

and perhaps show their proper place in the philosophy of mind, knowledge, and language. Since the systems were often depicted with a rather heuristic clusters of more or less vague properties, in the literature they were given deliberatively austere names "System 1" and "System 2" (Stanovich, 1999). Below is a generic list of characteristics associated with these two systems:[4]

System 1	**System 2**
evolutionary old	evolutionary recent
shared with animals	uniquely human
independent of general intelligence	linked to general intelligence
independent of working memory	depends on executive function
universal	variable across individuals
subpersonal/biased	personal/normative
unconscious/preconscious	conscious
implicit	explicit
intuitive	reflective
automatic	controlled
associative	analytic
parallel	serial
fast	slow
rigid	fluid
high capacity/low effort	low capacity/high effort
perceptual	propositional
holistic pattern detector	compositional structure
nonverbal	linguistic
heuristic	algorithmic
contextual	detached
pragmatic	logical
modular	single system
default process	inhibitory
domain specific	general
experience based	consequential

[4] Adapted from (Wilson, 2002), (Evans, 2008), and Chapters 1,2,5,6, and 13 appearing in an edited volume (Evans, 2009). I have collected lists of attributes mentioned in these papers (which, in turn, are integrated from several earlier works) and left out any features that appear in only one of them; thus the list represents something like a tentative consensus at least in one branch of the research program.

The characteristic properties above are clustered somewhat arbitrarily, and it is not clear which entries should be actually considered as the same. Does for example "unconscious", "implicit", "automatic", and "intuitive" mean the same thing, or are they distinct characteristics? Certainly they are related but are they related conceptually or empirically? These were the problems I was set off to solve.

In my quest to find the common core of the two systems, I dismissed the evolutionary considerations at the outset. It is tempting to think of System 1 as primitive and something that any cognitive creature might need while thinking that controlled reasoning is sophisticated and a distinctively human faculty. However, I found hardly found any sound grounds for demarcating the processes precisely in this way. Indeed, no one knew the nature of these systems (that was the problem), so how can one say that no other animal has Type 2 processes? Some non-human mammals and even birds are capable of something that at least resembles controlled serial cognition (Toates, 2006), and since no one had a a clear picture of the capacities of System 1, it seemed rather premature to claim that those are prevalent in the animal kingdom. True, conditioning and associative learning seem to be shared between humans and other animals, however if *that* is the key characteristic of System 1 processing, then it is perhaps advisable to work with that hypothesis and, for the time being, abstain from making strong claims about its evolutionary history.

Moreover, the putative Systems 1 and 2 apparently do not to appreciate the putative distinction of old and new structures in the brain. For example the so-called *belief bias* in reasoning is a fast and automatic response, characteristic of system 1 processing,[5] which involves the activation of Brodmann areas 11 and 32 in the ventromedial prefrontal cortex (Goel, 2007). According to Jonathan Evans (2008), this shows that not all System 1 processes are ancient since prefrontal cortex is heavily developed specifically in humans. Then again, at least primate mammals do have these areas (Passingham & Wise, 2012, Chapter 2); yet, it is unclear what the brain development implies about their functional differences between the species. But I don't care really. Mostly the dual-process theory taps into higher cognitive functions, which are either

[5] *Belief bias* is the tendency to evaluate the validity of arguments on the basis of the credibility of their conclusion rather than of their logical form (Evans et al., 1983). The phenomenon is relevant to dual-process theories since the implicit dismissal of the form can be overridden at will, which happens easily when the conclusion is unconvincing. Cf. "All plants need water, and roses need water; therefore roses are plants" and "All guns are dangerous, and snakes are dangerous; therefore snakes are guns".

clearly lacking or virtually impossible to study in other animals, and hence these considerations are mostly irrelevant to the empirical evaluation of the theory.

Inheritance and correlations with cognitive ability may be important clues for pinpointing the cognitive basis of System 1 and 2 processes. Especially the latter is significant evidence of the theory, since there is a good correlation between cognitive ability and propensity to Type 2 reasoning, while this correlation is lacking in Type 1 processing (Stanovich & West, 1998; Stanovich, 1999; Stanovich & West, 2000). Nonetheless, since I am interested in cognitive mechanisms, these characteristics are somewhat off-topic. Working memory capacity has been strongly linked to general intelligence (Chooi, 2012), and for my purposes the link between System 2 and cognitive ability serves only to corroborate the actually interesting (i.e. cognitive architectural) connection of system 2 and working memory and other executive functions rather than constituting independently relevant characteristic for system 1/2 demarcation. The above list contains other notions which are not strictly architectural; e.g., being ”fast” versus ”slow”. Still, in combination with what else we know about the mind and the brain, these features are indicative of underlying cognitive processing. For example, an oft-cited argument for connectionism against serial symbolic computationalism is that complex actions that are performed in a few hundred milliseconds must be executed in heavily parallel processing without recursion. This follows simply from our knowledge of neuronal firing latencies (Feldman & Ballard, 1982).

Those considerations out of the way, I sought what most philosophers would probably do to find something to complain about: consciousness. If you take the notion as the so called *phenomenal consciousness*—the stuff made from qualia: the subjective feeling of joy, the experienced redness of red, and so on—you are bound to create more confusion than solve. But ”consciousness” can mean many things, and in dual-process theories the relevant notion is *access consciousness*. We are (generally) aware of our intuitive responses but not how and why we reach them. We are denied an access to the process. Often we are able to explain our intuitive decisions but there are good reasons to believe that the *post hoc* rationalizations often misrepresent the process that generates them (Evans & Wason, 1976; Nisbett & Wilson, 1977; Haidt, 2001). Thus, ”consciousness” *per se* is too ambivalent and potentially confusing; yet properties associated with access consciousness such as *implicit*, *automatic*, and *intuitive* seem to hang together as an important cluster for dual-process theory. ”Intuitive” is also a vague notion which pretty much means ”implicit

and automatic". Hence, I figured that "intuitive" and "reflective" may perhaps serve more informatively as tags for the putative two processing types, rather than as their characteristic features, and focused on this particular distinction.

1.2 Putting things together the wrong way

There is also a branch of consciousness research that is interested in finding structural analogs between cognitive processing and descriptions of conscious experience. For example (Petiot et al., 1999) contains several texts that aim to show how aspects of cognitive processing can be mapped onto Husserlian phenomenology. I made a brief detour to acquaint myself with the works of Husserl, mainly via second-hand sources in analytic phenomenology tradition. So I read cursorily Husserlial works such as Dagfinn Føllesdal (1966), and especially David Woodruff Smith and Ronald McIntyre (1982). I got fascinated by the way the two latter authors had described Husserl's account of object perception as an act that constitutes the content (or *noematic Sinn*) to what is perceived.

This content constitution is not merely an act of classifying objects but an active process that involves the invocation of an intricate pattern of meanings that are more or less implicit in the act of perception. Every perception has content, which is constituted against a "horizon," a background of further possible experiences and properties prescribed by the singular experience. We see an object as a whole; as having a backside, even though what we sense is only its visible side. In this sense, the content of perception is *transcendent*, for the act of perception transcends or goes beyond the information given. In cognitive terms, the experience is a system of tacit expectations, delivered by our background knowledge.

These expectations are not subjectively obvious because, somewhat paradoxically, they are precisely something we take for granted. They are *too* obvious to be noticed; they are at work when we go around an object to casually examine its backside. We do not explicitly infer that there *is* a backside but simply take it as given, and that implicit givenness is an inherent constituent part of the object perception and object oriented action. These expectations do not give all the specific details, of course; for example, what is the history of the object and what actually is there on the backside? This indeterminacy is the very reason for further examinations. Although the horizon of experiences is partly indeterminate, it is still structured. It predelineates the range of sensible blanks and fillers, or questions and answers, that the perceived ob-

ject affords while retaining its identity. These tacit expectations can become manifest, however, when something violates them and the perceived nature of the object suddenly changes—e.g., when on closer inspection a tree proves to be a stage prop. Husserl says that in these moments our experience "explodes" (Føllesdal, 1966). We sense a fragmentation of the horizon of tacit expectations, which then become visible. Ordinarily, however, what we are aware of are just passing specific stances or perspectives towards things and events that constitute the totality of our lifeworld.[6]

Husserl's account of intentional content provides useful insight into the constitution of conceptual *understanding*—something that seems to be lacking in the analytically oriented philosophy of mind. The description of how we bring our background knowledge to an experience hints towards the subjective genesis of meaning which is necessary for intuitive grasping of our surroundings and which comes about via our continuous interaction with the world. Without any attempts to naturalize phenomenology or follow scholarly interpretations of Husserl, I gathered how this analysis could be reconciled with dual-systems theories. The idea is that intuitive System 1 automatically filters and processes information and makes situationally relevant information available for reflective processes. Generally, events activate memories encoded in previous comparable situations. Much of that information remains implicit but salient to conscious reflection. This gives a sort of fullness to our conscious experience. The sense of understanding comes about by this pre-reflective cognitive processing, which brings all kinds of presuppositions and anticipations into lived situations. Explicit manipulation of meaningless symbols would lack this sort of phenomenological richness, simply because there is no automatized filling of the blanks going on behind the scenes. Hence, someone without any meaningful background knowledge about dealing with this sort of symbols would be situated like Searle in his famous Chinese room thought experiment (Searle, 1980). There are symbols, rules, and calculations but no meanings anywhere, because meanings, phenomenologically speaking, are the systems of intuitive expectations, shaped by our previous encounters with the world and are pre-reflectively brought into situated action—mental acts included.

The description of how experience changes when expectations are violated reminds the *default-interventionist* account of dual-processing. In this model, System 1 automatically guides our action as long as everything proceeds smoothly. However, when our tacit expectations are violated, the system fails to deliver appropriate responses. Then the intuitive flow is interrupted,

[6] See Smith & McIntyre (1982), or McIntyre & Smith (1989) for a concise discussion.

and the reflective cognition kicks in to attain a grasp of the changed situation. We engage in conscious deliberation and action control until the issue is resolved, the new situation is comprehended, and the sense of normal flow of events is restored.[7]

After these contemplations, my research forked into two branches. As a sort of spin-off project, I investigated the possibility that the dual-system account could be turned into a cognitive theory of conceptual understanding, which would describe how preconscious processes shape our explicit comprehension. If carried out successfully, the project should be interesting because one fundamental problem in cognitive science has been the modeling of common sense, which seems to have little to do with explicit calculations but immensely with intuitive understanding and pre-reflective assessment of relevance. Not for long, this side project took the lead and turned into the present work. Along the way, the central role of dual-process theories faded into the background but many of the theoretical planks that aroused the philosophical excitement addressed above—along with the actual empirical results—can be found in some form herein.

I thought that it would be a good idea to think of Systems 1 and 2 approximately in terms of connectionist and classical computational systems. In the above list (and especially in the shorter one on page 27), the majority of the properties of intuitive and reflective processing align with the general characteristics of connectionist and classical systems, respectively. From the empirical perspective, I thought that assimilating similarity-based category representations with System 1 contents and theory-based conceptual structure with System 2 representations is perhaps workable hypothesis because they also have properties that to go well with this kind of an arrangement. These theories are elaborated in Chapter 3.

In the philosophy of mind and cognitive science, we encounter yet another vaguely similar dualism. So called *two-factor conceptual role semantics* (see the next chapter) aims to explain the origins of mental content by two independent factors: (a) a causal connection between mental representations and their putative referents, and (b) inferential dispositions of the cognitive agent that track conceptually relevant relations between mental representations. The connection to dual-process theories is remote, but the causal factor ostensibly

[7] The alternative is the *parallel-competitive model* (Sloman, 1996; Smith & DeCoster, 2000). It assumes that both systems are active and compete for control in any given task. This model has worse empirical and theoretical standing than the interventionist model and hence will be ignored in what follows; see (Evans & Stanovich, 2013).

handles content determination for concrete referents and the inferential factor for abstract and theoretical content; hence the link to similarity and theory-based accounts of concepts in cognitive psychology. In the introduction of *The Origins of Concepts* (2009, 5) developmental psychologist Susan Carey endorses "in broad strokes" some form of two-factor conceptual role semantics or related dual theory of concepts.[8] A somewhat similar dual account of concept representation is also proposed, e.g., in Armstrong et al. (1983); Allen & Brooks (1991); Smith & Sloman (1994), and importantly in Sloman (1996), a seminal paper in current dual-systems theory, which also proposes close links to connectionism and rule-based computationalism.

The two-factor conceptual role semantics is unfit for the reasons detailed later but, briefly, the problems concern its background assumption of a universal, privileged system of concepts which is independent of the concept using organism. The theory (like the related proposals derived from the analytical philosophy of language) dismisses the needs, goals, and capacities of the organism and pretty much the world around it—or, at least, considers these of only secondary importance. Whether this is a good metaphysics of concepts or not, it makes a bad philosophy of mind. Moreover this sort of conception of concepts is mostly incompatible with both empirical results and the genetic reading of Husserlian phenomenology. I do find it important that a full theory of human intentionality requires both: an account of content that is engendered by causal interaction with the environment, and an account of the role of public and private reasoning in content determination. However, for those ends I promote enactivism and linguistic pragmatism instead of causal and conceptual role semantics.

I use the term "enactivism" in a broad sense that accommodates situated and embodied approaches to cognition. According to Evan Thompson the central idea of the embodied approach is that "Cognition is the exercise of skillful know-how in situated and embodied action" (Thompson, 2007, 11). Cognition, intentionality, and meaning arise from continuous interaction between the world and the agent. The approach is independent of but highly compatible with both connectionism and similarity-based theories of conceptual represen-

[8] Although overt references to that book are sparse here, it should be mentioned that *The Origins of Concepts* profoundly influenced my research. There are similarities at least with to the proposed mechanisms of concept learning and chance, employment of some form of dual theory of conceptual content, and the existence of diverse sources of content determination: at least individual learning, cultural process, and evolution.

tation, especially the prototype theory.[9] It is influenced by phenomenology (e.g., by Merleau-Ponty and perhaps Heidegger more than Husserl, although Husserl influenced both) and bears clear similarities to pragmatist tradition. According to Robert Brandom "a founding idea of pragmatism is that the most fundamental kind of intentionality (in the sense of directedness towards objects) is the *practical* involvement with objects exhibited by a sentient creature dealing skillfully with its world" (Brandom, 2008, 178)—which is pretty much the founding idea of enactivism, too.

Instead of symbolic representations and reasoning, enactivism emphasizes the perception–action-loop and the operations of our sensorimotor system as the fundamental building blocks of cognition. Cognition is embodied in the way specific to the organism and developed to serve its needs and capacities. This view holds that cognition is not static system but it is shaped by the specific experiences that accumulate in the interaction between the organism and its environment. The strength of this approach is that it allows factoring in the situatedness, contextuality, specific contents, and individual learning in content determination without lapsing into solipsism: Similar constitutions of body, cognitive make up, and the environment forge similar conceptual systems. The weakness lies in its playing down of the significance of symbolic thought. The more the contribution of the dynamically coupled sensorimotor system is stressed, the harder it gets to explain how abstract reasoning is even possible.

One way out of the impasse is to embrace both enactive and symbolic cognition in their own right and let research sort them out in the long run. There is no *a priori* reason why there needs to be continuity from sensorimotor to symbolic processes and why cognition needs to be modeled based on a single principle. Although the issue with embodied/enactive versus symbolic thought is somewhat orthogonal to the debate between connectionism and classical computationalism, these paradigms are facing similar problems: Connectionism is especially suited to modeling low-level perceptual and motor processes, while computationalism is developed to deal with planning, symbolic reasoning, etc.

Taking all the above points into consideration, I opted for the following, albeit short-lived, working hypothesis and method: To resolve the debate between enactive/embodied research and the classical view of reasoning and concepts, put all embodied, fuzzy, online, pragmatic cognition in System 1; and

[9] E.g. (Varela et al., 1991), a seminal work on enactivism, explicitly sides with connectionism and includes Eleanor Rosch as one of the authors who was also a pioneer of prototype theory of concepts, discussed in detail later.

abstract, clean, theoretical thinking in System 2. Assess what empirical findings (especially in concept research) naturally goes with what system, without strong pretheoretical commitments about the underlying cognitive architecture. Evaluate the characteristics of cognitive processing and representations in the resultant hypothesis about the putative systems.

So basically the idea was that faculties of abstract and theoretical thinking, such as those employed in doing science, go to System 2 and intuitive common sense reasoning and the like to System 1. In light of the reasoning research that fueled the two-systems theory, the above hypothesis is quite reasonable. A key part of the methodology is to find tasks that prompt subjects to make conflicting responses (Sloman, 1996). This research shows that often the intuitive responses violate normative models, while subjects are capable of reaching the correct solution when they think the problem through (Evans et al., 1993; Evans & Over, 1996; Gilovich et al., 2002; Johnson-Laird, 2008). Importantly, the intuitive responses are not random but predictable across subjects, leading to specific stable *biases* in reasoning. These biases are often interpreted as to show that intuitive system employs quick and dirty heuristics that work efficiently in everyday pragmatic reasoning but are deficient in formal, abstract thinking.

1.3 From errors to expertise

But that didn't go very well. Basically, the problem was that theoretical and other declarative knowledge is not insulated from practical experience with concrete things. People can have all kinds lay theories derived from practical experience, and theoretical knowledge influences how things are perceived and conceived: If you think that vitamin C cures flu, a simple experiment, using yourself as a subject, will most likely confirm this. However, it does not matter what vitamin you take because the odds are that you will get well anyway. The background assumption about the vitamin may lead you to anticipate the expected outcome, to attribute it to the purported remedy, and hence to encode the spurious causal relation as genuine. In general it is complicated to tell if knowledge is *de dicto* or *de re*, or theoretical or practical. If you avoid spoiled food because you know it is a health hazard, which type of knowledge is this—especially if you have never experienced food poisoning? Is learning chess declarative knowledge about rules and strategies or situated knowledge about how to manipulate the pieces and concrete board positions to reach your goals? Is a mathematician's deriving proofs on paper not a sort of concrete

practice? At least the cognitive skills involved are most likely developed during laborious thinking, experimenting, trial, and error, that is derived from a long track of experiences gained in actual material events.

The dual-process social psychological research is also hard to interpret by conflating explicit processing with theoretical or abstract knowledge. Often the subjects make certain decisions and then give conflicting commonsense justifications for them, not displaying conflict between theoretical and practical inference but between implicit decision-making and its explicit *post hoc* rationalization, framed with solid, albeit irrelevant, practical reasoning (Wilson, 2002). Similar rationalization, or confabulation of reason, is seen in reasoning studies discussed in Chapter 5.

What is common in these social psychological studies and reasoning research, aimed at uncovering intuitive biases, is that the subjects often engage themselves in unfamiliar tasks. However, another strand of implicit reasoning and judgment research exists that is somewhat overlooked in dual-process theories in favor of the heuristics and biases tradition, namely expertise research that emphasizes the *proficiency* of intuitive decision making.[10] In this tradition, human expertise is taken to be a product of long and laborious top-down learning. First, we follow explicit rules to acquaint ourselves with a new problem domain. Gradually, we accumulate a large stock of memories about specific situations, actions, and outcomes. If randomness is involved, the recurring situations equip us with a good grasp of the range and proportional frequencies of expected events and ways to deal with unintended outcomes. The situation representations are assumed to be quite concrete, in the sense that they do not contain summary structural information about the problem domain but rather specific information about actual events, associated actions and their outcomes. A stock of these memories is implicitly activated when an expert encounters a familiar situation.

This sort of *recognition-primed decision-making* (Klein, 1998) is nicely captured by Herbert Simon: "The situation has provided a cue; This cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition" (Simon, 1992, 155). Since all this happens effortlessly without conscious intervention, we see a dissociation of implicit decision-making and explicit reporting, similar to that of untutored biased reasoning. Experts tend to make competent instead of biased decisions; however, especially in the case of complex tasks,

[10] See de Groot (1965); Dreyfus & Dreyfus (1986); Simon (1992); Klein (1998); for a concise dialogue between these two intuitive decision making traditions see Kahneman & Klein (2009).

they are often quite unable to explain their choices much better than advanced amateurs. Again, in Dreyfus & Dreyfus' (1986) words, "*When things are proceeding normally, experts don't solve problems and don't make decisions; they do what normally works*" (30–31) ... "They cannot always provide convincing rational explanations of their intuition, but very frequently they turn out to be correct" (34).

This sort of intuitive know-how is very similar to the assumed System 1 processing, and it also bears at least superficial similarity to Husserlian theory of the construction of perceptual content. To obtain pretty much the default-interventionist theory of dual-systems reasoning, we only need to add that in case intuitive comprehension fails to deliver useful information in accordance with our goals, the behavior is interrupted by controlled processing. When this happens there is no necessarily a "lift" to more abstract or normative thinking, but often the subjects lapse to the state of an amateur and are forced to use whatever cognitive tools and skills they have to think the problem through (Dreyfus & Dreyfus, 1986). It is generally incorrect to assume that, e.g., the comprehension of a mathematical problem happens by translating it into a content independent formal representation for explicit mental faculties to compute. Intuitive comprehensions seems to work in a way that the specific problem contents foremost activate associated *procedural* information.

The following example are from (Ross, 1984) where subjects were tasked to solve several probability problems:

1. There are six judgeships on the local ballot to be voted for. Each office has eight candidates running for it, one of whom is the incumbent. If a person randomly chose for each office (with a 1/8 chance of choosing any of the candidates), what is the probability that she or he would vote for one or more incumbents?

2. Two 8-sided dice, a green one and a red one, are rolled. Each dice has a 1/8 chance coming up 1, 2, 3, 4, 5, 6, 7, or 8. What is the probability of getting either a 7 on the green die or a (2 or less) on the red die, or both?

The subjects were then taught to solve the first problem by using the formula $1 - [(c-1)/c)]^t$, where c = number of choices (8), and t = number of tries (6); they were instructed to solve the second problem by adding the probabilities of the two events and then subtracting the probability of their co-occurrence: $P(x) + P(y) - P(x) \times P(y)$; i.e. $(1/8 + 1/4) - (1/8) \times (1/4)$

in this case. Then they were given, e.g., the following problem, which is a restatement of Problem 1 by borrowing the thematic content from Problem 2:

3. Six 8-sided dice are rolled. Each dice has a 1/8 chance coming up 1, 2, 3, 4, 5, 6, 7, or 8. What is the probability of getting a 8 at least on at least one of the dies?

For Problem 3, most subjects incorrectly try to apply the solution for Problem 2.[11] What they have learned previously is a procedure to compute probabilities associated with dice rolls rather than a general method of computing probabilities of n-independent events in t tries—or even a capacity to recognize the structure of the problem accordingly. More precisely: Of course, we do learn such capacity because we can appreciate that some procedures lead to incorrect and some to correct solutions, and that the reason is the underlying problem structure. However, I maintain that the structural understanding comes principally from associating a correct procedural knowledge to the task and not from initially forming a formal and decontextualized problem representation (see Chapter 5 for details). Finding the correct procedure can be guided by explicit search, especially for relevant analogies.

Task specific skills and domain understanding build up from these automatically activated, content-sensitive procedural memories. When the surface features that guide the memory search are misleading, this can lead to predictable errors similar to the ones the above. However, if the cues reliably activate useful and valid expectations, the result is highly efficient, domain specific cognitive competence, which does not tax working memory or other capacity limited cognitive resources. When expertise is cultivated, the focus gradually shifts from superficial features to more important structural properties of the domain. Does the latter mean that more abstract representation of the domain is generated or that the understanding of structural properties is simply constituted by associating generic procedural knowledge to a wider variety of superficial situation representations or surface features? Quite likely both are true; yet, the idea that I'm selling is that the phenomenological click of comprehension is produced by tacit processes that recruit practical knowledge, and understanding abstractions is similarly rooted in specific representations

[11] Note that these problems are actually identical in the sense that the solution procedure for 1 works for all these tasks. However, the solution for task 2 gives wrong probabilities if there are more than two variables; in other words, $P(x_1) + \ldots + P(x_n) - P(x_1) \times \ldots \times P(x_n)$ is an incorrect solution when $n > 2$.

of concrete instances.[12] This renders examples, metaphors, and analogies so useful. Moreover that is why it is not possible just to read a book of calculus and immediately understand its contents. Instead, you need to labor through examples and exercises to obtain know-how about the domain. This does not mean that concepts, in any philosophically interesting way, can be reduced to sensorimotor contents or to human psychology in general; only that the human conceptual cognition does not track (at least directly) abstract intellectual or metaphysical essences but affordances and other functional properties. Fundamentally, conceptual understanding is constituted by tacit know-how rather than explicit know-that.

Given these remarks, we now can review my initial (and now abandoned) working hypothesis that all embodied, fuzzy, on-line, pragmatic cognition goes into System 1 and abstract, clean, theoretical thinking to System 2; and the associated characterizations of intuitive and reflective cognition.

To recap, we know that expert competence rests mostly on a tacit cognitive skill-set to negotiate a specific domain. This know-how builds gradually through numerous encounters with specific problems.[13] Expert reasoning displays the effects of context, content, and specific learning history similar to pragmatic reasoning. Often expertise is accompanied by rather poor transfer of learning, meaning that acquired cognitive skills do not reliably cultivate general reasoning capacities; this even holds true for scientific competence. Basically, all that resemble standard idea of System 1 processing. If we accept that intuitive cognition is inherently deficient in abstract reasoning, as heuristics and biases traditions often suggest, we should expect scientific and other theoretical intuitions to be reliably biased. Unfortunately, scientific cognition has not been studied extensively; however, we know that training in statistics or formal logic, for example, does not eliminate biases completely and that experts from scientists to judges are susceptible to common reasoning errors (Griggs & Ransdell, 1986; Peer & Gamliel, 2013). Although logical and statistical reasoning errors are less frequent within subjects with mathematical training, the observed deviation of performance from normative models is

[12] Note that "concreteness" here should be understood as *practical familiarity*, and not that you always need to explain abstract concepts and principles by referring to concrete things—although analogies and metaphors are often used this way. Examples usually serve to introduce practical familiarity by showing how to apply a schematic mathematical principle to specific numbers of symbols.

[13] According to oft-cited estimates, acquiring excellence in complex tasks takes about 10,000 hours of practice extended over more than a decade involving a huge amount of repetition (Ericsson et al., 1993).

often still substantial (Tversky & Kahneman, 1971, 1983), and an education level as such is generally a poor predictor of normative performance (Jackson & Griggs, 1988). Nevertheless, the whole notion of "expert" implies that the person outperforms the average.[14]

Now, scientific education may equip us with at least three things: 1. Formal methods, such as mathematics, that can be applied to several specific domains, 2. metacognitive attitude to suppress reliance on intuitive judgment, and 3. social norms and practices to engage in critical thinking. These items likely result in more efficient cognitive tools and increased reliance on reflective reasoning. Hence, one might suspect that it is this reflective component that makes people good scientists. Undoubtedly, that plays a crucial role, but scientific conceptual domains are complex and sometimes quite vague. This is especially the case in humanities and social sciences; which does not mean that these disciplines are less scientific but that they deal with hard-to-define complex phenomena. All this sophisticated knowledge needs to be applied to actual use, and what we know about human expertise and reasoning in general suggest that theoretical competence requires considerable implicit learning and processing. Moreover, the norms and practices associated with critical thinking and academic discourse are similarly subtle skills that cannot be wholly captured by rules of thumb; instead, they need to be exercised to be learned.

The last point applies to other "intellectual" domains, such as the extensively studied chess, and hence the point is more general. You cannot begin to play chess from the scratch without explicit rules and metacognitive control; yet, to play *good* chess, effective deployment of the intuitive system for the task is needed. Highly competent players do not explain their moves by appealing to deductive arguments or computations of the most effective actions; they, instead, refer to the strategic situation on the board. They certainly do reflect on their play, but they seem to instantly disregard irrelevant and bad moves and consider viable ones only. (de Groot, 1965, 305–307) Indeed, they

[14] There is, however, a caveat in this seemingly obvious point. Experts are susceptible to believing in spurious causal relationships like anyone else. Sometimes the whole problem environment is essentially random, which makes it impossible to cultivate any reasonable skill in predicting long term outcomes. That is why we have professionals from witch doctors to fund managers whose status as experts is largely socially instituted, while their usefulness is largely at suspect (see e.g. Kahneman, 2011, part III). Note that randomness *per se* does not forbid understanding the domain. In fact any competent fund manager with training in economics should know why efficient market is impossible to beat consistently. It's just that these experts are often recruited for assignments to which their expertise is not actually very well suited, like making reliable forecasts of investments or of political developments.

seem to be completely unaware of the detailed rationale of their actions while still maintaining proficient performance; this is common in human expertise in general, and with complex tasks such as chess, it is easy to see why it is necessary. The complexity of the mandatory computation of the best move from all possible options quickly exhausts the performance capacity of the deliberative mind. The fact that it is *possible* to employ intuitive cognition to overcome these limits is more difficult to explain. The candidate explanations seem to be either that (a) the intuitive system carries out the same rule-based computations than the metacognitive system but only subconsciously and more efficiently, or (b) that the system exploits different kinds of information and processing. For example, intuitive cognition might perform some kind of global pattern matching to recognize strategically significant situations on the board and exploit complex regularities that associate specific situations to an efficient line of actions.[15] The latter option is discussed in detail in Section 4.3.

In any case, the psychological research uncovers substantial competence/performance mismatch: Even if you have passed a course on logic, you are still as susceptible to elementary reasoning errors as anyone else (Hoch & Tschirgi, 1985). Having specific (explicitly acquired) conceptual tools does not instantly give you skill in putting them to good use. Even if you have learned abstract reasoning schemata, they are automatically employed only in specific contexts. For example, you need to recognize a logic problem *as* a specific logic problem to understand what inference principles to use and how to use then, which requires practice. Even choosing the right tools for the task does not mean that you can excel with them. Think of proving a very complex theorem, for instance. Abstract, like other forms of reasoning, requires cultivated intuitive skills that can be only achieved by substantial practical experience.

Another point to note is that commonsense reasoning is sometimes rather abstract. For example, the folk is prone to reasoning errors, such as affirming the precedent or failing to commit *modus tollens* with logical conditional *if A, then P*, but display far fewer deficits with schematic deontic conditionals, e.g. *if the action A is to be taken, then the precondition P must first be satisfied* (Cheng & Holyoak, 1985). Note that the abstract placeholders A and P need not be specified. As long as subjects understand that they are dealing with inferences concerning conditional permissions, they often display quite fluent logical competence.

[15] Note it is possible to learn complex hierarchical plans by reinforcement learning used in AI systems. In other words, associations can be learned to map situations to *policies* of actions in addition to mapping them only to singular actions. (Russell & Norvig, 2010, 856)

Thus, my point does not specifically concern scientific cognition but that the assumed System 1/2 processing differences reflect more reliably the difference between familiar and non-familiar rather than concrete versus abstract content. In fact, I think that people casually use the term "abstract" to refer to notions that are both sufficiently non-concrete and *non-familiar*. For example, the notion of "permission" does not appear very abstract because we are all familiar with how permitted actions and their preconditions causally play out in various specific contexts. Elementary particles, I presume, are fairly concrete; yet, their nature might be rather abstract if not conceived as a kind of tiny billiard balls colliding with each other or in other familiar terms. There is also a social dimension to this: My bet is that many scientists agree that they operate with abstract concepts even if the concepts are completely familiar to them; however, this perhaps reflects the fact that the notions are foreign to most people, and hence it is appropriate to consider them as abstract in the common sense of the term. The point is that if we are to find the proper place for abstract and concrete concepts in cognitive theory, the issue is orthogonal to the ontological status of the referents. It is more useful to pay attention to the *actual practices* in which the concepts participate, and it is important to bear in mind that these practices are not always universally shared. Hence, ontological "abstractedness" is perhaps psychologically and phenomenologically irrelevant. What is relevant is that many abstract concepts are foreign to the untutored, and we initially need explicit definitions and rules to work with them. Hence the close pre-theoretical association of abstract concepts with explicit reflective thought.

Moreover, the need for controlled explicit thought is hardy linked to the complexity of conceptual domains. Commonsense reasoning is, in fact, astonishingly complex. One of the significant challenges in cognitive science has been the modeling of common sense. Our behavioral repertoire is very flexible but still systematic and reasonable in respect of our needs, goals, and situational demands. Even when facing uncertain and incomplete knowledge, narrow time frames, and abruptly changing contexts, we are able to parse meaningful wholes from logically unconnected information and fill in all kinds of missing blanks. All this is usually accomplished swiftly and seemingly without effort. For example, if I leave a note for a friend saying, "I'm making some bigos tonight. Could you check what's in the fridge and drop by the store? Some cash and the car keys on the kitchen table", the recipient almost certainly understands how these logically unconnected statements make a meaningful message. In case you wonder what ingredients you may need for bigos, you can congratulate

yourself on conducting a perfectly valid practical inference. Practical reasoning is so pervasive and effective in everyday life that it disappears. However, the deceptive easiness of it hides the convoluted system of tacit presuppositions and expectations that make it work, and the resultant complexity sabotages any attempt to model it by explicit and formal propositional reasoning systems, such as classical logic.[16]

For the people committed to the classical computational theory of cognition, this might suggest that the human mind is an exceptionally powerful reasoning engine. Hence, it was surprising for many that the psychological research during the past half-century or so revealed that our cognition is rife with biases. Our reasoning and decision-making markedly diverge from formal normative models. We tend to estimate probabilities wildly incorrectly and neglect the base rate, even when all relevant information is available. We evaluate credible correlations and outcomes as more probable than they are, to the point of violating the fundamentals of probability calculus. We tend to violate the principles of utility theory and make decisions often inconsistently. We do not respect the Bayesian rules of belief update and tend to disregard the logical form of arguments and let the content affect our reasoning in logically irrelevant ways. (Evans & Over, 1996; Gilovich et al., 2002) This discrepancy illustrates the fact that the intuitive mind does not deploy formal rules for everyday reasoning, and hence the question is what does it do then? The answer is outlined above. We simply take the notion of skills as the fundamental building block of not just expert competence but of conceptual cognition in general.

1.4 The working hypothesis

In this work, the following hypothesis is defended:

A.1. Expert and commonsense reasoning are both grounded in the same cognitive processes.

A.2. Conceptual understanding builds up as an adaptive cognitive skill. Its cognitive basis is in the intuitive system that gradually learns to exploit context- and goal-relevant regularities in the environment, especially the effects of our own actions in specific situations.

[16] See Davis & Marcus (2015) for a brief review of state-of-the-art of commonsense reasoning in AI.

A.3. Often relevant environments are, at least partly, socially constructed. In the case of theoretical concepts, in particular, the relevant regularities are in large part inferential and other discursive commitments. Hence, abstract concept learning is a special case of functional/causal learning in (broadly) social or cultural contexts.

A.4. Initial competencies depend on concrete examples and their surface features or specific content. Extensive learning of procedural knowledge results in a gradual shift of focus from surface cues to structural features of the conceptual domain.

A.5. Through experience, the abstract/practical distinction dissipates both psychologically and phenomenologically.

A.6. Still, even later competence is affected by the specific learning history and content. Hence, the result is not an acquisition of general formal reasoning capacity but a gradual transformation in domain understanding.

A.7. Therefore, the process produces domain-specific conceptual competencies by a general adaptive and praxis-oriented learning mechanism. The resultant domains are products of our needs, goals, capabilities, learning environments, culture(s), etc.

The fundamental idea is that interacting with the environment forges our basic conceptual ontology, and the pragmatic knowledge that accumulates through experience provides us with the know-how that shapes our understanding of how things generally hang together. The latter is the actual conceptual *content*. If all goes well, biases level out when the implicit faculties adapt to the actual statistical and causal structure of the domain of interest. When the tacit capacities become reliable basis for intentional action, the domain gradually becomes semantically transparent. The associated intentional contents are constituted by automatically activated material knowledge that tacitly fills all sorts of blanks and produces anticipations of how things unfold without conscious reflection. This is what I mean by conceptual *understanding*; Husserlians probably call it something like ”pregiven construction of meaning”.

But what sort of a thing the intuitive cognition is? As a starting point I adopt the model from generic dual-process theories, but the list of properties requires a revision. We need to drop references to biases and other subpersonal and non-normative attributes. While research on biases is methodologically important for the study of intuition, normative and conceptually structured

processing cannot be exclusively attributed to explicit cognition. Moreover, explicit reasoning can also be biased if there is something wrong with the underlying intuitions that support reflective thought.

Common sense is not a universal system but a local set of dispositions forged by the surrounding culture(s) and is, hence, variable across individuals, times, and places (Bruner, 1990). Therefore, there must be individual variability, at least, in the content if not in the process of intuitive reasoning. The pragmatic/abstract division we already abandoned, and item A.3. above imply that language is not a sound demarcation criterion of intuitive/reflective processes. In all, the idea that the intuitive cognition is a universal engine of quick and dirty heuristics while the explicit system is a cultured, normative, and linguistic symbolic processor needs to be abandoned.

For now, I skip the discussion about the processing properties of the two systems appearing in the list on page 9. Properties associated to intuitive reasoning, in particular, are briefly discussed below and extensively in the subsequent chapters. Basically, the array of features boils down to the following:

B.1. *General characteristics of intuitive and reflective cognition*

Intuitive system	**Reflective system**
independent of working memory	depends on executive function
high capacity/low effort, fast	low capacity/high effort, slow
associative	systematic
rigid	flexible
parallel	serial
implicit	explicit
automatic	controlled
situated	detached
default process	inhibitory

In essence, reflective processing constitutes an overriding system that is characterized by fluent control with low-capacity working memory. Intuition is based on fast, context cued parallel memory search and a slowly adapting knowledge base. The basic idea parallels the recent trend in artificial intelligence that trades sophisticated reasoning systems for rich data sets and dumb but powerful learning algorithms that extract goal-relevant regularities inherent in the data.

I opened this chapter by reciting the Gideon Keren's ((Keren, 2013)) concerns about the long term prospects of dual-process theories. In the same volume, Evans and Stanovich (2013) reply to critics with an overview of the past research, reaching the same characterization of reflective processing as the above. They propose, however, that non-conscious system is a collection of autonomous processes of at least three distinct kinds: (a) encapsulated input modules of the Fodorian type (Fodor, 1983), (b) associative learning system for extracting world knowledge, and (c) habituated processes that once were explicitly controlled.

My agenda is to find the cognitive basis of conceptual understanding and sense-making. Therefore, I no longer aim for a firm stand on how dual-process theorists should formulate their theoretical apparatus. My focus is on domain general learning mechanisms that deliver domain specific conceptual competencies. This framing rules out any low-level hardwired input analyzers, instincts or other "Darwinian modules" (Cosmides, 1989), and the like from the following considerations regardless of their status in the dual-process theory. Although, if there are any such things they do shape our behavior and therefore our conceptual system. At least I do believe that there are innate, specific learning mechanisms that support capacities engendered by general learning. For example, I will later explicitly rely on the assumption that we are predisposed to learn at least about causes, agency, and language in a manner that goes beyond mere association (see Carey, 2009). However, if we take the stance that skillful action is the basis of conceptual cognition and hence world knowledge, I fail to see how the above planks (b) and (c) constitute separate processes. I later show how they are integrated aspects of concept using and concept producing basic mechanisms.

What comes to the psychology of concepts, the pragmatic/enactive stance also means not to focus solely on category representation. I advertise a kind of inferentialism that is somewhat distinct from how philosophers usually construe the notion. I claim that content is derived from our capacity to *causal* induction and reasoning. Concepts are induction machines—aggregates of causal and functional knowledge about things and events—rather than symbolic nodes in a decontextualized analytic network.

From the empirical standpoint the above is hardly more than a statement of a problem. Hence a tentative model of intuitive processing is proposed along the following lines:

B.2. *Specific processing and representational assumptions about implicit cognition:*

a) Inductive (Bayesian) search for causally predictive and pragmatically relevant regularities in concrete situations:
   - Category representations are (prototype) feature clusters structured by causal relations between constituent features.
   - Basic ontology is formed around affordances and features relevant to event–event causation.
   - Basic situation representations are quasi-perceptual model type structures.

b) Instance-specific encoding of event/action/outcome exemplars:
   - Context- and goal-dependent causal expectations are generated by associative (pattern matching) memory retrieval or analogical mapping from exemplar-based situation representations.
   - Without valid expectations, control shifts to exploration or explicit reasoning.

c) Basic reasoning mechanisms are forward causal inference, simulation of situated action, and abstraction by exemplar-based analogical transfer:
   - Surface features are the primary retrieval cues.
   - Valid analogies bind different tasks under shared pragmatic schemata.

Thus, the basic level of representation and processing is considered to be highly concrete. The problem, then, shared with other enactive or pragmatistically oriented theories, is to explain abstract thinking. The key is the analogical transfer of causal and procedural knowledge. In Sections 3 to 5, I explain how the capacity for *taxonomic* and *schematic* abstraction is based on bottom-up learning. It does not necessitate constructing more abstract representations but focusing on commonalities between specific instances. In case of schematic contents, in particular, the commonalities are found in shared pragmatically relevant causal structures, under which different situation variables are interchangeable in relation to the agent's goals. The resulting schemata are projectible to novel situations through similarity or analogy. To understand especially schematic concepts (such as "permission" or "medication"), it is essential for the agent to understand relevant *event–event* causation which makes *situation* a core notion in concept representation.

Theoretical abstract content is construed more along the lines of classical inferentialism as the use of linguistic and other symbolic tokens in reasoning. But again, the hypothesis is that the underlying cognitive capacity tracks event–event causation in broadly understood discursive practices. Perhaps the expression "language games" best conveys the basic idea of these practices.

To start playing, one often needs explicit instructions about principles, rules, and conventions of the game. Therefore, reflective controlled processing is needed. No presumption is made that control has much to do with inherently logical thinking but simply with suppressing inappropriate behavior, following arbitrary instructions, and decontextualized thinking which is insulated from existing beliefs. The important capacity of reflective cognition is its possibility to implement an arbitrary algorithm in principle. You take a system of rules and representational tokens and get going. By using a pen, paper, and other external props, the system's inherent limitations can be relaxed—especially what comes to the working memory. By learning to associate an appropriate action with a given task by repetition, you can save computation time. Hence, the idea is that through experience recurring discursive, theoretical, and other arbitrary "algorithmic" activity can be partly internalized by intuitive capacities through top-down learning.

This, of course, is a sort of dual-process theory of cognition. Still, it is remote from any two *minds* hypothesis, since the two processing modes are highly intertwined, and they often carry out different *aspects* of the same tasks. Another link to the generic dual-process account is that the intuitive mind is conceived as very concrete- and practice-oriented. Its mode of operation is largely to serve as a filter that makes only contextually relevant information salient for the task at hand and hence limits the scope of all the possible knowledge and actions that the agent may take into consideration. There also lies its weakness because for the tasks where one should be open to taking a new look on things—such as in politics or science—this may lead to serious tunnel vision especially with highly practiced experts with heavily entrenched intuitions. Lastly, I do not wish to claim that the implicit heuristics recognized in the literature are nonexistent, but only that they often are results of interpretative capacities of the intuitive cognition rather than rules of thumb for shortcutting laborious inferences. That is, what are generally considered as reasoning biases are sometimes actually semantic problems in task understanding, and subjects do not necessarily *reason* any worse when they display biased responses but rather commit inappropriate goal setting and information selection.

In closing, we can restate the hypothesis A.1.—A.7. in epistemological terms:

C.1. The human conceptual system tracks affordances and other causal properties that have pragmatic relevance to us as embodied and active organisms.

C.2. Intentional content is fundamentally procedural competence to exploit the resulting know-how of what things do and what can be done with them.

C.3. Expertise in abstract expert domains (e.g. science) can be thought of as extended common sense, and common sense can be thought of as a specialized learned skill.

C.4. Cognitive contents of theoretical and formal concepts are constructed by the agent through social interaction. While grounded in the same capacities, discursive and concrete concept learning often differ qualitatively: Discursive conceptual domains track explicit and implicit communal conventions; this makes content intersubjective and normative.

C.5. Two people share the same intuitive understanding as far as they share the same practices, goals, needs, capacities, discursive commitments, environmental demands, affordances, and culture, or, in brief, the same practical reality.

2 What is intentional content?

The concept of *concept* has multiple meanings in both philosophy and psychology, but I agree with Edouard Machery's (2009) summary that, roughly, philosophers think of concepts as constituents of propositions whereas psychologists treat them as mental representations or other capacities that enable our higher cognitive competences, such as categorizing, learning, inference, decision-making, and so on. I don't know what propositions are, but usually they are considered as something that represent or stand in for states of affairs, something that can be either true or false and expressed by declarative sentences. If propositions are abstract entities serving as meanings of sentences, then concepts are abstract entities that correspond roughly to word meanings.

Thinking of concepts as constituents of sentential compound expressions implies a crucial distinction between *logical* and *non-logical* concepts. In the analytic tradition of philosophy of language, the latter is usually thought of as consisting of object and category concepts. Objects are individuals that can be tagged with a label, such as persons, mountains, Eiffel tower, your dining room table, and perhaps individual numbers and other abstract entities. In what follows, what I mean by an *object (or concrete) concept* is a category of things that can be expressed with nouns. Terms referring to this kind of categories all belong to non-logical concepts as do concepts corresponding to properties such as *red* and *tall*, and relations such as *taller than*, *next to*, etc. Logical concepts consist of negation and the sentence connectives, quantifiers, and related terms that can be used to express relations between propositions and non-logical concepts. Non-logical concepts roughly refer to the meanings of nouns, adjectives, and verbs, and sentence-level propositional contents can be build from these by applying logical concepts that express abstract structural relations.

To be sure, there are theorists who consider sentence-level propositional meaning as basic, but I leave that aside for now.[17] My intention here is not to do justice to the full spectrum of philosophical theories of content. The introductory exposition here is supposed to offer an approximate standard model to which most traditional analytical philosophers adhere or at least find familiar. Either way, the philosophers who see the propositional meaning as fundamental often at least think that propositional meaning has structure, and word meanings can be defined by substitution: What changes in the meaning

[17] But see Brandom (2000, Chapter 4) for discussion about how this idea stems from Kant and Frege and extends to the subsequent philosophy of language.

or truth conditions of a sentence, when you substitute term x for y in sentence α, determines the difference of meaning between terms x and y. The main difference, therefore, is that the *propositions first* theorists think that words do not intrinsically mean anything unless embedded in propositional expressions, and therefore the meaning constitutive relation between words and sentences goes the opposite direction. This camp maintains that conceptual meaning is ultimately determined by the whole language or some relevant subset of it; for example, that contents of theoretical terms are fixed only in relation to other terms of the theory. These kinds of holistic accounts are often varieties of inferential semantics, which will be discussed later.

The so called *classical account of concepts*—so named because it was *the* account of concepts from the time of Aristotle to the early 20th century—postulates that contents of concepts are determined by their definitions; that is, by the necessary and sufficient conditions that entity or phenomena needs to satisfy in order to count as an instance of a particular concept, which generally holds in holistic theories of meaning because introducing a definition is a way to fix the term's inferential potential. In the classical framework the idea is often interpreted as implying a hierarchy, or at least two kinds, of concepts: primitive concepts that can not be analyzed further, and the rest that can be expressed through biconditionals or other definitions. For example in classical physics $f = ma$; that is, *force* is defined as mass times acceleration. "Mass" is a primitive term, and *acceleration* is defined as a change in *velocity*, which, in turn, is defined by primitive terms "time" and "distance".

In the example above primitive terms are theory related. But what are primitive concepts, generally? The answer depends on what we are asking. For example, of you are a concept realist and think that concepts are abstract entities, then you probably expect some kind of a metaphysical answer. However, what interests us here are the possible primitives of the human conceptual system: how the contents of the concepts that we use in thinking and doing are determined, and what the fundamental building blocks of meaning are, if any. Logical empiricists, for example, thought that sensory primitives—i.e. sense data, like a patch of red hue present in the visual field—are unanalyzable primitives, and every meaningful complex proposition can be analyzed or translated into logical terms describing sensory experiences or observable properties. This was quite natural development of the empiricist epistemological tradition in the modern European philosophy. The logical empiricist project infamously proved unworkable; yet, the associated theories of conceptual con-

tent found their way to the naturalistically oriented branch of the philosophy of mind and cognitive science.[18]

The basic insight has deep connections with the fundamentals of formal logic, and it is basically identical to the modern expression of the model theory of predicate calculus.[19] Model theory is essentially a formalized version of classical theory of concepts: A model of a (first order) language consists of a set of objects—called *the universe of discourse*—and *interpretation function* which respectively maps primitive non-logical predicate and relation terms of the language onto the sets and pairs (or generally n-tuples) of the objects in the universe. If the language contains proper names, they are mapped onto individual objects. Logical constants assume standard fixed interpretations, and sentences are constructed by combining primitive terms with logical operators. The fixed interpretation of logical terms and the interpretation function of non-logical expressions guarantee that every well-formed formula has an interpretation: They are either true or false in the model. Regardless of how the interpretation is initially chosen, you can always introduce new concepts in the same way that you can construct any compound expressions (i.e. sentences). As a standard example, assume that we have an interpretation for the terms "adult", "male", and "married". Now, we can introduce the term "bachelor" by definition: x is a *bachelor* if and only if x is an *adult*, x is a *male* and x is not *married*; that is, the extension of the concept *bachelor* equals the set of adult males that are not in the set of married persons.

All this is probably completely familiar to any analytical philosopher. The important point is that when at least some of the concepts are interpreted, the rest can be understood as sentential structures built from these semantic primitives as per the classical theory of concepts. Moreover, the primitive terms, combined with the logical expressive power of the resultant representational system, strictly determine what concepts the system can represent. Now, if we model the human cognition as a sort of logic engine—as per the *classical computational theory* of mind—we need only be concerned about finding a theory of the interpretation function for primitive terms. As far as conceptual representation is concerned, the logical syntax run by the cognitive system will take care of the rest. After the semantics is fixed, we can utilize standard logical computations to carry out deductions and produce and interpret basi-

[18] There are many variations of logical empiricism. The account explained here mostly resembles the one put forward by Rudolf Carnap (1956).

[19] The modern formulation was published around the same time as the Carnap's paper referred in the last footnote; see Tarski & Vaught (1957).

cally limitless amount of novel conceptual constructs. The bottom line is that representation comes first, and reasoning and other intentional acts will follow.

Of course you don't need that interpretation function for any actual computational system to run (like a computer and perhaps the brain), but you need it to explain how the system's internal operations and its resultant overt behavior are *intentional*; that is, *contentful* and *purposeful*. In the logic-based computational story, you determine a mapping from the internal states of the cognitive system to the system of (primitive) concepts, and there: You just explained how the system has internal *representations*, which have genuine *referential content* and therefore bridged the gap between the realms of genuinely intentional and the mere mechanical.

Well, it's not really *that* easy. In what follows, I try to show how almost every piece of that old story is flawed. Still, it has produced a theory of mental content that has been well received in the philosophy of mind and has deep roots in the tradition of analytical philosophy of language. It merits particular attention, mainly because it serves as a background and contrast to the account of intentional content advertised in this work. The reader should take a note of the important distinction between primitive (or fundamental) concepts and combinatory ones, which can be analyzed into meaningful constituents. Regardless of the ultimate merit of that distinction, it captures something intuitively appealing in the phenomenology of human understanding: Concepts seem to play two kinds roles; we have *intuitive concepts* that we need *in order to* understand complex novel expressions and the world around us, and then we have *reflective* concepts that we can analyze intentionally and understand by explication. Thus, concepts are both means and ends in the acts of understanding. I think the dynamics of these means and ends has not fully received the attention in deserves, perhaps because the distinction is trivial if one adheres to the doctrine of primitive and constructed meaning. Along the way I sketch an account where intuitive concepts do not have much to do with primitivity or any kind of conceptual or epistemic foundationalism. The gist of the story is that intuitive conceptual understanding is largely a learned skill that underlies our conscious reflection, and hence it is a malleable joint product of the agent and its environment. Reflective concepts (in the sense mentioned above) are something we are trying to learn or to explicate by resorting to our pre-existing semantic and procedural intuitions.

More specifically, the theory I that will advance proposes that learned pragmatic knowledge is encoded in intuitive concepts and that this knowledge underlies our everyday reasoning capacities. In the intuitive cognition, con-

cept use and conceptual content conflate, and the resulting semantic intuitions mostly determine our explicit understanding of analytical concepts. The idea is hostile towards any fixed set of primitive non-logical concepts and universal conceptual systems. While we may or may not be innately armed with some kind of formal or structural concepts that,for example, reflect the syntactical categories of our natural language, most approximately word-level concepts are learned through our activity in our everyday environments. Thus, our conceptual system is shaped by our environment and behavior, and potentially any concept is subject to change and refinement throughout our life if our environment and/or behavior changes. Moreover, any concept that we use in explicit reflective analysis can be learned intuitively when its use becomes automatized. I maintain that *understanding* is based on these automatized ways of utilizing concepts where the use covers any intentional action ranging from overt behavior, including linguistic and social practices, to private thinking. When we engage in conceptual analysis, we are bound to use whatever intuitive conceptual resources we have, and when we use those analyzed target concepts for further reasoning, the practices of using them gradually form implicit intellectual habits. Thus, conceptual understanding is an acquired cognitive skill set, and acquiring that understanding tend to change the cognitive tools we engage with in further reasoning and analysis.

*Mis*understanding can be roughly defined as a conceptual capacity that supports systematically unsuccessful behaviors where an important special case is non-normative use of a concept in communication and public reasoning. Thus, I advocate a version of inferentialism or use theory of concepts in which "use" is conceived in the broadest possible sense of the term. I conceive concepts primarily as instruments of doing rather than of thinking. I also try to show how concepts in this sense are complex psychological phenomena and that insulating category structure from category utilization (such as reasoning) leads to a distorted understanding of conceptually structured cognitive processes. The principal claim is that in the typical case of routine reasoning, we exploit conceptual competencies to form expectations in given contexts to initiate an appropriate action. This is not a controlled process whereby we apply decontextualized formal principles to propositional knowledge but a skill whereby we apply habitual practical know-how that mostly works in familiar situations. Conversely, failures in reasoning are often failures of intuitive understanding of the functional structure of the task and not failures to apply inference principles properly. This knits the philosophical and psychological discussions of concepts, for, in this framework, there is no conceptual

gap between cognitive skills and understanding of propositional content. I acknowledge the importance of our ability to make explicit controlled inference, but it has mostly a peripheral theoretical importance to our understanding of how commonsense reasoning and expert know-how work. I propose that both of these capacities are mostly executed by our intuitive faculty wherein conceptual understanding, contextual reasoning, and practical know-how are not strictly separable.

2.1 Intentionality in the old school computationalism

A few of us today would take the Platonic idea of an independently existing conceptual realm very seriously. As we shall see soon, though, that has not stopped some philosophers of language behaving like Platonist. In the philosophy of mind, at least, it is more commonplace to think that concepts are something that somehow mediate the relation between the mind and the world. In this picture concepts in a sense have two poles: the other attached to our intentional mental states and the other to worldly phenomena. Here, the relevant mental states are the usual belief/desire type propositional attitudes: (agent) *a believes* that *p*, *desires* (that) *p*, *fears* (that) *p*, and so on. In the influential computational–functionalist theory, somewhat a synonym to the *cognitivist theory of mind*, the idea is that the attitude part of a mental state—i.e. *belief*, *desire*, etc.—can be defined by the state's causal role in the agent's cognitive system. This roughly means that fears, hopes, beliefs, desires, etc. have characteristic effects to our behavior and thinking, and these causal effects are what makes some states as fears and other as beliefs and so on. For example if I *fear that p*, then I *desire that not p*, which makes me to act in such a way that *p* does not happen, if possible. This is an application of the time-honored practical syllogism, which explicates the conceptual relation between intentional mental states and intentional action. The placeholder *p* here is usually thought to be a propositional mental representation, essentially a sentence or similar meaning bearing structure that represents some states of affairs.

The last clause above reveals how functionalism emphasizes the cognitive nature of intentional processes in the classical sense of "cognitive"; that is, the relevant states deal with knowledge or belief. The idea is to model our folk or belief/desire psychological explanations of human behavior by postulating a causal mechanism that operates on a language-like representational system. If we manage to device a model of human psychology where the basic processes

are purely causal but respect the semantics of propositional representations, we have a naturalistic reductive theory of intentionality of human thought and behavior. There are other ways to meet this goal, but the basic idea in computational functionalism is that logic shows how you can define semantically constrained syntactical operations over propositions, and computers demonstrate that you can execute these processes in purely causal system, which does not care about semantics. The key semantic constraint in logic is the preservation of truth, i.e. that you can derive only true sentences from true sentences. The following is perhaps the major reason why cognitivistic theories are often framed around propositional content: truth perseverance is the clearest semantic constraint which we know how implement in purely formally defined systems, and propositions are the simplest elements in logic that have a truth value. In any case, the core idea of psychological computationalism is that the world is somehow represented in a computational system which, in turn, is implemented by our brain. If the representations refer to worldly phenomena, and the mental processes are crafted to respect their representational character (i.e. meanings) then when our brain is processing these symbols we are actually making inferences about the world. This aboutness—the intentional character of mental representations—is what builds the mind-world connection and enables us to function in sensible ways in our environment.

But does our commonsense psychology actually work like this? It seems that many propositional attitudes are about objects, or classes of objects, but not about propositions. For example, I may be afraid of snakes with no reference to any specific states of affairs that includes snakes. For cognitivists, this is not a very pressing issue, though, and generally they conceive attitudes like fears and desires as behavioral dispositions toward objects. While this is a seemingly innocent detail, it may have far-reaching implications if we primarily focus on behavioral dispositions towards things rather than on logical operations over propositions when we are formulating our theories of intentional content. Be that as it may, in the next section we will see how ideas form logical semantics turned into theories of mental content.

2.1.1 Causal theories of mental content

The computational theory of mind in itself does not explain how the contents of mental representations are determined. Common wisdom concerning formal systems is that the mere syntax of symbols does not determine their semantics, and, as far as I can tell, no one has advocated computationalism where content

derives from syntax without any need for further specification. However, it is not uncommon to think that semantics is somebody else's problem or that adverting to intentional content in cognitive science is anyway a non-starter.[20] However, if you intend to explain intentionality within the cognitivist framework, then some account of mental content is clearly mandatory. Prevailing philosophical theories in this regard are causal theories or so called conceptual role semantics. These frameworks are not necessarily exclusive. On the contrary, it is typical to demand that adequate mental semantics requires both causal and conceptual role factors. I explain these theories separately at first and then discuss them together.

As the name suggests, causal semantics postulate that the content of mental representations is determined by a causal connection between representations and their putative referents. In brief, the core assumption is that the occurrences of a specific phenomenon or object a systematically cause a token of specific symbol or symbol structure α to occur in one's cognitive system. For example, whenever I see a cat, allegedly some representation α pops up somewhere in my cognition. If the presence of cats causes this systematically, then there is a nomological causal link between cats and α, and it is *prima facie* reasonable to conclude that α stands for (i.e. represents) cats in my cognitive apparatus. Causal theories hold that the same representational tokens that occur in perception can be used elsewhere in the cognitive system; thus, referential mental content can be generally defined by the causal connections of thought contents and extramental reality.

To be precise, it is not reasonable to assume that it is object a as such that causes the occurrences of α but a combination of specific features of a that allow for its categorization; for example its characteristic looks, sounds, smells, etc. Thus, pictures of cats, for example, could correctly elicit tokens of α, while they evidently are not actual cats. What is required for the causal link to do its job is a sufficient set of more or less primitive features associated with cats to be available for perception. Although causal theories of mental content come in many guises, the more plausible and popular formulations specifically postulate that there is some primitive (albeit, unfortunately, invariably unspecified) set of mental representations that correspond to some simple atomic features

[20] For example Jerry Fodor (1980) and Stephen Stich (1983), respectively; though both have later changed their hearts about these matters; see Fodor (1987) and Stich (1996). Also a host of theorist reject both the belief/desire psychology as a proper *explanandum* of cognitive science and classical computationalism as viable *explananda* (e.g. Churchland, 1989), but these views are off the point right now.

whose semantics are grounded by a causal connection, as specified above. In practice, though, some theorists are quite liberal in what counts as a proper primitive feature or property. For example, Jerry Fodor, a major proponent of causal semantics, holds that any property at least potentially counts (see Fodor 1987, 1990). He uses examples such as the property of being a horse as a valid causal source. In his theory it is precisely properties, and not objects, that serve as relevant causal factors. Others are more stringent. For example Robert Rupert (1999) frames his theory around simple natural-kind terms, which we supposedly learn in the early years of our development before we acquire language and a full-blown adult conceptual system. He leaves the question of whether these representations match lexical items in our natural language to the developmental psychologists. In any case, in causal accounts the primitive representations are actually *indices* in the Peircean sense rather than conventional symbols since their occurrences are caused by perception not substantially unlike smoke is caused by a fire.

Not all mental contents are directly grounded by a direct mind–world-connection, however. More elaborate representations that stand for cats, presidents, and black holes, for example, can be synthesized from atomic constituents in the cognitive system, given that these are not treated as primitives in the theory.[21] The role of the computational theory of mind is to explain how complex mental representations such as propositional ones can be constructed from primitive symbols. Formal languages in general—and logics and programming languages in particular—contain recursive rules to compound primitive symbols from a finite set to an infinitude of well-formed complex expressions. These rules enable the interpretation of all well-formed symbol structures that constitute the formal language, provided the semantics of the atomic symbols are determined. By the same principles, the causal account of mental content, combined with the computationalist theory of mind, explains the productivity and systematicity of thought; that is, our ability to entertain an infinite amount of different thoughts, and our ability to understand the meaningful

[21] Note that if we are too liberal in what counts as a proper primitive concept, we quickly run into problems with distant causal connections: Is *a tiger* a primitive term, or theoretical concepts like *an electron*? Many of us have encountered tigers only through stories, TV, and other media, and while electrons are all present, our cognitive representations of them come through scientific instruments, textbooks, etc. One needs to be very cautious here because if we restrict primitive concepts to sensory primitives, the theory essentially becomes a form of logical empiricism. If not, we need to explain how people can have primitive concepts that are not caused directly. Well, distant causal connections are causal connections, but a good case can be made that this issue undermines the the whole project, see Kroon (1987).

but arbitrary recombinations of structured conceptual representations, such as novel sentences. Productivity means that you can use already interpreted concepts to build new ones, like x IS A UNICORN $\Leftrightarrow$ x IS A HORSE, AND x HAS A HORN IN ITS FOREHEAD, or $f = ma$. Systematicity means that if you can think that *the cat chased the mouse* you can think that *the mouse chased the cat*.

Productivity and systematicity are arguably the hallmarks of the human conceptual system and also serve as evidence that recursive symbol processing is taking place for it is otherwise tricky to explain these phenomena. In brief, the meaning of the proposition "the cat chased the mouse" reduces to the meanings of "cat," "mouse," and the past tense of the transitive verb "x chases y." Thus, you can use the hypothetical computational mechanism to generate and parse any well-defined formula using these expressions. If the formalism works much like logical languages, the meaning of all symbol structures reduce to meanings of their constituent elements and, eventually, to the meanings of primitive concepts that obtain their meaning from the causal connection.[22]

Not all versions of causal account invest much in the computational theory of mind but instead focus on the key issue, that is the nature of the causal relation itself. The version I present here mostly resembles the theory that Jerry Fodor developed during the 1980s and 1990s (Fodor, 1987, 1990). Another notable advocate of the causal account is Fred Dretske (1981; 1988) who relies on the notions of information and function instead of Fodor's law-like causal connection. It is not my intention to do justice to the whole spectrum of these theories but to present a version of causal semantics that I consider to be as concise, representative, and plausible as possible without going into specifics. Unfortunately, all formulations of this theory that I know of suffer from the same three flaws: (a) because of the notorious *disjunction problem*, they cannot handle misrepresentation in a non-question begging way, (b) they

[22] The idea that mental representations form a language-like system, *the language of thought*, has deep Chomskyan roots (in Chomsky, 1957, 1965) and was developed before causal semantics (see Fodor, 1975). Note that in computational functionalism, at least in the Fodorian version, the computational mechanism plays a dual role: the content of an inner representation is causally determined, but the fact that it is a representational token in the first place (such as the content of a belief or a desire) is determined by its computationally implemented functional/causal role in the psychology of the organism that uses it (see Fodor, 1987, Chapter 1). That is content functionalism, which is different from the explanatory role computational mechanism for parsing the meanings of complex expressions. This can cause some confusion since apparently, at least in the received view of the theory, it was not realized that these two roles of computational mechanism are separate and thus computational theory is actually independent of functionalism, see (Piccinini, 2004).

are applicable only to a very limited set of concepts, and (c) they are empirically inadequate and philosophically dubious since they dismiss the role of knowledge-based inferences in determining mental content.

The disjunction problem is that, in principle, dog thoughts could systematically be caused to someone by foxes in poor lighting, by poking the brain with electrodes, or by asking "what was your pet when you were a child?" Moreover, there are causal factors that necessarily enter between dogs and dog perceptions, such as neural processes in our sensory organs. Since none of the mentioned causes of dog thoughts involve actual dogs, they should not count as a part of the dog concept. Otherwise *dog* would mean *a dog, or a fox in poor lighting, or neural event x, or ...*; but, unfortunately, this is what causal theories seem to imply. Thus, the problem is to separate content- determining causes from non-content-determining ones. Despite several intriguing attempts this has proven extremely difficult without smuggling some semantic concepts into the putative reducing theory, thus rendering the sought reduction of semantics to causality circular and making the content determination dependent on something other than causal relations.

For example (in 1990) Fodor tries to solve the problem by introducing a further asymmetric dependency condition: Assume that $X \rightarrow \alpha$ (read: "X causes a tokening of representation α in one's cognitive system, and this is a law like regularity.") The disjunction problem arises when there is also a nomic connection X *or* $Y \rightarrow \alpha$, where $X \neq Y$. The asymmetric dependency conditions means that if α means or refers to X, then the connection $Y \rightarrow \alpha$ must be asymmetrically dependent on the fact that $X \rightarrow \alpha$, in the sense that if you break that connection, then the $Y \rightarrow \alpha$ goes too but not the other way around.

The theory is supposed to work like this: Assume that $H_2O \rightarrow \alpha$ holds for someone. Later, that person runs into a sample of watery substance XYZ that she cannot tell apart from H_2O; thus, we also have that $XYZ \rightarrow \alpha$. Assume also that she uses the word "water" to denote α.[23] Now, two options are available here, where Fodor's theory may lead us.: 1° As a matter of empirical fact, the person is unable to make the difference in any circumstances and does not care. Then her concept α simply refers to the disjunctive set H_2O *or* XYZ. Her "water" concept is arguably different from our, even though she uses the same word as we do. No problem here since there is no misrepresentation, just

[23] This is the famous "Twin-Earth" case devised by Hilary Putnam, which features regularly in these discussions. The uninitiated may want to consult his (1975) for elaboration of this thought experiment.

different concepts. In another case 2° she maintains that "water" refers to the single substance H_2O and learns her mistake. In this case $XYZ \rightarrow \alpha$ is still in place but (read this quote carefully):

> ...intention to use "water" only of stuff of the same kind as the local samples [of H_2O] has the effect of making its applications to XYZ asymmetrically dependent on its applications to H_2O ceteris paribus. Given that people are disposed to treat "water" as a kind term (and , of course, given that the local samples are all in fact H_2O) it follows that—all else equal—they would apply it to XYZ only when they would apply it to H_2O; specifically, they would apply it to XYZ only when they mistake XYZ for H_2O. (Fodor, 1990, 115)

Now, it is one thing to claim that you need causal connections to fix extensions, but it is quite another thing to claim that it is enough to fix the meaning of mental content. Here Fodor says that you need intentions to use the concept as a kind term, and presumably you need to know that H_2O is different from XYZ. Right at the next page he refers to "a settled policy of using 'water' as a kind term" and presumably we need to know that our policy is that water refers exclusively to H_2O. The message is *not* that you can fix the contents of your concepts by their law-like causal co-variation with their referents only. To settle the basic meanings, Fodor explicitly exploits the notions of inference, *intentions to use* concepts in specific ways, and commitments to linguistic practices.

Fodor is implicitly invoking a version of a theory of reference called *causal descriptivism*. In a nutshell, the theory states that the term α refers to a if and only if a is causally connected to α *and* satisfies some identifying description D. E.g., in the above example, D says that *water* is a substance with specific molecular structure etc., and there is a causal link between our use of the term "water" and the substance that flows in rivers, which happens to be H_2O rather than XYZ. In the cognitivist version, this reads that a causes an occurrence of representation α in one's cognitive system and satisfies cognitive content D associated with a, like intentions to use α certain ways that are proper with respect to a. Arguably, this account resonates better in case of theoretical terms than the pure causal theory (Psillos, 2012). We then need a theory of content for descriptive apparatus employed in D. In what follows, we see how this can be done by introducing a hybrid theory that combines inferentialism with causal theory. After all, a mere causal relationship seems to be insufficient

to define the mental contents, and restrictive principles are needed to track the conceptually relevant causes. Later we see that this problem seems to contaminate all kinds of cognitivist theories of content, including the hybrid account.[24]

Moreover, the causal co-variation seems unnecessary or even inherently wrong kind of principle for a large class of concepts. Take abstract concepts such as *democracy* or *transfinite set*. Causal semantics seems intuitively reasonable with concrete object concepts, but no specific entities instantiate democracies with which you can be in causal connection. Things like the constitution, decentralization of power, and social equality are essential constituents of democracies, but it is far from clear how to reduce these kinds of phenomena to causally efficacious entities with which you can interact. Better yet, many mathematical concepts do not even have instances. Sure enough, there is a causal story of how, e.g., I came to grasp the mathematical concepts I know but attending to math class does not make you to participate in a causal process that emanates from prime numbers, transfinite sets, logical connectives, and so on. To acquire these concepts is to learn how to use them.

It is hard to see how a pure causal theory of content might work, apart from perhaps fixing the content of some very primitive perceptual features; and if conceptual content reduces to sensory primitives and their logical combinations, we are back at flat-out logical empiricism. Another pressing problem of Fodorian version of causal semantics is its empirical inadequacy. This theory aims to explain the productivity of thought and language, and as we saw above, it does this by resorting to the classical definitional theory of concepts. In that respect the causal factor is a bit of a red herring here since it serves only to get the representational system off the ground. Therefore the *computationalist* causal semantics, at least, seem to stand and fall with the classical theory. While we do use definitions in concept formation, this is generally true only of a very limited class of concepts that are mostly of technical nature. In both philosophy and psychology[25] it became clear during the 20th century that everyday natural concepts are structured rather differently. We will have review these theories in the next chapter; however, it is worth discovering what we

[24] But note that the following discussion is not supposed to make an argument against causal descriptivism in general but only against the discussed cognitivist version. This is because the cognitivist version here is supposed to be a reductive theory of intentional content while causal descriptivists in general seem to be interested only in fixing referents of kind terms.

[25] See Wittgenstein (1953) and Rosch (1978), respectively, for standard sources.

achieve by augmenting the causal theory with inferentialism. First, we examine pure inferential semantics and its problems.

2.1.2 Conceptual role semantics

Causal theories may seem a natural choice for those in the cognitive camp who stress the *reference relation* in content determination; however, variations of conceptual role semantics (CRS) are more obviously *computational* accounts of mental content. These theories have their roots in Wittgensteinian "meaning is use" conception of language and in the proof-theoretic semantics. The core idea is to provide the meanings of syntactic tokens by itemizing the relevant inferences that their putative interpretations would enable. For logical operators, this strategy is famously simple and effective: Consider a language with a generation rule "If α and β are sentences, then $\alpha \wedge \beta$ is a sentence." Then the following inference rules for connective $\wedge$ define it as the logical conjunction: "If α and β, then $\alpha \wedge \beta$" and "if $\alpha \wedge \beta$, then α (and likewise for β)". These rules mirror the semantics of the conjunction. By the tools of axiomatic set theory, one can define essentially all the interesting mathematical concepts similarly. Advocates of CRS reckon that a similar method can also be used to define non-logical concepts.[26]

The core idea is based on the fact that any concept A has several relations to other concepts $B_1, B_2, \ldots, B_n$, wherein the relations could be logical, constitutive, causal, etc. These relations can be stated in terms of inference rules. For example, if A refers to chairs and B to furniture, then "if x is A, then x is B". If disease D causes symptom S, we can state that "if x has D, then x has S". In the case of constituent relation, we have that "if x is A, then x contains B", and so on. Next, consider that interpretations are given to concepts B that appear in the consequent of the inference rules. Then by substituting the occurrences of concept A by uninterpreted syntactic token α in the conditionals, we may obtain, for example, the following expression: "if x is α, then x is a *smallish predator mammal* and a *typical pet*, x has *four legs*

[26] Pure versions of CRS without any use of a referential relation or a notion of truth have perhaps never been very popular, but at least William Rapaport (1995) has advocated that kind of CRS. Some critics, such as Field (1977), posit that Gilbert Harman (1974) advocates a version of the theory presented here. Harman (1987) has objected that in his theory the possession of a concept is, at least in some cases, associated with perceiving a thing and acting appropriately towards it, and therefore this is a misinterpretation. Because of that tenet Harman can be loosely grouped together with the two-factor theorists discussed in the next section. At least it is questionable whether his theory could work unless it implicitly involves two-factors (see Block, 1986).

and a *tail*, x chases *mice* and says "*meaow*". It seems that α stands for the cat. Perhaps we can expand the list in the consequent enough to capture all the essential features of cats or some total description of "cathood". Then the assertion that α="the cat" can hardly provide any further information. "The cat" becomes just a label for the uninterpreted symbol α, whose interpretation is provided by the description appearing in the consequent of the conditional. If this is correct, then the meaning of the concept *cat*, perhaps any concept—can be encoded as a set of inference rules, and the occurrences of the concept itself can be then replaced with syntactic token α without any loss of content.

In this manner we usually explain and introduce novel terms, and therefore CRS looks like typical definitional view of concepts. However, there are significant differences. First, definitional view holds that concepts can be defined by biconditionals that contain a list of conceptually necessary and sufficient features of the target concepts, but this usually is not the right interpretation of inferential relations in CRS. In fact, we typically do not know how to define even common concepts such as *furniture* in a non-controversial way, although we can use them with high efficiency. Well-defined concepts tend to appear mostly in technical, scientific, and related specialist contexts, and often we have problems in understanding precisely those kinds of concepts if we are not experts in the relevant field.

Thus, casual observation shows that well-definedness often indicates difficulties rather than fluency in concept use. Therefore whatever merits conceptual analysis may have, the definitional approach seems to be ill-suited for the concept of *concept* in the philosophy of psychology. However, inference rules in CRS represent a default body of inferences that a competent user of the concepts is disposed to make in certain circumstances, and those inferential dispositions typically extend well beyond conceptually necessary and sufficient features. For example, we may tend to infer that breaking one's legs implies sick leave, but this is hardly a *conceptual* fact about bone fractures or legs. (Though this might be a conceptually constitutive rather than just empirical fact about sick leaves.) Thus, inferential relations that constitute the concepts in CRS are typically thought to contain, *inter alia*, causal knowledge and default associations with thematically related concepts. Basically, the inferential rules encode a list of facts about the subject matter, and therefore concepts in CRS resemble miniature theories rather than definitions. This is psychologi-

cally more realistic than the definitional view[27]; however, it also constitutes a problem that will be explained below.

Second, the definitional view can hardly serve as a reductionist theory of meaning, since it relies on already understood terms if it is not augmented with a theory of fundamental meaning, such as the causal theory. The previous cat concept example hardly looks like a reductionist definition either, for, although a syntactic token α replaced the term "cat", the consequent in the conditional contains interpreted terms such as *mammal* and *chases mice*. The point in CRS is that, e.g., the term "mammal" has to be also ultimately substituted by a syntactic token, say β, and its semantics provided by a set of inference rules, such as "if x is β, then x is an *animal*". Again, the concept *animal* should be replaced by syntactic token γ, and its semantics be provided by a set of inference rules, and so on.

By iterating the above method, the result should eventually be an inferential network, wherein all the concepts are replaced by their respective syntactic tokens, and the interpretation of any symbol is laid out *solely* by it's inferential relations to other symbols in the web. Thus, the contents of symbols are not associated with the syntactic tokens as such but with their interrelations, and the upshot is that no concept has any intrinsic content. Remove a symbol from the web, and it loses its contents. This may sound counterintuitive but often if one is prompted to explain the meaning of any particular concept, it seems that the only option is to give its description in terms of other concepts. This observation is captured in the classical account of concepts. It is difficult to say what else the possession of a concept would mean other than the ability to use it properly in thought and communication, and the advocates of CRS contend that there really is nothing more to it. Finally, the theory postulates that the inferential relations between concepts are realized in the cognitive system as computational rules that the cognitive agent is disposed to execute in the relevant circumstances, thus completing the reductive theory of mental content by merging it with the computational theory of mind.

Conceptual role semantics may be rather natural theory with abstract concepts, but there appears to be concepts whose contents cannot be readily understood by a description but only by personal acquaintance, such as *red*

[27] A theory paradigm of concepts is a relatively recent innovation in cognitive and developmental psychology and it lacks a canonical formulation; see e.g. Chapter 4 in Machery (2009) and Chapter 6 in Murphy (2002) and Carey (2009) for good recent summaries and discussion. Nevertheless it has become one of the standard textbook accounts and has a good empirical standing.

and *pain*. It is not that one cannot use these concepts effectively without prior contact with their instances—that is, without ever seeing red or feeling pain—but the problem is that mere knowledge about the rules of use of these concepts hardly suffices for really knowing what they are about. Hence, it appears that the blind lack some content of the concept *red*, which the others do not; even though the blind may equally well understand and use the concept just like most of us can understand, e.g., the concepts x-rays and radio waves even though we cannot sense them. Thus, arguably, there is content that does not reduce entirely to inferential links, and hence, there seems to be some, even if very limited, room for primitive unanalyzable meaning after all.

Nvertheless, there are more focal problems with CRS that affect all concepts regardless of their variety. First, the theory is inherently holistic, which causes all kinds of trouble. The holism comes about by the fact that inference is a transitive relation: if α affords an inference of β, which in turn affords γ, then there is an inferential relation from α to γ. For all I know, it is reasonably plausible that there is a chain of inferences between any two concepts, however laborious, and thus the content of any concept depends to some degree on every other by the tenets of the CRS.

Conceptual holism is somewhat odd, but how bad is it? Jerry Fodor and Ernest Lepore (1991) have pointed out that some may believe that cows are dangerous, and therefore possess the inference rule "if x is a *cow*, then x is *dangerous*." This would render dangerousness as a part of one's cow concept. If this is the case, then in general concepts should be highly idiosyncratic, since we all have different experiences and beliefs about many things. The problem is that the inferential networks between people tend to be dissimilar, and since the content of any concept reduces to its position in the whole network, it seems that by the holism of CRS two people cannot, strictly speaking, share any identical concepts. Even if they superficially agree completely on what cows are, for example, they might disagree what animals are and thus have different cow concepts. This may still be acceptable but the situation becomes quickly untenable. Because of the holism, even disagreement about, say, black holes imply different cow concepts. The bottom line is that this makes the inferential networks, and thus mental contents, incommensurable between people. If the networks diverge *at any point*, then you cannot identify any node α with node α' in the other system, because the sole identity criterion is the nodes' place in the network, and different networks simply do not have identical places. Moreover, if you cannot compare the conceptual systems between persons, then on what basis can you compare any such conceptual network with extramental reality?

Besides the inferential relations, additional criteria are needed to determine what the nodes of the network refer to. Hence, CRS in its skeletal form is clearly incomplete.

Perhaps the remedy is to find corresponding inference clusters in different conceptual networks and invent a similarity metric to compare the contents and the identity conditions of their tokens. Do people not have different conceptions almost about anything to a degree? The answer is probably in the affirmative; however, there are still hard cases. Consider the story of ancient people who thought that a canopy covers the sky at night and stars are holes that let the celestial light shine through. Whether this story is true or not,[28] it seems that those people and us, armed with a modern scientifically informed world view, have nothing in common concerning the concept of stars—save the perceptible bright dots in the night sky. Thus, the prospect of finding similar local conceptual structures between them and us is highly improbable. In that case, on what grounds were the putative star thoughts of ancient people actually about stars? Then again, on what basis our thoughts are? What is *the* conceptual structure that sets the standards? This point is worth stressing since CRS seems to collapse into the core problem of content in computationalism: the symbols as such or their formal relations do not determine contents, which is a necessary fact of formal systems, and the demand stands for some extraneous act of interpretation. Either, we need causal theory after all or another non-inferential procedure that we share with ancient stargazers and other people with different conceptual systems to fix shared referents. These are reasonable demands; however, they are not available for pure versions of CRS. In my opinion, the only viable naturalistic theory of mental content is a version of inferentialism, which is framed with references to (human) practices. But before going into that let's first see how CRS fares in combination with the causal account.

2.1.3 Two-factor theories

Although some authors consider the causal and inferential semantics as rivals (e.g. Fodor & Lepore, 1991), a more fruitful attitude is to view them as complementary. The intermixed theory is usually called *a two-factor conceptual role semantics*, and the division of labor between causal theory and inferential account are frequently dubbed as domains of wide (referential or external) and

[28] As a side note, this may be the actual conception of stars at least of the ancient Israelites (see Wright, 2000, 54–56).

narrow (inferential or internal) content or more metaphorically as long-arm and short-arm aspects of mental content, respectively. As noted on page 43, this sort of hybrid theory may be considered a cognitivist version of causal descriptivist theories of reference (see e.g. Psillos, 2012). The version of the theory I present here is basically that of Ned Block's (1986). Also Hartry Field (1977) and Gilbert Harman (1987)[29] have introduced relevant and broadly similar accounts.

Earlier we saw that causal and inferential theories both have types of concepts which they seem to cover better that the other. As the first approximation, it could be reasonable to let the causal theory handle concrete terms and CRS the abstract terms that do not have physical referents, such as *freedom* and *Ackermann function*. Then concepts would be divided in two clusters: the other containing concepts with causally determined content, and the other handled by the CRS. This move would be too swift, however, since abstractedness comes in degrees, and it is impossible to draw a definite line between the concrete and abstract. Some concepts do not refer to physical entities as such but can be introduced by ostension of concrete exemplars, e.g. *parity*. Moreover, concepts such as wait and perhaps verbs, in general, are in a sense abstract but something that we can learn through concrete experience. Concepts such as a doctor have corporeal instances but could not be properly possessed without having some relevant inferential knowledge. In other words, you cannot categorize doctors just by perception, and it would be odd to say that someone has the concept of doctor without knowledge of how one can become a doctor, what doctors do, and why. It is safe to say that the proper possession of most of the concrete concepts requires an ability to make relevant non-trivial inferences about them. Thus, although the putative two-domains may not overlap completely, at least they intersect to the degree that makes their categorical separation very dubious.

Therefore dividing concrete and abstract concepts into two separate clusters is not a very good idea, and the focal point in two-factor CRS is that a mental representation can have two kinds of determinants of meaning: (a) referents that are picked out by a causal connection and (b) cognitive content that is determined by a set of relevant inferences. Consider in incommensurability problem in CRS. If you allow concepts to be at least partly defined by their referents determined by causal relation, the problem, if not solved, is at least considerably mitigated. This is because you then have a criterion to fix at least

[29] Although it is somewhat contentious whether Harman advocates a one- or a two-factor theory; see footnote on page 45.

preliminary referents for a host of syntactic tokens in networks, rendering them comparable. For example, you can anchor the mental representations caused by bright dots in a night sky in both ancient and modern astronomers as *stars* regardless of how wildly their cognitive contents of the dots may vary. This, however, does not resolve the problem in the case of abstract concepts. For example, if two people with very different conceptions of democracy are debating on the matter, then on what grounds are they using the same concept? Is it enough that they use the same word and that they recognize at least some common features associated with it? It is not possible to fix referents by a reference to democratic states, for example, since what the putative extension actually includes is often precisely the problem.

Regardless, not only does the causal theory alleviate some problems of CRS, but the theories are mutually supportive both ways since the disjunction problem of the former can be relieved by CRS to some extent. Consider a neurosurgeon who activates dog thoughts in one's mind by stimulating the brain with microelectrodes. By the causal theory, the electrode assembly is included as a referent of the concept *dog*. By the CRS, on the other hand, the referents of the concept *dog* are mammals that bark, chase rabbits, have four legs, and so on, which microelectrode assemblies most definitely are not. Therefore, there arguably are non-question begging grounds for claiming the installed microelectrodes are a spurious causal source of the dog thoughts. Thus, besides providing means to handle abstract concepts, CRS may also resolve the problem of misrepresentation—the most pressing problem of causal theories. Unfortunately, however, the problem is tougher than it looks.

The central problem in the two-factor CRS concerns the relation of the factors. As explained above, the mental content can be considered a composition of two constituents: referent, which is determined by the causal theory, and inferential content, which is determined by the CRS. The principal point of postulating two distinct determinants of content is to provide means to fix the extension of concepts when CRS fails to do so and, at the same, time handle the problem of misrepresentation infesting the causal theory, and thus to produce a prescriptive theory that explains how we can be in error when applying concepts. This is vital if we are to describe human behavior in terms of propositional attitudes that rely on conceptual content. If the theory of mental content renders every use of concepts correct by default, there can not be false beliefs. In that case, there are hardly any grounds for applying the concept of belief and other propositional attitudes, and hence no effective use for propositional attitude ascriptions. The point in two-factor theory is that the factors

are supposed to determine different contents in specific cases. Otherwise the other factor would be redundant, or we have to resort to the aforementioned unattractive two-domain theory.

If the factors should always determine the same content, we would need to explain how and why they are aligned in this convenient manner, which ostensibly is not any less opaque a problem than the one considered here. Moreover, the factors would not relieve the problem of misrepresentation, since they both would always fail and prosper in complete correlation. If one of the factors is primary, we will need a theory to help us understand the circumstances in which we should apply the secondary one, instead. If neither factor is primary, we will need an account to help us know the factor on which we should rely when they determine different contents. The problem is that there is no obvious solution to guide us to know how to do this in two-factor theories, and the problem with disjunctive categories persists.[30]

Let us assume that I systematically mistake elms for beeches. I associate the correct inference rules with elms but beech trees cause the tokening of my elm representation. Now, how can we decide whether I am really confusing the referents and not that I have some misconceptions about the disjunctive category that contains elms and beeches? If this kind of confusion happens in perception, we may find it natural to say that the object is a beech after all and that disjunctive categories make poor concepts—hence we wish to avoid preferring the option that I have the concept *beech or elm* with improper inference rules that pertain only to beeches. This is the basic intuition behind the theory; however, what we really want is to explain this kind of semantic intuitions and why we have cross subjectively robust preferred ways in grasping the reality. The two-factor theory is not helping much. There seem to be no satisfactory way around this problem but to rely on semantic intuitions on which would be the appropriate factor in a given situation. Thus, we need either justified semantic intuitions or a theory of content *in order* to apply the two-factor account properly, which is an untenable situation because the sole point of the two-factor CRS is to be a theory that explains such intuitions or alternatively removes the need to resort to them.

[30] For a more detailed discussion see e.g.(Perlman, 1997). Note also how at this point a theorist might be lured to exploit dual-process theory of cognition to resolve these issues: the alignment and primacy problems may turn out empirical if one associates System 1 with concrete concepts and the causal factor, and likewise System 2 with abstract concepts and the inferential factor. That's what I tried and failed.

Deep down the problem is that although the causal factor may help us fix referents for some terms, we still have the problem of determining what counts as correct use. Without restrictions, every use of concept counts as content determining, and hence necessarily correct. Moreover, although incorrect use cannot be content determining, we do not want every possible correct use to count either. For example, not every empirical fact we happen to believe about a target concept can count as content determining; this is because some of our conceptions may turn out to be false without changing our concepts, and if correct use of a concept requires the tracking of all the facts of its referent, then we barely grasp correctly any concept at all. In any case, the wholesale conflation of meaning determining factors with contingent facts sounds a bit too radical move. This becomes very clear in the case of abstract concepts, which do not have empirical content and can not be defined by using natural or scientific kinds, such as socially constructed contents. It seems that two-factor theorists need a criteria of correct use that is independent of actual use and empirical facts but what could that be?

Mark Perlman (1997) has concluded that the fundamental problem with the two-factor framework is that it simply cannot work without invoking synthetic/analytic or some effectively similar essentialist distinction of content. Otherwise, it is impossible to frame any cognitivistic or related internal conceptual role theory without making meaning completely idiosyncratic and identical with belief, thus rendering all use error-free by definition. The way out is to employ some non-cognitive criteria of what counts as conceptually proper use, and in CRS it is implicitly assumed to be analyticity.[31] Therefore cognitivist concept realism is untenable unless analyticity is built in to our cognitive system.

How fatal is this problem? Cognitivists in general are probably quite hostile toward Platonic concept realism; no longer does anyone believe in strict analytic/synthetic distinction, and blatant psychologism or logical empiricism does not sound that great either. So what to do? This provides an excellent reason to abandon the two-factor account as well as the causal and CRS theories. This, however, would be a severe blow to the attempt of explaining mental representation on the basis of cognitivism. Perhaps you do not need to be an old-school computationalist; however, it may be difficult to see how you can be naturalist and realist about intentional content without some reductive account that identifies intentionality with cognitive contents: If the origins of intentional content are not inside our minds, where could they be residing?

[31] See also Fodor & Lepore (1991) for similar remarks on inferentialism.

At this point, philosophers should probably turn to psychology and try to find out if empirical research can shed some light on the issue, and this has fortunately somewhat happened. Historically, this chapter has covered the philosophical discussion of concepts from the early 20th century to the late 1980s and mid-1990s. Since then, there has been an emerging trend to abandon the idea that mental processes should be conceptualized after linguistic and logical models. There are at least three good reasons why this is a good development:

1. The use theory of concepts is independent of the referential and truth-conditional theory of meaning, and it can survive without strict content determination and bivalent true/false assignments of propositions. It seems that the analytic/synthetic distinction creeps into the theory from the back door because of the unquestioned and perhaps mostly tacit demand that correctness of use is a categorical notion. Perhaps we should think of concept use and conceptual understanding as a learned skill where, instead of veridical reference, correctness is a graded notion about the successful deployment of concepts in action, including communication and reasoning.

2. In that case, we make advance if we know the etiology of that skill and also what we learn when we acquire that competence since that should tell us what phenomena our conceptual competencies actually track. The whole point in conceptual role semantics was supposed to be about our actual use of concepts and not about whether meaning is analytically or metaphysically construed. The problem remains to distinguish content determining use from non-content determining; however, as per item 1°, the notion does not need to be necessarily categorical, and the target of our inquiry should be on the practical reality of the cognitive agent rather than objective conceptual reality.

3. Moreover, when analytical philosophers probe their intuitions about what (correct) concept use is, they are actually using that very skill. This does not necessarily imply vicious circularity; however, it is reasonable to discover the empirical nature of that activity. The point is that if we conclude in our analysis that we should replace the idea that concept use is applying *a priori* analytical knowledge with the idea that it is an acquired psychological skill, it becomes more reasonable to investigate its

nature empirically, similar to any other cognitive capacity. Otherwise, we are left with nothing but speculative or introspective psychology.

If someone sees an epistemological problem here, all I can say is that this is a philosophical result after all.[32] The result is not that we should completely abandon philosophy in our quest for intentional content but that we should abandon strict internalism, logic-based modeling, and cognitive reductionism. Instead, we should look for phenomena that structure the world as meaningful to us and leave it to the psychologists to answer the internal questions of how our concepts are structured and how our cognitive capacities utilize them. This project is remote from a reductionist cognitivist theory of thinking with propositional content and that is precisely the point.

The philosophers who began with the theory of content by pondering how to fix the semantics of propositional thought to explain conceptually structured behavior (including thinking) got the project completely backwards. You need intentional content *in order* to have meaningful thoughts, and, in my opinion, the only way you can have that (empirically but perhaps not conceptually speaking) is by way of intentional action. This is how we shall proceed. Before our empirical discussion, we briefly look at two naturalistic but non-computationalist theories of conceptual content, namely neobehaviorism and neopragmatism. Both of these accounts diverge from the theories thus far considered, in that they place the origins of intentional content outside the cognitive system. The point in these discussions is to lay the conceptual foundations for a pragmatistically oriented version of inferentialism that will be developed along the way.

2.2 The origins of intentionality reconsidered

In the context of the history of psychology, the term ”neobehaviorism” often refers to the second generation of behaviorism that reigned mostly in North-America from the 1930s to the 1960s; it is most prominently associated with the works of Clark L. Hull, Edward C. Tolman, and B. F. Skinner. Thus, the name refers to the research tradition that is usually called simply ”behaviorism”—a

[32] And any philosopher, who thinks that we need a good analytical understanding of the phenomena under scrutiny before empirical investigation is even possible, probably believes in fundamental analytic/synthetic distinction and therefore does not buy the above argument against cognitivist CRS anyway. I mean to limit this discussion to the philosophy of psychology and the mind. Whether the philosophers of language, for example, should take note is their concern. In the next chapter, though, I give some reasons why they perhaps should.

convention I will follow. I use the term *neo*behaviorism to denote a philosophical theory of the semantics of intentional terms that carries no direct implications to empirical psychology. Indeed, the account is entirely compatible with the methodology of cognitive psychology. My use of the term, like that of *neopragmatism*—a.k.a. *linguistic pragmatism*—is borrowed from John Haugeland's (1990) well-known paper "The Intentionality All-Stars" in which he compares three main outlines for a philosophical theory of intentionality. I do not treat these views consistently as separate because I think that neobehaviorism and neopragmatism are commensurable and even complementary.

I believe that there is no preferred foundation of conceptual content, if not that concepts, at root, are capacities for systematic behavior and their contents depends on who uses the concepts, how, and to what ends. These capacities are malleable and adaptive, and their contents are determined by our needs, environments, and capacities as biological, social, and active creatures. Later, I will explain how this means that there are multiple determinants of intentional content and that actual conceptual systems are always somewhat idiosyncratic, although a shared way of life tends to suppress the diversity between agents. Hence, the more our practical realities coincide, the more easily we understand each other; this is because the similarities of our practices mark the similarities in our learned intuitive capacities to grasp how things hang together in our environments. Thus, at the end of the day I actually will be selling a sort of enactivism; which is the central philosophical theme of this work. I argue that by exploiting insight from both neobehaviorism and neopragmatism it is possible to parse a theory of intentional content that is far more illuminating than its internalist rivals. Apart from the naming conventions, the following exposition of neobehaviorism, and especially of linguistic pragmatism, borrows heavily from the above-cited Haugeland's (1990) article. Below, I will explain how I think these accounts are mutually supportive and dependent on each other's insights.

2.2.1 Neobehaviorism

As the name suggests, neobehaviorists claim that propositional attitude terms refer to the behavior of agents instead of their internal mental states. The difference from the psychological behaviorism is that this stance does not pertain to the empirical methodology or theory formation in psychology but is simply a thesis about the semantics of our everyday folk psychological par-

lance. Although this standpoint may take many forms,[33] the neobehaviorists generally agree that it is necessary to consider the internal cognitive processes of organisms in explaining the etiology of intentional behavior (which marks the difference in comparison to the hard-line psychological behaviorism). Accordingly, this position is compatible with computationalism, but the principal difference between neobehaviorism and representational cognitivism is their different accounts of the semantics of intentional terms. The point is not to deny the importance of internal processes in explaining thought and behavior but insist that they are irrelevant in our attributions of beliefs, desires, etc., to cognitive agents.

What neobehaviorists and cognitivists generally agree on is that we can effectively use the so-called folk or common sense psychology and its propositional attitude terms to describe, explain, and predict others' and our own behavior. The gist is that this strategy works exceptionally well even without any knowledge of the internal organization or workings of the human cognitive system. Indeed, we can use the same ascriptions to make sense of the behavior of widely different systems that definitely have different internal organizations. Daniel Dennett is a prominent proponent of this idea, and his favorite example is a chess machine (Dennett, 1971). Consider you are playing chess via a computer terminal against an opponent who plays decent game. You can make sense of your opponent's moves by postulating that the opponent is *trying* to win the game and *considers* dominating the middle of the board as a good strategy and therefore *decides* to get the queen out early, and so on. For this explanative and predictive stance to work, it does not matter whether you are playing against a competent human being or a competent computer program. What matters is that the opponent exhibits systematic behavior that is overall means–ends rational.

Further, even if you knew that the opponent was a machine and knew its program code and all the details of the physical mechanism that makes it run, the intentional strategy would still be the most effective way to make sense and predict the system's behavior, provided its program is complex enough

[33] Philosophical neobehaviorism is not hugely popular, so there are hardly many competing formulations. Daniel Dennet (e.g. 1987) is by far the most referred author of this view. Other notable theorists include Robert Stalnaker (1976) and, up to a point, Gilbert Ryle (1949). Despite its limited popularity, the instrumentalist account of intentionality inherent in this theory has been adopted by several authors (e.g. Pettit 1996 and Clark & Charlmers 1998). As a side note, B. F. Skinner had similar views about the status of folk psychological ascriptions (1965, Chapter 17), although he resides in the tradition of psychological behaviorism and supposedly would have disagreed with philosophical neobehaviorist on many counts.

to render the machine code explanation too laborious and that its behavior is sufficiently systematic and rational to make the intentional explanations effective. The point of this rather trivial observation is to invalidate the argument, put forward e.g. by Jerry Fodor (1987), that the success of commonsense psychology in explaining and predicting human behavior gives us at least a *prima facie* good reason to assume that propositional attitude ascriptions reflect the factual ontology of our cognitive system. The Fodorian argument states that we use propositional attitudes as hidden variables in causal explanations of others' behavior, and since those explanations generally work very well, then, in the spirit of scientific realism, we have a reason to assume that the intentional terms correspond to true causal variables in the etiology of human behavior. However, if it is true that we can apply commonsense psychology successfully to describe and predict the workings of chess machines, animals, and a wide variety of other agents, the argument seems to lose its edge.

The upshot is that if folk psychology does not discriminate different kinds of agents, then it does not make difference between different types of theories of human cognitive processing either. This is because if there are any stable and real phenomena that commonsense psychological ascriptions pick out they cannot be the elements of internal structure and processing. According to neobehaviorists, however, you can be intentional realist even though it turns out that nothing in the cognitive system corresponds to propositional attitudes because the referents turn out to be real dispositions or patterns in overt behavior. For example, we can say that "The cat believes that there is food in the box" and take that statement as literally true without committing to the view that there are representations in the cat's cognitive system that stands for the concepts of *food* or *box*. The corollary is that you do not have to be able to mentally represent the concept *food* in order to have wants or beliefs about it because beliefs and wants are manifest in overt action and not, at least principally, in thought. This does not mean to deny the existence of intentional thinking, as some radical old-school behaviorists would have it but to insist that acting as such is intentional, thinking is a kind of acting, and concerning to intentionality acting is ontologically more fundamental than thinking. Moreover when we try to decide whether our intentional descriptions are adequate or not, we do not inquire into hidden internal causes but try to figure out whether the overall behavior of the agent matches our intentional interpretation of its behavior.

Slight deliberation reveals that the neobehaviorist strategy of intentional explanation works a bit too well. Why not explain the rock's rolling down

the hill by its purported desire to reach the bottom? Perhaps rocks are not good candidates for things that exhibit intentional behavior, but programmable washing machines, for example , behave at least in some ways comparable to chess machines. Should we, therefore, conclude that the washing machine wants to get the laundry dry and considers spinning it furiously for a few minutes contributes toward that end? If that is absurd, why is it all right then to adopt the intentional explanation stance towards chess machines? Many of us find it hard to dislodge the intuition that while some complex non-mental systems may have a sort of as-if beliefs and desires, it is humans (and perhaps some other animals) who really have the right stuff. The neobehaviorists maintain that these intuitions are unwarranted because intentional explanations actually are instrumental regardless of the fact that they refer to actual patterns of behavior.[34]

It is an empirical question of whether an when specific folk psychological explanations work; however, when they do, they are concerned with real behavioral dispositions. Since there are no as-if behavioral dispositions, there are no as-if intentional attributions, either. It's just that in case of rock behavior you don't gain anything by postulating rock-beliefs and rock-desires because simpler and more effective explanatory strategy is just to say that rocks, like physical objects in general, tend to roll down the gravitation gradient if there is nothing blocking their way. In the case of chess programs and people or other entities that exhibit complex behavior, we gain the most effective explanatory and predictive framework by imposing on them intentional states rather than resorting to laws of physics. It is not a matter-of-fact issue about which descriptive strategy and terminology to use but a matter of explanatory and predictive economy.

So, in the realm of intentional explanation everything is permitted because nothing is, strictly speaking, true? Dennett has stressed that instrumentalism implies that there are good and bad intentional construals of behavior and the good ones are genuinely predictive. It is, after all, real patterns that they should pick out. Moreover, there is a certain logic intrinsic to intentional terms that the explanations need to respect. For example, you cannot explain someone's preference for celery stalks over ice cream by saying that he hates celery and loves ice cream. However, if you add the fact that he tries to eat healthily, the preference readily makes sense. On the other hand, if he often consumes large

[34] See Dennett (1987, Chapter 2). Note that theory instrumentalism is generally considered to be incompatible with realism. Dennett defends a sort of mid-ground position, which he calls "mild realism" (Dennett, 1991b).

quantities of bacon, the explanation becomes defective unless it is augmented with the hypothesis that he believes bacon that is bad for health. Regardless of whether the latter belief is warranted or not, it again renders the behavior rational and enables us to derive a further prediction that, all else being equal, he would prefer bacon over ice cream. There are two points here: (a) there is a specific logic intrinsic to terms like "belief" and "desire," and (b) the correct use of intentional descriptions presupposes that the target systems truly exhibit patterns of behavior that are systematic and sufficiently rational under that logic. According to Dennett (1991b), the target system can satisfy multiple intentional explanations; however, some of them may imply predictions that turn out to be systematically false. Hence, Dennettian instrumentalism is a sort of "mild realism" about propositional attitudes rather than rampart nihilism.

Fair enough, but how is it that we really seem to utilize internal propositional representations? I really can entertain a thought that it is raining and evaluate its implications; however, these mental episodes do not necessarily manifest themselves in my behavior. Moreover, if we want an accurate intentional description of others' behavior, is it not usually the most straightforward way to just ask them? They may not always tell the truth, of course, but at least they *know* the truth. I take that these rather obvious facts render the internalist accounts of intentionality so compelling. There are two distinct matters here: The first concerns the nature of thoughts and the other is about the privileged epistemic access to our own intentional states. Thus, the issue here is that while thinking probably is a kind of doing, you cannot evaluate all the mental acts on the basis of overt behavior. Therefore, if intentional phenomena are tied to manifest behavior, how can private thoughts be intentional? Here the story becomes a bit tricky and unavoidably "just-so". My preferred version is roughly what Daniel Dennett (1991a, Chapter 7) speculates about the development of language.

There are more refined accounts of linguistic development available but for the sake of parsimony we do not go into details. An interested reader might appreciate the reference that story bears relevant resemblance to the theory put forward by Michael Tomasello (2009) and his colleagues (Tomasello et al., 2005). See also (Fitch, 2010, Section 4) which presents a more broad outlook of themes discussed below, especially about the different varieties of the possible protolanguage. The *developmental* aspect of story that follows is not necessary; it is just illustrative and should be taken with a grain of salt. The point is to show how the system of communicative expressions can grow out from shared non-linguistic intentions that an agent wants to communicate

to another in situations where they both share some common practices, which make it possible for the recipient to understand what might be the point of the expression (see also Grice, 1957).

The story assumes that language initially developed from signaling practices that were tied to the social behavior of communication about things in the immediate environment. Certain deeds were made to pass warnings, requests, and threats, and to direct joint attention and to coordinate material exchange. These communicative practices could have comprised sounds or gestures, but they were anyway instruments of social behavior, not unlike what can be observed in present-day primates. Over time, these practices evolved into a protolanguage, which enabled communicative acts to be used to refer to things not immediately present.[35] Thus, utterances or gestures were used to pass and ask for information about things that were absent or not necessarily even materialized yet. A simple example of this type of practice is when I point to a direction and signal a particular sound and you respond selectively with a signal that is used in the presence of food when facing danger. This sort of behavior could ostensibly engender the representational function of utterances, which evolved from the primary coordinative function. Before this stage, the only normative element of language use was how successfully it was applied to the joint task at hand, but now another factor entered into linguistic practices, namely truthfulness, which became relevant since the new communicative function made it possible to misrepresent reality. It was now possible to lie and to be mistaken. The emergence of this sort of symbolic practices also enabled one to represent the intermediate consequences of actions before they were taken and, by induction, alternative chains of actions to reach distant goals. Thus, while the chains might have been built only from associative links, it became possible to practice some kind of reasoning and planning jointly by producing those chains in conversation.

Now, if you can talk with your comrades why not with your self? The social linguistic practices could have first transformed into private talking and then people just dropped the overt speech when they learned to carry out the discourse entirely in their minds, perhaps in the same way that we can mentally simulate overt actions. The neobehavioristic account of propositional or conceptual mental representation is precisely this. Speaking (or gesturing or

[35] This capacity is manifest in the gesturing of at least some great apes (Lyn et al., 2014). See e.g. (Jackendoff, 1999) and (Arbib, 2005) for arguments for vocal and gestural origins of protolanguage, respectively.

whatever) evolved into a system that enabled referring to things not necessarily present, and thinking is nothing but speaking internalized.

The complete story of the evolution of language is undoubtedly far more complex and presumably involves concurrent biological evolution. These details are not much relevant to our discussion, except perhaps the possibility of the commonly held idea that language came about as a means for expressing our thoughts. Not surprisingly that has been a more influential view among old-school cognitivists like Fodor and Noam Chomsky (Harman, 1975). Computationalists armed with two-factor CRS hold that we speak out our thoughts whose intentionality originates from within. For neobehaviorists intentionality is tied originally to behavior. This includes social practices, and when they extended to the communal use of symbolic language, it became possible mentalize intentionality. This neobehaviorist stance bears a strong resemblance to linguistic pragmatism. Before probing that in detail we have one loose end, still: How neobehaviorism fares with our purported subjective authority and access to our own intentional states?

The above story answers the question in case of linguistic thoughts but how it fare with propositional attitudes in general? First off, it is instructive to mind the difference between explicit commitments and dispositions such as beliefs and desires. Jonathan Cohen (1992) has offered an influential analysis of intentional terms in which he claims that beliefs and desires are in a sense are cast upon us. We cannot choose them, we often do not know where they came from, and we do not always even know that we have them until we find ourselves in a situation where they manifest themselves. He adds that another breed of intentional states is accepted commitments, which are explicitly adopted and affect our behavior as far as we deliberately enforce them. The difference seems to be very relevant to the analysis of propositional attitudes within the two-system framework (Frankish, 2004). The point is a blunt but easily overlooked fact about the proper use of propositional attitude ascriptions: Often we come to believe what we accept and sometimes we come to acknowledge our beliefs and desires; however, the explicit endorsement of our attitudes is not built into the logic of proper use of intentional vocabulary, and a deliberative control over them is actually absent as a norm.

To illustrate the difference, Cohen cites a lawyer who cannot help believing that her client is guilty and desiring that the client gets punished. Due to work ethic or whatever, the lawyer may accept the policy of treating the client as innocent and act accordingly. This is a conflict of intentions but not a violation of practical syllogism: The lawyer desires to win the case and be-

lieves that to do so, she should treat the client as innocent and hence presents the client as such. At glance, this kind of self-controlled behavior is problematic for the Dennettian account. In terms of explaining the lawyers behavior, ostensibly, the most parsimonious intentional explanation would be that the lawyer believes the client is innocent. However, as per the hypothesis, this would contradict the facts considering the lawyer's beliefs. But if we look at the lawyers behavior in the long run, we may discover that her behavior has been consistently professional in court cases. If that is the case, we better assume that the lawyer is simply doing what a good lawyer is expected to do and resort to the practical inference just explained; we restrict ourselves from hypothesizing about her beliefs about her client and explain her behavior by referring to her putative beliefs and desires about proper professional conduct. Moreover, if we share the same information that made the lawyer believe the client was guilty, we should expect that she also considers the client not innocent regardless of the lawyer's actions in the courtroom. This is because if we compellingly conclude that the client is guilty, then, under the rationality assumption built into intentional stance, we should expect the lawyer to reach the same conclusion accordingly.

The problem the neobehaviorists are facing boils down to the status of the putative *actual* beliefs and desires of the lawyer. The issue is that even if we do not have authority over our beliefs and desires, do they still reside inside us as states of mind that we can access and report or do our reports on our own propositional attitudes stem from taking an intentional stance towards ourselves? The latter would mean that when we introspect we actually treat ourselves as a subject of behavioral theorizing. The idea may appear odd but at least Wilfrid Sellars (1956) has proposed something along these lines. His idea was that intentional terminology is fundamentally behavioristic; however, after the practice of using it to describe others' behavior was been established, the next step was to extent it to explain our own behavior to others and to ourselves.

The point is that while reporting our propositional thoughts (expressed in internalized language) resembles reporting our propositional attitudes, the surface similarity is misleading. The latter is actually grounded in taking a theoretical stance toward our own behavioral inclinations in a vocabulary that makes sense to others. Interestingly, the facts support this position. The classic example is Richard Nisbett and Timothy Wilson's (1977) experiment. They arranged a fake consumer survey in a commercial establishment where they invited passer-bys to evaluate five pairs of stockings. They found a marked

preference for pairs positioned on the right-hand side of the table. When inquired about the reasons for their preferences, the subjects referred to various properties of the stockings, even though all the pairs were actually identical. All the subjects denied that the location had any effect on their preferences; yet, it did for reasons still unknown. Nevertheless, they were not lying or consciously making up reasons.

The phenomenon is called *confabulation* which (in this context) means sincere rationalization of one's actions. The subjects were confident that they were reporting veridical reasons for their behavior, which was actually caused by non-conscious processes that may elude intentional (in the sense of rational) explanation altogether—which likely is the case in the stockings experiment. No normative reason exists for preferring objects on the right-hand side of the table, although the effect is so evident that there must be a causal reason for it. Wilson's *Strangers to Ourselves* (2002) demonstrates how surprisingly meager our self-understanding truly is and how often we devise sincere and reasonable, albeit non-veridical, explanations of our behavior. Although none of this refutes the reality of so-called internal mental states, not even propositional ones, our knowledge of the psychology of intentional attributions attests to the correctness of Sellars idea.

2.2.2 Neopragmatism

Neopragmatists aim to derive conceptual contents not from the individual mental states or patterns of behavior but from the cultural practices that constitute the way of life of linguistics communities. They maintain that conceptual meaning is produced or determined by using words in a familiar manner. This can be specified in many ways and the version that is relevant here is the one further claims that meaning is normative in essentially social sense. Along this line of thinking, the original intentionality is often seen as attached to tools, performances, and other shared objects and practices. The idea is that fundamentally norms are not explicit rules or conventions but shared communal practices of doing things in specific ways. Norms are established by tracking what others do and how they do it, conforming to those practices, and penalizing deviant behavior. This sort of conformism is arguably necessary in prelinguistic social coordination, since it is hard to do things together if we have different expectations about how to accomplish joint goals and cannot communicate our intent or negotiate our methods.[36]

[36] See also Tomasello et al. (2005); Tomasello (2009).

Neopragmatism is a multifaceted postmodern version of classical pragmatism that emphasizes linguistic practices in determining intentional content. The term is often associated with the philosophy of Richard Rorty and has been influenced by continental philosophy, especially Martin Heidegger and Jacques Derrida, and thinkers from the analytical tradition, such as Willard van Orman Quine, Donald Davidson, Sellars, and Wittgenstein; it has also been influenced by early American pragmatism, especially the works of John Dewey.[37] The neopragmatism discussed here inclines toward the analytical facet, and my exposition is based on the versions formulated by John Haugeland (1990) and Robert Brandom (1994; 2000).

Humans tend to form social groups. The potential behavioral variety of human is higher than other animals; however, as social creatures, we also have tendency to suppress this variety by imitating others and adopting their habits. Thus, people tend to adapt to mainstream practices and, moreover, we persuade others to conform by peer pressure or sometimes force. This results in a dynamic coupling between the group and its individual members, which tends to level the initial multitude of behavioral dispositions and forge a group behavior with relative homogeneity. Group behavior comprises of a more or less consolidated set of norms that shape society's culture and way of life as a whole. Here "norm" does not mean moral code only but whatever behavior deemed normal in relevant circumstances. In this reading of "norm," non-normative behavior is deviant by definition but not necessarily wrong. Sometimes breaking the norm exhibits criminal, lunatic, or heretic behavior, but sometimes an innovative act of a reformer or genius. Often, it is mostly trivial. Why do certain behavioral dispositions end up as the norms of society while others do not need not concern us here. The reasons are several, including a mere chance. According to the neopragmatists, it is essential that the norm-welding process does not initially require explicit thinking or discourse. The process is solely the result of our innate mode of group behavior, more like flocking or herding than negotiating.[38] In fact, to neopragmatists, there is no language or concepts to negotiate and reason with befor community's norms have consolidated, because they are the source of all meaning.

According to Robert Brandom (1994, Chapters 1 & 2) the foundational idea of linguistic pragmatism rests on the tenet of Immanuel Kant that all

[37] See Haugeland (1990) and introduction in Rorty (1991). Note that it is the later works of Wittgenstein that are relevant here, especially Wittgenstein (1953).

[38] Or, as Haugeland puts is, "When behavioral dispositions aggregate under the force of conformism, it isn't herds that coalesce, but *norms*." (Haugeland, 1990, 405).

conceptual activity necessarily has a normative character. In perception we take something to be true, and in action we make something to be true. What makes these events intentional, in contrast to merely causal, is that there are reasons to take or make something true, and the concept of reason is inherently normative: things can be done for wrong reasons while they cannot happen for wrong causes. Often norms are understood as explicit rules that express how things ought to be done; however, neopragmatists conception of norms also cover socially enforced regularities that are manifest only in action. They claim that the latter type of norms (i.e. largely tacit forms of social praxis) are actually more fundamental and a necessarily precondition for explicit normative rules.

This necessary pragmatist factor relies on the Wittgenstein's observation according to which the act of *following* an explicit rule also has a normative dimension: In principle, there are conditions where rules can be applied incorrectly, and to explicate these conditions, we need further rules and, to cut the resulting regression finite, the chain of norms must be eventually grounded in practical normativity implicit in doing things in particular ways. The point is not that there must be practices that cannot be described in language but that some non- or pre-linguistic standard practices are necessary for the system of rules to get off the ground and make explicit norms intelligible (Brandom, 1994, 62).[39] Finally, the necessary social dimension sets in because mere regularity in behavior is not normative enough: for a performative regularity to be normative in the absence of explicit rules, there must be practices of social sanctioning that enforce certain actions (Brandom, 1994, 34–35). This last clause has, again, a Wittgensteinian flavor, because it is reminiscent of the famous (impossibility of a) *private language argument*[40] in that the idea of private norms is fundamentally incoherent. If we abandon the idea of transcendental, norms then the only remaining viable naturalistic account is that incorrect behaviors are eventually those that get sanctioned by our peers.

The story so far may sound identical to neobehaviorism with a caveat, in that it is not patterns of individual behavior that engenders intentional content but patterns of communal behavioral regularities. In other words, the critical difference is that neopragmatists ground meaning in successful conformism

[39] Note that the argument bears a close similarity to the one introduced in the famous "What the Tortoise Said to Achilles" by Lewis Carroll (1895): While you can always substitute inference rules for axioms, you can't introduce all the rules within the system. Some principles must be grounded as uninterpreted procedures. See Brandom (1994) onwards from page 22 for discussion.

[40] See Wittgenstein (1953) from §253 onwards.

rather than means-ends rationality to fulfill the individual's goals. This observation is not entirely accurate, however. Neopragmatism takes norms to be necessary for demarcating the causal realm from intentional because the latter is constituted by reasons. According to Brandom (2000, 106–110) it is essential to note that a mere selective responsivity to the environment is not enough to render behavior intentional; otherwise, we should accept that iron has the concept of oxygen because it rusts as a response to an exposure. In the preceding section, we dismissed this problem by treating intentional ascriptions as instrumental; however, neopragmatists maintain that a fundamental distinction can be made between intentional and non-intentional behavior.

Norms are essential to intentionality because they establish reasons for acts, claims, and conclusions. In order to do or say X intentionally, one has to have a reason for doing so, and the norms regulate what are good and what are bad reasons for X. This implies a kind of inferentialism that links premises, which work *as* reasons, to conclusions that are established *for* reasons. Here, inferences need not have anything to do with formal logic. In the intentional discourse, they are often codified as practical inferences where conclusions are usually actions and where perceptions can be considered a special case of premises (Brandom, 1994, 233-234). This is what establishes the logic of intentional vocabulary discussed in the previous section. To exhibit true intentionality, we need to be able to play the game of giving and asking for reasons. This means that we need to understand and be able to explicate the purported rationale of our actions. This parallels the neobehaviorist account of internal intentional states explained earlier.[41] The difference is that, according to neopragmatists, it is this ability to exploit the logic of established intentional vocabulary to make sense of the reasons behind actions that renders behavior conceptually structured. Indeed, neobehaviorists owe us an explanation where the logic of intentional terms comes from, which enables us to take intentional stance in the first place. According to neopragmatists it comes from the social fabric of reasons that determines what cultural niches various contentual tokens, such as ritual objects, customary performances, and tools, occupy (Haugeland, 1990, 404). However, conforming to social norms as such cannot exhibit full-blown intentionality since the neopragmatist position assumes that such conformism is initially operative without conceptually structured intentionality (i.e. prior to language), which would require the capacity

[41] And, in fact, Wilfrid Sellars is perhaps more clearly in the linguistic pragmatism camp than with the neobehaviorists for he holds similar views.

to articulate reasons and hence language. So, how does conceptual content emerge from social praxis then?

John Haugeland (1990) explains the idea of how practical norms relate to linguistic meaning by an analogy between words and tools. Consider tools like a saw or a hammer. They clearly are not symbolic representations and, as objects, they do not refer to anything. However, *as tools*, they are *for* something. Hammers are for driving nails and saws are for cutting wood. Moreover, there are correct and incorrect or intended and non-intended ways of using them. Hence, tools have something like proto-semantic properties that are derived from their intended or correct use, and, as tools, they in a sense point toward their purpose not unlike words point toward their meanings. Moreover the practices of nailing and sawing serve as means to other ends, such as constructing buildings. Thus, there is a close analogy between linguistic expressions and carpentry tools, for example. Words and sentences are *about* something whereas tools and practices are *for* something. Where the former are embedded in social practices, which constitute language (understood not as an abstract system but a practice of social coordination), the latter are embedded in an interlocking web of paraphernalia and practices, which constitute the art of carpentry.

The analogy goes one crucial step deeper, still: The major philosophical problem with the naturalization of intentionality is to explain how intentional phenomena can satisfy both semantic and causal properties simultaneously. Tools have this character, since an essential properly of being a tool is to take part in causal nexus of hunting, constructing, or whatever kind of work, and tools also retain their normative character of what they are for even in the absence of their intended practical use. Similarly, linguistic tokens are intertwined in causal acts of communication but retain their meanings independent of occurrences of their use and in the absence of their referents. In both instances this intended purpose or meaning is ontologically dependent on the underlying culture in which the tools or the words are put to use. Outside the actual specific use contexts they carry no interesting causal or semantic properties.[42]

Linguistic utterances, though, have a property that tools lack: they are tokens that stand for other things; that is, they have a symbolic function.

[42] For further discussion (outside strictly neobehaviorist tradition) the reader is advised to consult Enfield (2015) on treating language as both a medium for and a result of causal social interactions; the parallel to tools is further elaborated in Sections 4 and 5 and in the first few pages of Wittgenstein (1953).

However, here ”symbolic function” does mean an abstract reference relation but a context bound causal factor in social exchange. Words symbolize through their use in actual material social contexts. The point of neopragmatist use theory of language is not just that in order to make sense to others we need to follow some rules of language, but also that norm bound communal practices precede and establish all intentional content. Accordingly, concepts form a kind of inferential network where their contents are partly determined by how they relate to each other in reasonable intentional explanations. For example, explaining why one is cutting down trees by referring to one's intention to visit his relatives would presumably make perfect sense without further elaboration in a society where families are dispersed on several islands and it is a common practice to build boats. Specific ways of explicating our intent make sense while others do not, and our everyday communication requires figuring out a vast amount of relevant semantic and practical connections rooted in our daily practices. This essentially knits intentional content and practical reasoning.

Furthermore, this makes linguistic tokens function similar to tokens of mental representations in conceptual role semantics; that is, their content is determined by their role in the system of tracking reasons. The key difference here is that for neopragmatist, the ontologically fundamental domain of concepts is their use in communicating intent in social praxis rather than their use in mental processing as the proponents of CRS would have it. Thus, while both theories are varieties of inferentialism, the original medium is external and communal to neopragmatists whereas internal and individual to the advocates of CRS.

I take that these communal regularities play the role of ”settled policies” that in turn translate into ”intended purposes” that Fodor tried to smuggle into his theory to make sense of misrepresentation.[43] What we have finally found here are the grounds for demarcating correct and incorrect applications of concepts without invoking analytic/synthetic distinction or other notion of conceptual truth. There is a price to be paid, however. One of the main problems with internalist accounts was to explain how it is possible that different cognitive agents can have the same propositional attitudes and commensurable conceptual systems. In the case of CRS, the problem lapsed into determining *the* correct inferential network. If you go along with linguistic pragmatism, these problems disappear because of conceptual relativism built into the theory. The correctness in the use of concepts is dependent solely on how they are understood under the social conventions that determine what is locally rational.

[43] See the quotation on page 43.

Local rationality here means that while local norms are not necessarily arbitrary, their acceptance is fundamentally just a question of how they play out in social discursive practices. When we learn these practices, we simply learn to explicate reasons according to principles that are sanctioned by our peers. To neopragmatists, conceptual domains are constituted by human practices where particular objects and acts are *intentionally* bound together by their perceived normative statuses. This means that conceptual contents and domains are strictly dependent on the idiosyncrasies of the relevant community.[44] What is taken as obvious or (un)reasonable is likewise locally determined.

If you have postmodernist (or perhaps Quinean) inclinations, you may find this aspect of the theory appealing; however, I think many would find it hard to swallow that meaning and rationality are just a matter of communal inclinations, especially in the fields of science and formal disciplines. In any case, this view departs radically from the idea that reasoning is grounded in intellectual intuition of universal conceptual or metaphysical essences. Here, the concept of "norm" is philosophically very thin as it is. Ultimately, it is taken to mean just social expectations; basically what people with certain statuses are expected and allowed to do in certain situations, which is often far from the notion of ideal. Moreover, social expectations are supposed to emanate from learning social regularities, which can often violate explicit normative standards. But note that this "often" comes about because we usually notice the normative character of behavior only when our expectations are violated. Much of our everyday life is normatively structured but this eludes our awareness because social norms are mostly implicit and often quite trivial. What needs to be shown is that non-trivial, non-social, explicit, and often abstract norms, such as moral ideal, and those that govern formal rationality (both often violated in everyday behavior) are fundamentally the same ontological species as the more trivial social expectations. We will get back to this in chapter 5.

But how credible are the principles of neopragmatism? Inferentialism as such is too commonplace to require a dedicated discussion here; however, the idea that the relevant inferential relations are local discursive regularities is certainly far more controversial. I do not intend to delve into the philosophical merits and problems of this account. I will, however, question the claim that

[44] But note that certainly plain practical matters can engender normative statuses and often practical and other factors are intertwined. The use of clothing and dietary customs are excellent cases in point. Moreover practical rationality is supposedly constrained by our biological makeup, and therefore this account does not in principle rule out the possibility of more or less universal human practices and hence universal norms.

linguistic pragmatism can be the whole story of human intentionality. It will play an important role later when we discuss abstract concept learning. Right now I want to address the conjectural mechanism tracking and enforcing the relevant social regularities, which bears rather strong empirical commitments considering human social cognition. Another empirically risky claim is the comparison of tool and language use. Both planks are somewhat difficult to assess empirically; however, our scant evidence supports these ideas. Let us start with remarks on the connection of language and technological praxis.

Premotor cortex F5 in monkeys contains a *mirror neuron system* for grasping, which has common activation patterns for executed and observed manual actions. Area F5 is homologous to human Broca's area in left cerebral hemisphere, which is generally considered to be a speech area, or, more specifically, to execute tasks associated with production and syntactic parsing of language. Moreover, the monkey area F5 contains *canonical neurons* that fire when a monkey sees a graspable object. (Arbib, 2005) Recent evidence shows that in addition to speaking, Broca's area is also active in signing and when humans both execute and observe grasp. It seems to play a crucial role in action recognition and production, and in interpreting actions of others in term of goals. (Fadiga & Craighero, 2006; Fadiga et al., 2006) This strongly suggests that cognitive mechanisms responsible for social understanding and language use are closely related to manual praxis. Broca's area seems to be involved with a more general function than previously thought, which is not solely a linguistic capacity but a general function to extract action meanings by interpreting sequences (e.g. motor or phonemic) in terms of goals (Fadiga et al., 2006, 87). Naturally, such capacity is essential for dealing with meanings of actions that have sequential and hierarchical structure—which are hallmarks of both language and tool use.

There is also evidence that links archaeological records of increasingly complex tool making to progressively complex neural structures, which overlap language processing. Apparently, in stone tool making (and in many modern manual tasks), the division of labor of dominant hand performing small-scale, rapid processing, and the inferior hand providing stable postural support for the worked object is mirrored in the development of more general functional lateralization in the brain. Ancient Oldowan stone tool manufacture require less of this type of hand specialization, while newer Acheulean tools require more asymmetric specialization of contralateral hand coordination. These tasks recruit neural resources that respectively overlap with neurons involved in the word- and sentence-level processing in the left hemisphere and discourse level

processing in the homologous areas in the right. This suggests that there may be an important evolutionary link between the development of language and cognitive capacities underlying tool use and manufacture. (Faisal et al., 2010; Uomini & Meyer, 2013) Broadly comparable conclusions have been drawn by studying the teaching of Oldowan stone tool-making techniques by mere imitation and teaching with gestural or verbal support. The results indicate that verbal teaching has significant benefit over gestural teaching which is, in turn, superior to mere imitation. Hence, apparently the reliance on stone tool making generated selection for teaching and especially for verbal tutoring, thus engendering the gradual evolution of language tied to social and technological praxis. (Morgan et al., 2015)

The idea that tools, gesture, and language share a common evolutionary history and, therefore, common cognitive resources is not new and far from a settled matter. Nonetheless, it is becoming clear that the old discoveries of neural correlates of linguistic processing have actually found neural resources for more general capacities, which are involved in organizing and interpreting complex intentional action, which transcend mere linguistic tasks and are shared with technological and social praxis. This, and related recent research,[45] are important to our inquiry because these considerations highlight that regardless of the conceptual and factual disanalogies between language and tool use, their psychological connections and similarities as hierarchically and combinatorially organized causal domains are not superficial. Therefore it is empirically motivated, and not a mere philosophical speculation, to theorize about actual language use as ontologically similar to technological and social praxis, thus empirically supporting the neopragmatist conception of language.

The other open issue was the assumption that we exhibit a specific type of social conformism. It is, of course well, known that children are prone to imitate others and that most people tend, often unconsciously, to prefer individuals who are similar to themselves (Jones et al., 2004). However, the neopragmatist claim of conformism is far stronger since it assumes that we extract norms by tracking what our peers and authorities do and that we also tend to enforce these practices to others. Moreover, this disposition should be an instrument for social cohesion and coordinating collective behavior in complex tasks, and therefore it should interact with both perceived group affiliation and instrumental rationality in a social setting. It seems that for a long time imitation research has somewhat missed the social and normative aspect, and

[45] In addition to the above mentioned research the reader will find a brief and accessible summary of relevant discussion in Stout & Chaminade (2012).

therefore much of these phenomena have remained undisclosed; however, recent evidence reveals essentially this type of pattern (Over & Carpenter, 2012).

Children often readily learn what is relevant for achieving instrumental goals; however, they also pick up and replicate irrelevant aspects of others' actions, leading to over-imitation. There is ostensibly some inconsistency in this behavior, but at least four variables modulate it: (a) causal opacity of procedures, (b) instrumental versus ritual (or conventional) context, (c) peer pressure, and (d) a need to identify (or to "affirm a shared state"; Over & Carpenter 2012, 185) with a model. Thus, children faithfully replicate futile details if they are unsure what aspects of behavior are causally relevant or what the purported goal is supposed to be; but they may also do this knowingly—even when alone—if they want to identify with the model. Both the presence of the model and observing several peers participating in causally irrelevant behavior facilitate over-imitation. Hence, this sort of learning is often termed "ritualistic" instead of imitative. This tendency apparently serves many functions, such as learning causally opaque skills, learning cultural conventions, forging social cohesion, and affirming one's position in the relevant group (Over & Carpenter, 2012; Keupp et al., 2015; Legare et al., 2015). The underlying mechanism involved in all these factors seems to be that of tracking what perceived in-group members do and how they do it.

Of particular importance is the normative character of ritualistic learning precisely in the sense of normativity employed by neopragmatists. Children spontaneously enforce learned practices on others and protest against violations. It seems that they pick up regularities in ways of doing things as socially normative regardless of their causal relevance. The enforcement of norms is stronger if the context is considered conventional rather than instrumental and also if others are deemed to be in- rather than out-group members (Over & Carpenter 2012, 187; Keupp et al. 2015).

Abstaining from norm enforcement in instrumental contexts may appear contradictory to neopragmatism; however, this is not the case. If the fundamental function of conformism is to coordinate social praxis in the absence of explicit knowledge about how to achieve shared goals, then we should, somewhat paradoxically, expect that irrelevant regularities are enforced more strongly in a conventional context where participants are unable to explicate any reason for conforming to shared standard practices. In conventional or ritual contexts, the specific ways of doing things are the ends themselves. However, in instrumental contexts, what matters is that things are accomplished and any means–ends rational way is, in fact, normative if the procedure is laid out

such that it makes sense to the peers. Showing that the goal can be achieved in novel ways may be an essential part of making sense of one's idiosyncratic actions because it makes salient what parts of action sequences actually are relevant. In this sense instrumental rationality is multiply realizable in a way that conventional normativity is not. Therefore norm enforcement may be suppressed. On the other hand, perceived practical rationality is also constrained by (implicit) social norms underlying our understanding of what means and ends make sense. This is why in instrumental contexts, we still should expect people to sometimes protest the perceived violations of means–ends rationality even if their own interests are not directly involved and even if ends are met. This is because causally opaque procedures may be perceived as failing to behave in (locally) rational ways, thus constituting a norm violation—given that the procedure is not carried out by the model; i.e. authority or in-group majority. That is somewhat curious because it means that people may get irritated by completely effective behavior simply because they have learned to do things differently; however, this is precisely the pattern that has been observed (Kenward, 2012, 203).

Although many variables modulate the fidelity of imitation systematically, children are often incapable of explicating why irrelevant actions should or should not be performed. While children from two to three years of age can often distinguish morally bad from mere violations of conventions, they tend to encode observed actions as normative without encoding explicit reasons for this. (Kenward, 2012, 205) This kind of naivety is philosophically significant for the reductive explanation of (socially) normative intentionality because one only needs to be able to observe what others do and follow suit. One need not be able to understand why certain acts and choices are made. Rather, this conformism lays the foundations of social expectations and, hence, norms and eventually understanding reasons.[46] Essentially, it explains normativity through social normality. Relevant research discussed here is mainly involved children. It may be reasonable to hypothesize that these effects diminish when children mature and develop better conceptual tools to rationalize what is normative and why and gradually drop imitation and other ritualistic behavior; however, this is not the case. Implicit norm extraction and enforcement actually increases with age and remains with adults. I quote Ben Kenward (2012, 205):

[46] Although the imitative behavior requires goal directed intentionality in (social) action, see the next section and Tomasello et al. (2005) & Tomasello (2009) on etiology of norms, collaboration, and language.

> First, in over-imitation studies, there is no clear evidence available to the children as to why the action is performed. Second, in one previous study, using a very simple apparatus, the majority of over-imitating children, when asked why they would perform the unnecessary action, were unable to give any sort of coherent answer, although most could explain that a necessary but otherwise equivalent action was causally necessary (Kenward et al., 2011). These observations demonstrate that children are capable of encoding observed behavior as prescriptively normative without exposure to clear information as to why it is normative, or even as to within which domain the normativity of the behavior is determined, and without having formed an expressible belief about the reasons for the action's normativity.
>
> Therefore, I suggest that children may sometimes encode an observed action as normative without engaging in any reasoning justifying the action's normativity and without even believing the action's normativity is determined within any specific domain such as social convention or instrumental rationality. This suggestion is in line with evidence from adults, who are capable of holding views about the normative status of a behavior without being able to give coherent explanations for why the behavior should be proscribed or prescribed (Haidt, 2001; Hauser et al., 2007). The cognitive processes that produce such views are intuitive and not directly available to introspection, and norms can be acquired unconsciously as a result of observing others follow them (Cialdini, 2007; Haidt & Bjorklund, 2007; Sripada & Stich, 2007).

I believe that this is more than enough to conclude our discussion about the empirical credibility of the psychological nature of implicit norms as employed by neopragmatists. What remains to be shown is that practical social norms in this precise sense are constitutive of conceptual content. As I explained earlier, I will rather take that as a starting point later when discussing abstract concept learning and try to show how that idea can be utilized to yield a fertile theory of intuitive conceptual understanding, pertaining especially to theoretical concepts understood as a set of learned social practices.

2.3 The constitutive role of non-linguistic behavior

Whether or not there is a uniform class of psychological phenomena that can be labeled as "concepts", I'm confident that they are primarily not instruments of thinking but doing. With concepts, we can make generalizations which allow us to encounter the world as a multitude of ordered phenomena so we need not consider every thing and every situation as novel and singular without connection to other phenomena and past encounters. This, of course, is what makes goal-directed rational behavior possible in the first place. Therefore, concepts mediate the relation between sensing and acting, and it is natural to assume that the information that conceptual mental representations carry are primarily affordances and related functional information. In other words, the cognitive system does not just represent what *is* out there but primarily what *things do* and especially what *can be done* to or with them.

This point has been too often overlooked in the traditional philosophy of mind, and as a result there have been considerable problems in accounts of the content of mental representation. To put it bluntly, it is no wonder that there have been critical problems in attempts to determine mental content when the theories tend to leave half of the content out of consideration—indeed a very important half since the whole point of "what is it" part of mental content is to serve the "what can I do about it" part. There is no point in intentional content if it cannot be used to direct behavior. This, however, does not rule out that some of our concepts may be purely "intellectual" or "discursive". This is trivial since thinking and public reasoning are kinds of behavior. The non-trivial claim that I will pursue mostly in Chapter 5 is that the psychological content of even highly abstract concepts is at least partly constituted by functional knowledge about the actual use of these concepts in concrete situations, as per the principles of linguistic pragmatism.

A serious issue with linguistic pragmatism, which I take to be fatal for the orthodox version of the theory, is that it tacitly assumes—but ends up denying—the existence of non-linguistic intentionality. I find it rather evident that some non-linguistic intentions necessarily underlie the capacity to enforce and engender prelinguistic practices that institute norms in the first place. Brandom's philosophy carries an intuitively compelling implication that mere responsive mechanisms do not count as intentional agents; however, it is far more harder to swallow that non-human animals and prelinguistic infants do not count, either. Brandom certainly acknowledges this and also that infants and animals may be thought of as having *something like* beliefs that guide

their action (1994, 153-156). However, concerning non-linguistic creatures, he thinks that intentional ascriptions are merely derivative (p. 142). He adds that we make sense of the behavior of such organisms (and perhaps other systems) by applying propositional attitudes that, strictly speaking, apply only to linguistic agents. This is because to have propositional content, which according to Brandom is the only form of genuine conceptual content, one must have an ability to participate in the practice of giving and asking for reasons which, in turn, necessitates language. In his account, this marks the very important distinction between genuinely sapient and merely sentient creatures. For Brandom, plain perceptions and actions are not normative if they are not framed with discursive concepts, and therefore the behavior of organisms that lack language belongs to the realm of causality rather than of intentionality (Brandom, 1994, 233-234).

It seems intuitive (to me, at least) that creatures that exhibit selective responsiveness to their environment with accordance of their needs, pains, perceptions, and expectations have exhibit some form of intentionality even if they are incapable of doing deliberative judgments and decisions. This seems particularly true for organisms that ostensibly express primitive non-linguistic conformism by learning to keep their behavior in order. But, after all, it is philosophy's task to check these intuitions when they are unfounded. However, like e.g. Carl Sachs (2014, 69) I find that Brandom is selling the right account of linguistic intentionality, but there is the problematic move from this to a less credible linguistic account of intentionality. Nothing in Brandom's theory of linguistic meaning seems to imply that a notion of non-linguistic and non-propositional intentional content is incoherent. His theory of conceptually structured observation and action seems to entirely dismiss the relevance of other cognitive capacities, which makes it possible for organisms to respond to their environments differentially, systematically, and means–end-rationally with regards to their needs and capacities.

Brandom makes it a point that if we accept any sort of intuitive/discursive-concept distinction—where intuitions carry non-discursive content—we will face problems with normative restrictions on intuitive concepts and end up basically in the same muddle than with two-factor CRS: We either need to resort to supposedly given semantics facts or else we cannot explain what it means to apply concepts incorrectly. Therefore he rather rejects the need to offer an account on how intuition guides the use of concepts; he refuses to discuss

differential responsiveness in other terms than physiological; and, therefore, conflates intentionality with discursive reason rather than with agency.[47]

The most pressing issue here is not the intentionality of animals and infants (or the lack of it) but that if we take even rudimentary formulations of two-process theories seriously, pure linguistic pragmatism tends to render most of human behavior not conceptually structured and hence non-intentional, which is absurd. This is because most of our daily activities and expert tasks are executed either completely or at least partially as a tacit process that is generally supposed to be pragmatic, action-oriented, and independent of language.[48] More to the point, putative System 1 processes are inaccessible, and hence often we do not have any explicit understanding on what the supposed rationale our intuitive judgments manifest. As noted earlier, in some reasoning tasks especially non-experts often offer irrelevant, nonsensical, and even contradictory assessment about their presumed decision-making processes (Wason & Evans, 1975; Evans & Wason, 1976; Evans & Over, 1996). Experts obviously apply conceptual knowledge but are often incapable of articulating clear reasons for their decisions—or at least *better* reasons than advanced beginners—especially with complex problems.[49] Thus, even though people are capable of rationalizing and conceptualizing their behavior (either through confabulation or reflection), this capacity is often irrelevant to the behavior exhibited. Hence, it is an empirical fact that although concepts and knowledge guide perception and action, the associated cognitive processes are often not guided by discursive knowledge. In other words, the reasons we give in the practice of giving and asking for reasons are sometimes different from those that actually produces our behavior, and even if they often do conflate, it is still possible that the cognitive mechanisms driving the two (i.e., implicit production and explicit comprehension of behavior) generally operate independently.[50]

Of vital significance is the assumption that intuitive System 1 processes carry out the very same tasks as reflective System 2 because, considering the

[47] For more elaborate discussion see Sachs (2014, 72–82) and Pendlebury (1998).

[48] A proviso should be added outright. Much of what I intend to say later depends on the claim that some intuitive processes are linguistic. This does not affect the issue at hand, however. Many of my arguments also depend on the claim that very little intuitive processing is linguistic. This means that language is not a sound criterion for demarcating System 1 from System 2 processes.

[49] This observation dates back to the famous research on chess masters by Adriaan de Groot (1965). The phenomenon is discussed at length in connection with other tasks and domains, for example, in Dreyfus & Dreyfus (1986) and Klein (1998). See also Kahneman & Klein (2009).

[50] For further details, see the discussion in Section 3.1.

paragraph above, it would be borderline absurd to count behavior intentional only if is a product of System 2 deliberative processes. The upshot is that there has to be a *vast* intentional middle ground between discursive knowledge and mere causal differential responsiveness. But how can we analyze this kind of intentionality which perhaps eludes analysis into propositional constituents? It is worth noting how neobehaviorists shift the focus of intentional interpretation from articulated reasons to behavioral competence and hence substitute central elements of propositional representations with instrumental ones. This means to focus on goals instead of referents, success instead of truth, and means and ends instead of evidence and conclusion (Haugeland, 1990, 398). Brandom says that non-linguistic systems can only have "a primitive kind of practical taking *of* something *as something*" (Brandom, 1994, 33–34). However, to me it is clear that this "primitive" capacity can go a very long way in what comes to conceptually structured behavior.

I am not sure why the aforementioned capacity should be normatively determined especially in social sense. A good enough reason for taking particular action is to fulfill one's goals, even if the agent is incapable of explicating its rationales. Misapplication of discursive concepts can be thought as a special case of behavioral failure where the feedback that something went wrong comes from the social environment rather than the material one. You can't pass a course in logic if your behavior violates the norms of valid inference. Shifting the focus from misrepresentation to pragmatic failure blurs the line between the causal and intentional (in the Brandomian sense); however, the other option seems to be denying non-linguistic and non-human "primitive" intentionality entirely. I take that that would also mean that non-human organisms do not have goals or intentions, properly speaking. I guess one philosopher's *ponens* is another's *tollens*, basing my argument on this wisdom, I would rather go for the former option. What ever demerits this choice may have, I am convinced that they are balanced out by the advantages it offers in understanding how conceptual cognition works, including discursive reasoning.

Taking discursive misapplication of concepts as a special case of pragmatic failure also blurs the line between cultural and material: Social praxis is just a special case of material praxis. Learning discursive practices is based on tracking actual material causal events in the social domain. This position is generally endorsed in neopragmatism, and it precisely gives its status as a naturalistic theory of intentionality. The implication for cognitive theory is that even if humans have dedicated cognitive capacities to language and social cognition, more general causal learning should be involved in learning

and comprehension of discursive practices. This was, after all, the point in the above comparison between social and technological praxis. Often the core idea of neopragmatism is considered to be that everything contentful is normative all the way down; however, I am more interested in the aspect that conceptual understanding is based on causal induction, all the way up to abstract concepts. In short, causal cognition is the key to semantic cognition.

Given the discussion above, I propose that concepts or intentional mental capacities are primarily devices for systematic behavior and only secondary instruments of thinking and public reasoning. Thus, contents associated with object and event concepts also contain information about the environment and object affordances, possibly in the form of motor (or more abstract action) schemata and causal knowledge. Moreover, representational and pragmatic aspects of mental contents are frequently inseparable. Perhaps content determination in non-propositional mental representation must ultimately be fixed by terms of behavioral control in the perception–action loop because it is the action control for which mental contents are principally for. However, for my purposes it is immaterial whether this is strictly necessary. The important point is that to understand what content mental states carry, we must find out what behavior they bring about and how the world discloses itself to the active organism. This is basically an obscure (i.e. philosophical) way of saying that we need empirical research on what the organism tracks in its environment, and why and how it employs the conceptual resources so produced. If this is correct, the quest for abstract reference relation is generally inadequate for forging the foundations of contents of mental representation and intentionality in general.

Also, is that if propositional attitude ascriptions presuppose phenomenological or commonsense theoretical understanding of how we grasp the world (the ”propositional” part) and our psychological needs, capacities, dispositions, and so on (the ”attitude” part), it follows that folk psychological descriptions capture the actual essence of our own intentional contents almost by definition—since that is where propositional attitude terms receive their meaning; however, we can apply these ascriptions only derivatively and instrumentally to other lifeforms and systems. As we saw, this is what orthodox neopragmatists claim. The reason is that our interpretation of these descriptions depends on our conceptual system, and full appreciation of our conceptual system includes, *inter alia*, the discursive and inferential knowledge that it contains. And if there is no strict demarcation of discursive and other pragmatic knowledge, human concepts that are contaminated with discursive

contents (which is perhaps most of our concepts) cannot be projected on non-discursive conceptual systems without residue. But given this, it is essential to note that then these ascriptions apply only partly also to other people, especially from other cultures. Their pragmatic and discursive knowledge may be different from ours because of their different environmental demands and ensuing experiences. In the case of language users, this parallels the thesis of the indeterminacy of translation,[51] although its impact is somewhat different. This is because most of our conceptual content—as I will show later—is partly constituted by non-discursive pragmatic knowledge.

Therefore, like any discursive concepts, intentional ascriptions refer to the kind of intentionality or intentional content that we share with similar agents with similar worlds (or practical realities or *Lebenswelt* or whatever). In other words, our understanding of intentional ascriptions are dependent on our pre-theoretical understanding, gained through personal experience, of how we cope with our world. The extent to which we share with other people or agents the same capacities, needs, goals, demands, social institutions, and other practical frames that constitute our life-world approximately determines the extent to which our conceptual contents coincide. Moreover, discursive reasoning is a practice in its own right, and it may well be a powerful tool in widening the scope of our conceptual understanding. Mutual reasoning is a way to integrate conceptual systems (at least to a degree), and hence explicit discourse relieves us from being a hostage to our culture and autobiography, even if we cannot wholly share our different life experiences. These remarks about cross-cultural understanding should also pertain to subcultures, expert communities, etc., in one's *own* social environment. The implication is that intentional content depends on several factors: our biological and cognitive makeup, our experiences, and our material and social environment.

The reader may find that the above paragraph contradicts my earlier statement, that often we do not know the rationale behind our own behavior. Hence, a point of clarification is perhaps in order: What I'm claiming here is that the neobehaviorist story about propositional attitude ascriptions is, by and large, correct. However, language as such is not semantically transparent and folk psychological language (or perhaps all language) is too crude an instrument to accurately describe all the psychological, social, and environmental aspects behind the etiology of our behavior. Neopragmatism is necessary for elaborating the origin from which our understanding of the logic of practical reasoning

[51] See Quine (1960, Chapter 2). I see this as a rather trivial implication of locality of rationality, as explained earlier.

emanates. Our life experience makes that logic semantically transparent because the propositional attitude terms grow in and out of our practical realities. Hence, the "essence" what these expressions capture is not stable, and the same propositional attitude descriptions may describe different contents to different persons. It is important to note that even if the propositional attitude terms refer to behavior (and derivatively to thoughts, as explained earlier), we need to make sense of that behavior and that happens by (mostly tacit) recourse to our own experience as agents in the world. This does not mean that our folk psychological reports are always correct, even in the case of our own behavior, but that the involved *concepts* obtain their *contents* from our human conduct, including how we think and talk about it. We often misinterpret others' reasons, and the rationale of even our own behavior may sometimes be completely obscure to us and require a cognitive scientific explanation. Again, our understanding of these latter explanations is grounded in our scientific practices. Propositional attitude terms are explicit statements that capture aspects of our life-world and our orientation toward it but not necessarily the essential aspects of the cognitive capacities that enable us to think and behave in meaningful ways. So, if a lion could talk, we could not understand him and presumably, neither would Aristotle readily understand the Bayesian rational analysis of intuitive decision-making.

Furthermore this means that there are no strict objective facts about intentional attributions because different capacities, demands, and experiences will end up dissecting the world differently. However, as a matter of empirical fact there are gradual similarities between intentional contents and, therefore, more or less literal and derived ways of attributing them. What is implied is that there is no one privileged source of meaning and no possibility of general analysis or identification criteria of intentional content, at least in propositional terms.[52] Later, I will explain how this will pertain to abstract and discursive concepts and that all understanding is based on our ability to use concepts skillfully, including "intellectual" understanding. If this is correct, then the linguistic meaning is a form of (or at least mostly constituted by) pragmatic knowledge rather than a separate representational realm. This blurs the distinction of the two, and if one wishes to hold a strict discursive theory of

[52] To be more precise, there should be no hegemony of language, logic, non-linguistic practices, perception, or cognitive faculties in determining conceptual content. They all take part in constituting content both in subjective and in intersubjective spheres because they all take part in defining the limits and character of our practices. To what extent each listed factor is relevant depends on the specific tasks.

concepts (i.e., that genuine conceptual content is linguistic knowledge), we risk claiming that most if not all human intentional behavior is not strictly speaking conceptual, linguistic behavior and thinking included.

Even after dismantling the possibility of a determinate analysis of propositional content, the question about content similarities and differences still makes sense. However, the relative indeterminacy of content ceases to be a merely philosophical question and becomes an empirical one. Content determination may be a difficult task (depending on the case), but basically content determination problems now lapse into empirical underdetermination of factual claims, which we cannot entirely avoid in science anyway. My argument aims to show that (a) non-linguistic content often precedes linguistic content (I take the latter to be determined by enactive extension of neopragmatist inferentialism) both ontologically and psychologically, but (b) in some cases they are independent to a degree, but (c) even when they are, inferential use of language almost invariably recruits non-linguistic cognitive resources. I aim to blur the distinction of sentience and sapience as much as I can. Sadly, this distinction appears to be the last line of defense in the border of causality and intentionality. I hope the reader is already accustomed to the basic drift that I do not feel much entitled to defend any dualisms of the enlightenment, be they intuition vs. reason, mind vs. matter, subject vs. object, or thinking vs. doing. I do not think that what I am sketching here denies us of the understanding of these distinctions or of what the clear cases of the agency are. If it robs us the notion of clear boundaries between mechanisms and agents, so be it.

The above remarks rather define a problem than answers one. If we get rid of the aforementioned important frames of modern philosophy that have shaped our understanding of epistemology, ontology, subjectivity, and mental concepts what questions can be offered in their place? The problem becomes to determine exactly the pragmatically relevant variables and relations that conceptually structured cognitive processes are sensitive to, and what is the proper language to be used in describing the allegedly non-propositional pragmatic content. The first problem is empirical, and we can get a grip on the latter if we put the skills and functional knowledge in the foreground of our analysis and try to explain conceptual understanding as an adaptive skill pertaining to practical expertise.

2.3.1 Embodied and enactive alternatives

To recap the discussion thus far, I think neobehaviorism and linguistic pragmatism both get aspects of propositional attitude ascriptions right. Neobehaviorism harbors the correct idea that intentionality is not fundamentally about reasoning or propositional thinking but about doing. Neopragmatism provides a fruitful naturalistic account of discursive concepts and ontology of language. What is particularly important is the idea that linguistic meanings are social practices and that utterances can be considered analogical to tools in how they take part in concrete causal processes. Both of these approaches, however, are ill-suited for describing non-propositional conceptuality in its own terms and hence do not readily provide any analysis of the nature of intentionality that is operative when an organism couples with its environment through action. Fortunately, under the banners *enactive*, *embodied*, and *embedded cognition*, a body of research is emerging specifically addresses this aspect of intentional action. I will not discuss these positions in detail for reasons that are explained below; however, a review of the core characteristics of these overlapping trends is useful before we conclude this chapter.

Differentiating between these three approaches is not straightforward mainly because they share many common characteristics, and different theorists hold different views and emphases. The field is currently taking shape and lacks a canonical formulation. Often the collective header *4E* is used to refer to these research paradigms. The fourth E comes from "extended"—the idea some of our cognitive processes and memory systems are distributed in the environment (Clark & Charlmers, 1998; Clark, 2008). What can be considered the unifying theme is that cognition and intentional phenomena are active processes that are manifest in a dynamic coupling of the organism and its environment through its bodily action. The point is incisively paraphrased in the subtitle of Andy Clark's (1997) book as "putting brain, body, and world together again." The message is that mind is not a representational system attached to the brain, insulated otherwise from the world but through the body that works as the brain's input/output interface—like cognitivists often seem to think.

Often the terms *enactive* and *embodied* cognition are used interchangeably and neither do I mind the subtle differences they might carry. However, as a terminological choice, I would rather call my stance enactivist because I put more theoretical weight to action than to the body. My understanding is that some form of embodiment is at least implicitly presumed in any enactivist

stance. Although enactivism is relatively recent development (the term was introduced in Varela et al. 1991),[53] it has important predecessors; for example biologist Jakob von Uexküll (1926), developmental psychologist Jean Piaget (1952), perceptual psychologist James J. Gibson (1979), and, in this context, oft-cited phenomenologist Maurice Merleau-Ponty (1996).[54] Merleau-Ponty, in turn, refers to Edmund Husserl's notion of "operative intentionality" in an introduction to his anti-cognitivist theory of perception:

> When I begin to reflect, my reflection bears upon an unreflective experience; [...] Perception is not a science of the world, it is not even an act, a deliberate taking up of a position; it is the background from which all acts stand out, and is presupposed by them. The world is not an object such that I have in my possession the law of its making; it is the natural setting of, and field of, all my thoughts and all my explicit perceptions. (x–xi)
>
> [...] Husserl distinguishes between intentionality of act, which is that our judgments and of those occasions when we voluntarily take up a position [...] and operative intentionality (*fungierende Intentionalität*), or that which produces the natural and antepredicative unity of the world and our life, being apparent in our desires, our evaluations and in the landscape we see, more clearly than in objective knowledge, and furnishing the text which our knowledge tries to translate into precise language. [...] Through this broadened notion of intentionality, phenomenological 'comprehension' is distinguished from traditional 'intellection' [...] (1996, xviii)

This non-propositional intentionality is precisely what I am after: intentionality that is manifest in action and perception, that precedes our intellectual reflective understanding and furnishes our discursive reason with meaning. The notion and the importance of operative intentionality are again raised by Shaun Gallagher and Katsunori Miyahara (Miyahara, 2011; Gallagher & Miyahara, 2012) in their attempt to fuse neopragmatism with enactivism and to hence produce a somewhat similar account of human intentionality to what is developed here.

[53] Although psychologist Jerome Bruner already used the term in similar but somewhat more restricted sense in 1966 in his book *Toward a Theory of Instruction* (p.10–11) .

[54] Other notable influences/concurrent developments are, for example, autonomous robotics by Rodney Brooks (1999) and embodied cognitive linguistic of George Lakoff and Mark Johnson (1999).

In 1991, Francisco Varela, Evan Thompson, and Eleanor Rosch described enactivism as stemming from the observation that there are many ways the world can disclose itself to different organisms, depending on the structure they have and the kind of distinctions they can make. They wrote that "[...] cognition is not the representation of a pregiven world by a pregiven mind but is rather the enactment of a world and a mind on the basis of a history of the variety of actions that a being in the world performs." (1991, 9) Thus, the mind is a plastic and dynamic system, and a cognitive agent's world is a relational (in contrast to absolute) entity formed through the autonomous activity of the agent with its specific mode of coupling with its environment. The same conviction is echoed in the later works of these authors. Varela (1999) cites the works of Lakoff and Johnson and asserts that higher cognitive capacities "also emerge from recurrent patterns of perceptually guided action" and that "the world we know is not pregiven; it is, rather, *enacted* through our history of structural coupling, and the temporal hinges that articulate enaction are rooted in the number of alternative microworlds that are activated in every situation" (Varela, 1999, 17). These "microworlds" are basically situations or events with which we cope in our daily lives. Thompson (2007), in turn, stresses that cognition is the skillful exercise of embodied know-how in situated action. This is an important idea, which will be echoed throughout the rest of this work.

The enactivist framework focuses heavily on sensorimotor mechanisms that are engaged in all perceptually guided activity of the organism. The point is that cognition-in-action is not best thought of as a linear perceive-compute-act cycle but as real-time coordination of behavior and changing environmental variables. In this line of thinking the intrinsic activity of the organisms marks the difference between intentional agents and mere responsive mechanisms, such as vending machines or rusting blocks of iron. The upshot is to insist that all cognitive processes are truly cognitive insofar as they participate in the corporeal activity of the organism, including the ones that are often thought of as information pick up processes, such as perception and categorization. For these reasons, enactivist and embodied paradigms are closely aligned with mobile robotics, self-organizing systems, and artificial life research rather than artificial intelligence. This is not meant to be an all-out assault on reflective reasoning but to insist that action itself is meaningful, cognitive processes that take part in perception and action have intrinsic intentional content, and that operative intentionality is constitutive also of reflective thought and hence fundamentally conceptual. As Di Paolo et al. (2011, 39) remark:

> Regulation of structural coupling with the environment entails a direction that this process is aiming toward [...] This establishes a *perspective on the world* with its own normativity [...] Exchanges with the world are thus inherently significant for the agent, and this is the definitional property of a cognitive system: the creation and appreciation of meaning or *sense-making*, in short.
>
> [...] Organisms do not passively receive information from their environments, which they then translate into internal representations. Natural cognitive systems are simply not in the business of accessing their world in order to build accurate pictures of it. They participate in the generation of meaning through their bodies and action often engaging in transformational and not merely informational interactions, *they enact a world.*

The knowledge involved in these interactions serves us to respond to the environment selectively and anticipate how the world—or our perception of it—changes as a result of our activity; thus, intentional content tied to enaction is both meaningful and "goes beyond information given" while it remains non-propositional and non-representational because it is constituted from the viewpoint of the organism and intrinsically *for* the organism. Whether this kind of content can be considered conceptual or not naturally depends on the analysis of concepts one adopts, but I hasted to add that the matter is somewhat orthogonal to the intuition/deliberation distinction. There is no reason why conscious deliberation should necessarily involve symbolic representations because you can reason with mental images or other sensorimotor representations, and such thinking *may* constitute propositional but non-discursive and enactive reasoning. As Alva Noë has put it, "we need only to recognize that the concept of experience and the content of thought [as representing things as being thus and such] can be the same" (2004, 190). The issue is discussed in detail in Chapter 4

Neither does enactive intentionality cut across linguistic/non-linguistic demarcation, because this sort of mental content does not imply that all perceptual concepts should have lexical labels nor deny that some have. Some content involved in operative intentionality is readily communicable, although it clearly does not have meaning in referential semantics sense. Rather, this sort of content is something that comes out as the intuitive interpretation of propositional attitude ascriptions and which is, as explained earlier, intersubjectively com-

municable and understandable to the extent the participants share the same capacities and practical world as defined by enactivists.

Another key tenet that I will exploit comes from Valera with considerable influence from Dreyfus brothers (Dreyfus & Dreyfus, 1986). It connect enactivism with dual-process theories and highlights the importance of expertise in intuitive reasoning and casual coping with the world:

> My interest in immediate coping does not mean that I deny the importance of deliberation and analysis. My point is that it is important to understand the role and relevance of both cognitive modes. It is at the moments of breakdown, that is, when we are *not* experts of our microworld anymore, that we deliberate and analyze, that we become like beginners seeking to feel at ease with the task at hand. In this light one can say that computationalist cognitive science has been mostly concerned with the behavior of beginners and not with that of experts. (Varela, 1999, 18)

The moral is that if tacit and automatized commonsense reasoning is an expertise that pertains to our daily activities, then it is this practical know-how that cognitive science needs to study, instead of deliberative problem-solving and explicit know-that, if it intends to reveal to us the fundamental nature of cognition. Moreover, concerning intentionality and mental content, the meaning constitutive relation goes precisely the opposite direction in comparison to cognitivism or Brandomian linguistic pragmatism: Action does not inherit its meaning from acts of deliberation by the agent; it is the other way around. Enactivism comes close to neobehaviorism; however, is also contrasts with it, in that intentional content is not an interpretation of behavior but instead engendered in it. Ontologically, the only thing that incorporates meaning in the world is the intentional activity of organisms. Certainly, interpreting behavior *is* behavior, but the first order intentionality is realized in the action itself and for the agent itself. Therefore, the manifestation of intentionality does not necessitate a separate act of interpretation. It is difficult to tell whether enactivism strictly inclines toward the internalist or the externalist side of the fence since enactivist intentionality is supposed to be found in junction with inner and outer realms if analyzed according to the traditional philosophy of mind.

Because of these characteristics, enactivism has always been closer to phenomenological tradition than to logical analysis, even while it promises not to be a descriptive project about human experience but an empirical framework

for psychology. This has sparked hopes not only for phenomenologisizing the cognitive science but also for naturalizing phenomenolgy[55] and thus perhaps narrowing the notorious explanatory gap between the physical reality and the consciousness.

For the philosophers accustomed to analytical rigor, everything said above may appear vague and programmatic at this point. The enactivist framework is somewhat challenging to describe concisely because it is more a collection of ideas that indicates a problem space rather than a solution. Consequently, many authors have different interpretations of what enactivism and embodied cognitive science is or should be about. There are reasonable doubts that enactivism is too vague and radical and whether or not it can really provide a new framework for modeling cognition and not just a new vocabulary for conceptualizing mental processes in novel ways—which, of course, is an achievement in its own right.

Enactivism is closely related to a modeling framework called *dynamic systems theory* (van Gelder, 1995; van Gelder & Port, 1995; Beer, 1995; Thompson, 2007). Applying dynamic systems theory in cognitive science is not an entirely new innovation. It originates in Ross Ashby's general cybernetic theory of the behavior of organisms and other systems (1956; 1960), which influenced the second generation of neural network models (Rosenblatt, 1958, 1962) and later found its way back to the foreground of cognitive science through developments in connectionist theory in the 1990s[56]. The basic idea of dynamic systems modeling is to identify a set of variables that characterize the system and define how their values depend on each other. The state of the system can be defined as a point in n-dimensional space, defined by value ranges of the n-variables. The state space of the system can then be represented geometrically as a manifold, which is essentially an $n-1$ dimensional surface, formed by the points that represent possible configurations of values (i.e. states of the system). The shape of the manifold and the resultant dynamics is determined by functions that define dependencies between the variables and how the state of the system evolves as a function of time in each point. Thus, the manifold is a vector field, and the system dynamics can be represented as a trajectory through the state space on the manifold. Essentially, a dynamical system simply is this vector field.[57] The

[55] See e.g. Thompson (2007) and Petiot et al. (1999), but also Zahavi (2004) for critical remarks about the latter project.

[56] See Bechtel & Abrahamsen (2002, Chapters 8 & 9.)

[57] For an accessible introduction to basic concepts of dynamical systems see e.g. Abraham & Shaw (1992) and Norton (1995).

aim is to represent how a complex system of interdependent variables evolves in time. Inputs to the system are conceptualized as perturbations to its intrinsic dynamics, and if the system is self-organizing—like a living organism—then its internal dynamics can be considered to consist of triggered compensations to push the system back to acceptable regions in state space. As an explanatory framework, dynamic systems theory aims to capture the internal and external forces that shape these trajectories as they unfold in time (Thompson, 2007, 10–13). If the variables are highly connected (as in feedback loops) and the functions that characterize the system are non-linear, even seemingly simple systems may exhibit surprisingly complex behavior.

Dynamic systems theory in cognitive science approaches cognition as a temporal phenomenon, and depicts the internal dynamics of cognitive system as a self-maintaining causal manifold rather than an instruction set for computing symbolic representations. Dynamic systems formalism is commonly utilized in connection with (artificial) neural networks. In case of neural networks, the variables are neuronal units, their values are neurons' activation states, and the manifold is determined by the strength of connections that pass signals (i.e. unit activations) between the neurons. It is not always obvious what the relevant variables should be in enactivist dynamic systems; however, the basic idea is to identify the relevant causal components of the organism (e.g., brain states, joint angles, direction of gaze, hormonal levels etc.) and the environment and then model how the cognitive dynamics emerge as circular and non-linear interactions involving brain, body, and the environment. *Practically* that is pretty tall order, but this gives some conceptual tools to understand what "dynamic coupling with the environment" might mean in causal terms and how to analyze it. Work done with simple autonomous robots has demonstrated the power of embodied dynamics in some decision-making, route planning, and perceptual discrimination tasks. The complexity of many motor and perceptual problems can be significantly reduced if the corporeal activity of the agent is taken into account (Brooks, 1999; Beer, 2000; Floreano et al., 2004; Floreano & Mattiussi, 2008).

A powerful demonstration of utility of embodied dynamics is a simple vision discrimination robot designed by Floreano et al. (Floreano & Mattiussi, 2008, 474–475). The system uses a camera that can pan, tilt, and zoom and contains a single-layered neural network for shape recognition. It is a fundamental result that single-layered networks can only solve linearly separable problems (Minsky & Papert, 1988); however, the robot managed to achieve 100% accuracy in a linearly non-separable recognition task. What made this

possible was factoring in the same model with the system's cognitive architecture, the environmental variables, and behavioral capacities and how the processing unfolds in time. This changes the formal definition of the problem and hence the computational requirements. *All* these variables and not just the neural network do genuine explanatory work in the robot's visual recognition capacity, which remains unaccounted for if we restrict our analysis to its internal cognitive machinery. In general, the hallmarks of systems designed in autonomous and active robotics is that they lack central control and often do not allow a neat functional decomposition. Instead, the behavior they exhibit is the result of the whole system coupled with its environment.

The bad news is that although dynamic systems theory may be a powerful tool for modeling things such as active perception, insect locomotion, perception–action loop in tennis playing, and other smooth "online" behaviors, it is far less clear if it helps at all in understanding higher cognitive functions such as causal induction and "off-line" reasoning. Undoubtedly, the latter are causal processes unfolding in time, and it is not an issue whether higher cognition can be in principle modeled with dynamic systems framework. The theory can be defined in a way that trivially accounts for any causal system.[58] Moreover, it might be true that "when we examine very simple level intelligence we find that explicit representations and models of the world simply get in the way. It turns out to be better to use the world as its own model" (Brooks, 1999, 81). However, when we examine reflective reasoning, flexible problem-solving, planning, and other tasks that used to be at the center of cognitive research, it appears that the notions of conceptual content and representation are indispensable. To use Andy Clark's example (1999, 348), consider whether or not US gun manufacturers should be held liable for having knowingly manufactured more guns than the legal market could possibly account for. It is thoroughly mysterious how sensorimotor capacities could possibly account for all the moral and abstract reasoning involved. Still, human cognitive life is filled with this sort of "representational hungry" tasks. Perhaps sensorimotor simulation has a substantial role in human reasoning; however, it is hard to understand how to even define the relevant variables that the sensory processes

[58] In Beer (2000), for example, a dynamic system is defined as triplet consisting time series T, set of system states S and transition function $\phi(s,t) \mapsto (s',t')$, i.e. a mapping that determines the state s' of the system at time t' as a function of its state s at time t. There is no other restrictions on elements T, S, and ϕ other than that T should be ordered. This is clearly as abstract as any definition of temporal system can get, incorporating for example neural networks, Turing-machines, or any implementation of cognitive architecture one might think of.

track in higher cognition without some more conventional theory of conceptual representation.

This problem is often surpassed in autonomous robotics by keeping the variables predefined and low in number, which renders the analysis tractable even if conducted on a single neuron basis. With humans, however, we do cognitive psychology precisely because in practice we need to work with more abstract descriptions of behavior than modeling environment interactions at a neural level. To that end, we need to know what higher-order properties exists in the human cognitive system, and it is unclear whether the successes in modeling simple behavior can provide useful research program in this respect. It might turn out that higher cognition also needs to be modeled on a whole organism basis (this is what I am claiming in this work anyway); however, the successes of cognitive psychology are existential proof that functional decomposition and theory formation at a higher level of abstraction is a viable research strategy and arguably the best we have at the moment. Even while enactivist framework has shown prospect in explaining some higher cognitive phenomena (e.g. aspects of social cognition) it is for the aforementioned reasons that radical embodiment and enactivism have received well-earned criticism for excessive radicalism and because of the worries that it might get stuck into accounting only for marginal, even if important, issues in human cognition (Clark & Toribio, 1994; Clark, 1999; Di Paolo et al., 2011). In brief, enactivist program is quite hard to turn into actual working *cognitive* science. Philosophically, radical enactivism might lead to a new form of material eliminativism if it replaces the analysis of how intentional content structures human thought and behavior with a purely causal framework. This might be a tempting position for some but it is not a necessary outcome of either enactivism or dynamic systems account. To what analysis of the mind enactivism eventually leads to depends on how these frameworks are employed in the study of cognition.

2.3.2 Chapter summary

As things stand, the whole $4E$ paradigm is theoretically too heterogeneous to lend itself to a concise analysis, and it remains to see how radical enactivism fares empirically, especially with explaining conceptually structured higher cognition. Although I am convinced that symbolic propositional representation is the wrong unit on which cognitive science can be founded, I am still not entirely ready to give up on the notion of representation altogether. It is perhaps better to think of representations as a forming a continuum from subjective,

implicit, and contextual (or situated) sensorimotor expectations to intersubjective, explicit, and abstract symbolic concepts (see also Clark & Toribio, 1994). It is somewhat a matter of terminology whether the former can be considered representations because is not apparent what they actually represent. However, by definition, they do contain subjective knowledge for generating expectations about how concrete situations and the effects of actions unfold. What I am interested in is to explain how our conceptual understanding builds up from this form of practical knowledge that is drawn from personal experience. I advocate a pragmatist inferential theory of understanding which, in this case, means that understanding is not semantic knowledge about linguistic expressions but practical know-how about how things relate functionally to each other in specific concrete circumstances. As explained earlier, I assume that this ability is founded on tracking causally relevant variables in concrete environments and a capacity to generate expectations of our actions (including refraining from acting) based on the values of those variables.

This approach certainly comes close to enactivism, at least in spirit and it shares the very same problems, especially how to explain abstract conceptual thinking and related higher cognition. My solution is to integrate enactivism and linguistic pragmatism to yield a working model on conceptual cognition, which also includes discursive reasoning. I favor a reformist rather than a radical stance on enactivism. That is, I assume that dynamic systems theory or related approaches might be the best way to model sensorimotor skills in executing appropriate behaviors at appropriate times as external and internal situation continuously change;[59] however, nothing in my account depends on this. I am interested specifically in how our conceptual understanding is constituted by the ways we cope with concrete situations and how this affects our reasoning capacities and understanding of abstract concepts. If those capacities can be strictly reduced to sensorimotor coupling that's very exhilarating, but I make no claims that they can.

My strategy will be to make my case by analyzing rather standard research on cognitive psychology and closely related fields. At first, this might appear like an odd way to pursue a theory that is principally influenced by neopragmatism and enactivism, especially since the latter is supposed to be a radical departure from standard cognitivist research. However, on closer examination it becomes clear that the real target of enactivist criticism is internalist cognitivism and especially computationalist theories that rely on language-like mental representation. A large body of research on cognitive psychology is

[59] As Randall Beer (1995, 174) defined the central task of autonomous agents.

quite antagonistic to this sort of cognitivism anyway and reconcilable with theoretical insights drawn from enactivism; for example, this is how Eleanor Rosch (1999) sees her classic work on prototype theory of concepts. Moreover, recent advances in causal inference and concept research support pragmatism rather than logics oriented accounts of (intuitive) reasoning and conceptual structure. Keeping the discussion approximately at the level of cognitive processes and representation that pertain to commonsense reasoning is interesting from both philosophical and cognitive scientific standpoints. This is because common sense is notoriously difficult to model and the possible novel theoretical insights are directly relevant to several philosophical discussions considering, for example, the philosophy of language and epistemology. Moreover, the enactivist/embodied framework should account for the established research at some point anyway, and I consider it a healthy scientific conservatism to incorporate existing knowledge as much as possible into novel theoretical frameworks. I am more interested in empirical adequacy of the resultant account rather than in the theoretical orthodoxy subservient to a particular philosophical project.

The following chapters employ theoretical ideas drawn from enactivism and philosophical pragmatism to re-examine and integrate selected research programs in cognitive, developmental, and social psychology and, at the same time, to advance a philosophical theory of concepts by showing how conceptual understanding derives from accumulating personal experience via pragmatic action. This contrasts with seeking conceptually sufficient and necessary conditions or other decontextualized and explicit basis for semantic knowledge. The argument aims to solve the problem of how competence for abstract reasoning can be explained by pragmatic contextual knowledge. To that end, the key is the neopragmatist idea that a causal exchange in social interaction engenders discursive meaning.

As explained above, making that case necessitates an account of nonlinguistic conceptual cognition, which will occupy most of the discussion in the rest of this work. Tracking and fluently exploiting relevant causal regularities in real time is far from trivial, and it is vital capacity to both common sense and expert competence. The epistemologically significant idea is that while abstract reasoning is usually associated with deliberation and expertise, the skillful use of abstract concepts requires intuitive encoding of a vast amount of specific concrete situations where the concepts are applied and that the process is similar to the accumulation of practical common sense through personal experience. It is probably wrong to say that intuitive concepts in this sense are information structures that we *use* in reasoning. They rather par-

ticipate in situated action of the agent. These aspects suggest that situations, context effects, and practical skills should be at the fore of concept and category research. These convictions parallel very closely the ideas expressed by Rosch (1999). The resultant explanation blurs the demarcation of abstract and concrete, and theoretical and practical reasoning. These and related distinctions are often considered important in System 1/System 2 classification, and therefore there are implications for dual-systems theory, as explained in the introduction.

3 Bringing the philosophy and cognitive psychology of concepts together

The plan for this chapter is as follows: I begin with an overview of the methods and aims of the empirical part of this work with the further development of some points covered in the previous one. This is followed by a review of the three main strands of concept research in cognitive psychology, namely prototype, exemplar, and theory accounts. I discuss how all these theories capture fragments of human conceptual cognition; an integrative account is subsequently proposed. The basic idea is that category knowledge is represented as feature clusters that coalesce around causal properties. The role of exemplars is to encode situation representations that both provide a source of causal knowledge and mediate the access to category information on a contextual basis. They connect tacit reasoning to category content, and this forms a crude explanation of how a human intuitive grasp of reality works.

In the first paragraph of the previous chapter I referred to Edouard Machery's (2009) summary that philosophers and psychologists have different aims and therefore different definitions of concepts: Psychological concept research is mainly concerned with the nature of information representation that supports higher cognitive processes while philosophers often think of concepts as constituents of propositional content and try to determine the general conditions under which one can have propositional attitudes. While I think this is a broadly correct description of the explicit aims of the respective fields, I argued in passing that there are reasons to believe that the philosophical project can not proceed strictly independently of empirical research (as a matter of fact, not as a point of logic). In the first half of this chapter, I continue with this theme.

The discussion in 3.2, *Category theories in cognitive psychology*, begins to unravel the Machery's (2009) main argument that although the aforementioned three concept theories in cognitive psychology are thought of as rivals, they all tap into equally real, albeit, distinct phenomena. However, they contain different assumptions about information structure, processing, and memory, to the extent that there are no scientifically interesting shared features that characterize all these three capacities. In conclusion, the candidate cognitive mechanisms of concepts form a disjunctive class and therefore the notion of "concept" should be abandoned in cognitive psychology since it does not stand for a specific natural kind but refers to at least three different kinds of processes.

After the short review of these three theories, I will show how they actually may describe integrated rather than an independent set of processes. However, my aim is not to criticize Machery's work but rather build on it. I will not touch his arguments, except for some scant remarks in the concluding chapter. The reason to mention his *Thinking Without Concepts* in these introductory remarks is that I have found this work intellectually liberating because it relieves us from pitting against each other in vain the three theories that describe component processes of conceptual cognition. Therefore I will not devote much effort to argue that prototype, exemplar, and knowledge accounts describe real and independent processes. I take consider this effort properly done already. I will also acknowledge, at least tentatively, that these processes might encode separate conceptual representations; however, I take that this sort of compartmentalization presents limiting cases at best. I am interested in exploring how these processes interact to engender intentional behavior and thought jointly. Although the jury is still out, the recent research supports rather strongly the view that similarity-based categories (i.e. prototypes and exemplars) are not independent of causal knowledge but that the latter forms a fundamental constitutive part of the similarity structure (Medin, 1989; Hampton, 1998; Ahn et al., 2000b; Rehder, 2003, 2009; Luhmann et al., 2006).

Machery's (2009) main conclusion is not contested here. If the integrative theory that I offer is right, then in cognitive psychology it becomes impossible to identify concepts with any specific type of information structure. Instead, possessing a concept should be seen as a product of interlocking capacities where use cannot be strictly separated from content. This means that we should shift focus from static *representation* to active *processing*. This was the main conclusion in the last chapter considering the philosophy of mind, and if the argument in this chapter goes through, it brings philosophical and psychological concept research a significant step closer together.

My contribution, mosty laid out in Chapters 4 and 5, is twofold:

First, I separate prototypes and exemplars to serve different functions in conceptual cognition. The idea is that while prototypes carry summary information about categories, exemplar effects reflect implicit input and output processing, such as selective access to the prototype memory, retrieval of contextually relevant information, causal induction, and generation of causal expectations. In short, exemplar effects arise from processes that support intuitive reasoning.[60] I discuss at length the constitutive role of exemplars and

[60] I use the term "reasoning" mostly as a catch-all term for any conceptually structured cognitive processing. It should not be understood as a explicit rational inference.

implicit skills in conceptual understanding. The proposal rests on a rather standard cognitive psychology and does not postulate novel types of mechanism or representation. The proposed etiology of the exemplar effects is not my idea but stems from Gregory Murphy (2002, 85–88), albeit with slight differences. Murphy proposed that exemplar effects might result solely from memory access and that there might be no stored exemplars at all (2002, 88). He was talking about taxonomic categories; however, I assume that there are stored exemplars of events, which are used to classify situations, interpret stimulus, and support causal inference and learning. The exemplars encode contextual goals and event stimuli, interpreted in terms of those goals and knowledge of past actions and their outcomes. A pragmatic knowledge base is composed of these very concrete and specific event representations. My assumptions about the exemplar representation format and processing are effectively identical to Gordon Logan's (1988) idea of memorized "processing episodes," which constitute the basis of implicit skills. I also draw on a body of empirical and theoretical work that has established a close link with memories of specific exemplars, pattern recognition, and intuitive skills (de Groot, 1965; Dreyfus & Dreyfus, 1986; Dreyfus, 1992; Ross & Kennedy, 1990; Simon, 1992; Palmeri, 1997; Klein, 1998, 2008; Kahneman & Klein, 2009). A prominent feature of these theories is that they consider expertise as knowledge intensive and not computation intensive phenomenon. Expertise is achieved by exploiting associative memory search, cued by superficial features, to retrieve a match to the current situation and using analogical inference as a default reasoning mechanism. This combined with analogical transfer based schema abstraction is basically the cognitive theory of pragmatic knowledge that follows.

My *second* contribution builds on the first and accommodates discursive reasoning into the pragmatic knowledge framework. The idea is based on neopragmatist inferentialism by treating discursive knowledge as a special case of procedural and causal knowledge. Much of our declarative knowledge is not about causal facts, of course, but if inferentialism is the correct theory of discursive meaning, then our conception of discursive content is engendered by our ability to carry out discursive practices effectively. That ability, I maintain, rests on tacit causal and procedural knowledge about discursive acts embedded in wider cultural praxis. If this is correct, it has significant implications because then theoretical reasoning is, at least psychologically, a special case of practical reasoning. This implies that grasping theoretical concepts is not simply a matter of acquiring explicit know-that (although this might be necessary for learning many conceptual domains) but a learned skill that rests

on implicit know-how. Moreover, the hypothesis offers an empirical framework for explaining how and why our discursive reason reflects our local social and cultural practices. In consider the theory to be plain pragmatism, but to the extent that social, contextual, and constructivist character of meaning and reason are central to continental and postmodernist philosophy, this latter contribution can be thought of as something like postmodern or post-structuralist cognitive science, making a rather sharp departure from both empiricist and rationalist traditions in cognitive theory.

3.1 Theoretical and methodological motivations

What comes to the mutual constraints between philosophy and psychology, I here adopt an approach which is reversed in comparison the one in the previous chapter. That is, I assess empirical research of concepts through the lens of philosophical theories of intentional content. Basically this means to investigate the possible cognitive basis of enactive and neopragmatist intentionality as presented in the last chapter. This is the key methodological principle here, and before putting it to work we need to examine its rationale because there are reasonable grounds for skepticism about this kind of an approach. For one, cognitive psychologists may worry that conflating inference and related higher processes with category representation might needlessly degrade the analytic precision of concept research. However, the strict demarcation of process and content does not survive if the theory of concepts proposed here is correct. If use is constitutive of conceptual content, one should expect that some questions concerning conceptual representation remain elusive without consideration of the processes utilizing them. This might be apparent in the case of logical concepts, for example, that are inherently tied to reasoning. The remark, however, pertains even to concrete concepts and issues, such as conceptual combination and coherence, construction of *ad hoc* categories, contextual malleability of category representations, and representations of contexts and situations themselves. Moreover, as human reasoning is content-based, it might be intractable without references to inferential potential, which is directly encoded in conceptual structure, and to how this information is accessed.

After this chapter, I present evidence that category representations are active, task-oriented, and context-dependent constructs that encode, *inter alia*, procedural and causal information that is selectively activated in category retrieval. I further that utilization of pragmatic knowledge is an essential part of sense-making and intuitive situated reasoning mainly by automatically fil-

tering relevant information and generating causal expectations. This sort of tacit inference underlies explicit reasoning and conceptual comprehension and makes fluent commonsense reasoning possible in the first place.[61] Since propositional attitudes essentially are commonsense ascriptions, and commonsense ascriptions depend on our capacity for practical inference, it is theoretically motivated *not* to insulate at least certain aspects of active cognitive processing from concept research. Recalling Eleanor Rosch's point, that was iterated at the end of the previous chapter, I hold that the unit of concept research should be situations rather than precompiled information structures. Concepts participate in the making sense of situations, and it is not clear whether they have any relevant explanatory function in cognitive science, apart from their active interpretative role in situated intentional action.

3.1.1 Methodological discussion

The general point I try to make is that if philosophers intend to discover the conditions under which one can have intentional contents while psychologists try to discover the nature of our higher cognitive competencies, they are both actually studying the same phenomenon—given that the conditions in question happen to be precisely those competencies. This does not mean a wholesale conflation of these research programs, because they obviously have their respective aims and methods but that they should be mutually informative and constraining. Moreover, the theory advanced here contains strong pragmatist and constructivist elements that imply that intentional content cannot be reduced to psychological mechanisms. We also need to understand the lifeworld of the organism to understand how it represents its environment. Hence, the account is not internalist or rationalist. Moreover, we cannot get a grip on exactly what the organism's mind tracks in its environment without understanding its biological and cognitive constitution. Therefore, the account is not strictly externalist or empiricist either. I believe the key to concepts, both in philosophy and psychology, is *interaction*, which cannot be understood if either the environment or the organism's capacities is granted only a second-class status.

Certainly there is an important distinction between content and form, evident in *explicit* logical inference, especially when formal principles are involved.

[61] Something like this view generally shared amongst enactive and neopragmatist theorists; see also e.g. the introduction in Dreyfus (1992), Johnson-Laird (2008, Chapters 4 & 5), and the next footnote.

However, not even formal reasoning is psychologically content free in any clear sense, at least if it utilizes learned concepts rather than innate mental logic or other hardwired formal procedures. Although the computationalist theory that stresses the content/process distinction provided a powerful model for cognitive science for decades, it has not helped much in understanding human concepts and commonsense reasoning, and one of the most pervasive results in the psychology of deduction is that content affects human inference in logically irrelevant ways (Evans et al., 1993, 4–7). In the empirical discussion of this work I mostly focus on *implicit* content driven-inference, underlying commonsense and expert reasoning.

This tacit ability enables us to grasp our lifeworld as self-evident without explicit comprehension and often even awareness of the active interpretative processes taking place. It is not very controversial that at the heart of the cognitive sciences is to discover how this process works, but it is contentious to what extent and how exactly implicit processes are constitutive of explicit reasoning. This controversy dates back at least to the 1960s when Hubert Dreyfus (1965) criticized ongoing AI research for dismissing sophisticated and highly relevant implicit intelligence as seemingly primitive and marginal. Current dual-process theorists generally hold that tacit System 1 processes are the default mechanisms for executing routinized tasks; however, whether explicit reasoning is considered as independent or fundamentally dependent on implicit cognition is a matter of controversy.

Whether or not there is a categorical demarcation of implicit and explicit reasoning processes, I work under the assumption that there is no such strict division in higher cognitive capacities. Implicit know-how can become (partly) explicit, and explicit procedures can be tacitly learned. Regardless, as far as cognitive skills are concerned, implicit cognition is primary. Because I consider intentional content to be founded on tacit inferential skills (broadly construed) it follows that, by and large, implicit cognition is also fundamental in content determination. However independent explicit processes might be, they generally need content that is provided by the tacit system.[62] A quotation by psychologist Arthur Reber illustrates my position: "...to 'get the point' without really being able to verbalize what it is that one has gotten, is to have

[62] See Evans (2009) and Johnson-Laird (2008, Chapter 5). On the implicit/explicit distinction and the primacy of the former see e.g. Reber (1989, 1993) on implicit learning, Sun (2002); Sun et al. (2005) specifically on cognitive skills, Evans & Over (1996) on reasoning, and Karmiloff-Smith (1992) on conceptual development.

gone through an implicit learning experience and have built up the requisite representative knowledge base to allow for such judgment" (Reber, 1989, 233).

However, primacy should not be understood as total hegemony. Many tasks are executed as mixtures of automatized implicit procedures with explicit conscious control. Arithmetical calculation (Rumelhart et al., 1986, 45) and reading (Shiffrin & Schneider, 1977, 161) are oft-cited examples in literature. Both are examples of skills that contain learned automatized procedures that were initially carried out explicitly. For example, most people presumably learn single-digit addition by serial counting (perhaps with fingers) until the task is internalized by associating the correct answer to numbers to be added without explicit counting. This capacity is then further exploited in multi-digit addition by breaking down the problem into a controlled series of single-digit computations. This is perhaps one of the clearest examples of complementary contributions of implicit and explicit cognition where automatized and controlled processes execute different aspects of the task.

Thoroughly learned attention and control demanding tasks tend to form habituated chunks whose execution does not require explicit cognition. These chunks come in varying degrees of access and control. Complex tasks with multiple dimensions generally promote implicit learning, which yield competencies that are hard to articulate and even carry out by following explicit instructions (Sun, 2002; Reber, 1989). The habituated chunks can be exploited to organize complex hierarchical behaviors where control and attention are shifted to more strategic and general aspects of the task (Vallacher & Wegner, 1987; Christensen et al., 2016). Now, linguistic concepts just might be these sort of chunks that support various communicative, interpretative, inferential, and other complex tasks. If that is the case they can be procedurally opaque and semantically transparent at the same time, meaning that our phenomenology of conceptual understanding stems from our fluent intuitive use of these inferential chunks while we might be quite unable to explicate what their proper use is and how we actually use them. Complex higher-level tasks, which are embedded in wider contexts and require explicit thought, make sense just because these underlying component processes are already understood, i.e. effectively applied without reflection.

This might look like a promising avenue for a philosophy of concepts to pursue because it seems to respect broadly construed inferentialism and explain interesting aspects of the phenomenology of conceptual understanding. However, on philosophical grounds, a good case can be made that the inquiry into implicit cognition does not really amount to the research of concepts in

the sense philosophers use the term. This sentiment often stems from the past divorce of psychological theories from the classical account of concepts.

3.1.2 Problems ahead: Kripke–Putnam externalism

For example Georges Rey (1983) has argued that there is a substantial difference between *conception* and *concept*. The former pertains to epistemology and belongs to the domain of psychology, while the latter is a matter of metaphysics. Concepts, properly construed, are factual descriptions of essences or necessary and sufficient conditions for category membership. Until the 1970s this was also the received view among psychologists (Murphy, 2002, Chapter 2). But things changed decisively when similarity based theories entered the scene. The new paradigm held something that Wittgenstein (1953) had already proposed around two decades earlier, i.e. that most concepts hardly have any defining features but only more or less vague set of characteristics that generally are neither sufficient nor jointly necessary. Later, it turned out that even if there were such defining properties, they—as a matter of fact—are often not used as the basis for categorization (Murphy, 2002, 25–28). Hence, the human conceptual system does not track analytical truths nor metaphysical essences.

Of course everyone knew already that we are unable to define many concepts and that there are numerous inherently vague cases; however, this is hardy a conclusive argument against the definitional view, especially if one considers the defining characteristics as metaphysical and not epistemological properties. But the new psychological theories sought to explain where this ignorance comes from and, more importantly, held that conceptual competence without knowledge of definitions is a built-in property of the human mind. In this new outlook, the lack of definitions did not reflect ignorance but a normal functioning of conceptual cognition.

What follows is that if the classical account of concepts is correct, then psychological research cannot shed light on the true nature of concepts but only on how we, as the limited creatures we are, can conceive them or rather their degenerate substitutes. If so, the new empirical paradigm cannot refine philosophers' conception of concepts but it only proves that there is an irrevocable departure of psychology from philosophy since empirical theories fail to meet the essential desiderata of concept theory, which are, according to Rey, to explain the semantic stability, communicability, tracking of counterfactuals, metaphysical taxonomy, and certain epistemic functions. Because psychology only pertains to the last item, using empirical theories to refine the philoso-

phy of concepts would only amount to a naive conflation of metaphysics with epistemology.[63]

Underlying the above argument is the widely acclaimed Putnam–Kripke line of semantic externalism (Kripke, 1971, 1980; Putnam, 1975), which can be roughly summarized thus: Meaning is the extension of a word. Extensions are classes that are defined by the necessary properties that their members share and that makes them the entities they are. So far nothing new, but these properties, while conceptually necessary, are often not knowable *a priori* and in fact they might not be knowable at all. For example, modern science has shown that gold is an element with 79 protons in its nucleus, and it is this property that makes gold atoms as instances of gold. This particular essential property clearly cannot be known *a priori*. It is an empirical fact discovered by science. However, if it is a fact, it is a metaphysically necessary condition for something to be referred correctly as "gold" once the extension of the term is fixed. The theory is tailored to respect common intuitions about the semantics of natural kind terms: If category members share an internal structure, it is this hidden structure rather than observable surface properties that determine their identity as category members.[64]

Importantly, the theory also applies to other kinds of concepts that are often taken to be representative of the classical definitional theory, namely formal ones. For example, it is not known whether every even integer greater than

[63] There is also an influential internalist line of argument against similarity-based theories. The standard objection is this: Similarity based categories do not compose. That is, a feature set of a typical Pet fish is not a combination of features of typical Pet and typical Fish. But concepts do compose in the sense that the meaning of "pet fish" is a combination of the meanings of "pet" and of "fish", and this property, and respecting boolean functions in general, is essential in explaining productivity and systematicity of thought. (Fodor & Pylyshyn, 1988; Fodor & Lepore, 1996). Also, if the problem with the definitional theory is that we do not know the putative constituents of concepts (i.e. definitions), exactly the same issue plagues the similarity based theories, i.e. we are unable to articulate the features we rely for similarity computations (Margolis, 1994). I will ignore this particular discussion for the following reasons. First, on theoretical grounds the relevance of compositional systematicity as an explananda can be reasonably questioned, especially the role of logical processing as its necessary explanandum (Matthews, 1997; Bechtel & Abrahamsen, 2002, Chapter 6.). Second, even the early prototype researchers (Smith & Osherson, 1984) recognized the problem with conceptual combination; however, it was taken as an empirical issue worthy of a research program that has since yielded solid results (Murphy, 2002, Chapter 12). The standard objection gets it right that prototypes cannot be combined by simple extensional operations (e.g. set intersection or union) but goes wrong in assuming that extensional logic is the only viable route to conceptual combination.

[64] See Kripke (1980, lecture III) and Putnam (1975, 160).

2 can be expressed as the sum of two primes or not, and as things stand, it is unknown if this can be even proven to be either true or false. Still, this sort of property is an excellent candidate of something that is necessarily true or false of numbers.[65] But never mind, for the point of this theory is that we do not need to know these things to use concepts such as *gold* or *an integer* correctly. Conceptual necessity is metaphysical necessity to which human intuition has no privileged access (Putnam, 1975, 151). This means that intensions, conceptions, or cognitive contents are fundamentally irrelevant to meaning. What is relevant is that we track correct referents, not essential properties. According to this theory, concepts work like proper names: Someone names an instance of a category and then the use of the term spreads through cultural transmission. Whether folk's conceptions of the properties of the referents are correct or not depend on what properties the referents in fact have, but for semantics it is immaterial whether we can ever define or even recognize these properties or not, hence Putnam's famous catchphrase *"'meanings' just ain't in the head!"* (Putnam, 1975, 144)

Linguistic pragmatists agree that meanings are not in the head, but they are not metaphysical essences "out there," either but immanent social institutions. We are often unable to articulate exact word meanings, not because we lack access to metaphysical facts but because the underlying social norms that govern meaning engendering practices are implicit—including the instituted use of those very words. Note that while this implies semantic constructivism, it does not entail metaphysical anti-realism. The claim is that intentional content is engendered in interaction with the environment. For us, the relevant environment is partly socially constituted; however, the practices of human communities are also shaped by the physical environment and other material conditions. If these practices are the foundation of linguistic meaning then even

[65] Note that there are some technicalities involved with this example. It is contentious if knowing non-trivial mathematical facts equals proving them, and second, "provability," in the technical sense intended here, is not an absolute but an axiom-system related notion. Moreover mathematical antirealists may raise issues whether non-provable properties can be considered necessary. The reader may want to consult e.g. Clarke-Donae (2013) and Koellner (2006) about conceptual issues regarding absolute undecidability in mathematics. The papers discuss the continuum hypothesis instead of Goldbach's conjecture referred in the paragraph. Nevertheless, the point is that while after Gödel's incompleteness results it is commonplace to informally talk about "unprovable statements", it is unclear if, strictly speaking, there are any. Gödel's results state that for any finite and consistent theory T, which is expressive enough to formulate arithmetics, there is a true proposition p which can not be proven in T; however, this does not mean that there is an absolutely undecidable statement p, in the sense that it can not be proven in any such theory.

while our (linguistic) concepts are ontologically dependent on our communal deliberation, cultural practices, and historical traditions, it does not mean that their putative referents are or that human conceptual system is independent of physical reality. Scientific realism holds that this is generally true also of theoretical concepts which are negotiated by scientists and perhaps other experts. For example, planets as physical objects certainly exist regardless of our astronomical practices, even while their respective *status as a planet* does not. This is because "planet" is a socially negotiated concept for human purposes whose extension depends, among other things, on how specific expert communities see best to carve the nature at its joints. The ex-planet Pluto is a case in point. While the pragmatism advertised here is constructivist about conceptual content, it presupposes materialist metaphysical realism.

To be sure, in the "meaning of 'meaning,"' Putnam stresses many of the same points I try to argue here, namely that social factors and individual aptitude in concept use are important. But these factors are secondary to him on what comes to the determination of conceptual content. The position I am arguing is certainly more antagonistic to Kripke–Putnam externalism beyond these remarks, given that their theory is supposed to be a general theory of meaning. Their theory suffers from the same fault as Fodor's causal semantics. It presumes a form of intentionality without reductive (i.e. non-intentional) explanation, specifically that we have a practice of interpreting natural kind terms in certain essentialist ways. But perhaps Putnam, Kripke, and their followers are just articulating common linguistic intuitions in selected conceptual domains and trying to extend this to linguistic meaning a bit more generally. If that extension is not very wide, it does not threaten the method employed here. Neopragmatism gives us a general reductive account of linguistic meaning, and what we eventually want is a more fundamental theory that also covers non-linguistic intentionality, as explained in the previous chapter. If semantic externalists do not dispute the importance of that agenda, and only consider it to lie outside the scope of their theory, then our explananda simply do not align, and I fail to see why these projects should be contradictory: Then Kripke–Putnam externalists are concerned only with particular practices associated with a specific fragment of natural language, and not with meaning or intentionality in general.

Then again, neopragmatism and especially enactivism can be criticized precisely on the grounds that they are merely framework theories about how *subjective* understanding is engendered, and therefore they fall short of the philosophers' project because they disregard the essential desiderata of con-

cepts harbored in analytic philosophy. Well, clearly they do. In the last chapter, we saw that analyticity and universality go right out the window. What is even worse, I later argue that concepts lack even *intra*personal stability. This is, of course, an anathema to the adherents of the classical account. On the face of it, I could probably leave the issue at that because perhaps philosophers of language, at least in the analytic tradition, simply discuss different things that I have on the agenda. However, what is at stake here is this: By agreeing to disagree with the philosophical project, there would be little point in using philosophical theories of intentional content to constrain interpretations of empirical research, because they should track entirely different phenomena. Therefore, to counter the argument from Kripke–Putnam externalism by this sort of evasive maneuver is not really an option for me. Instead, I need to confront it head on and explain the semantic intuitions behind their theory within my framework. This will be addressed later (see p. 155).

In the last chapter, I argued that the philosophical study of concepts cannot be strictly separated from empirical research because, methodologically speaking, we often conduct a philosophical inquiry by grounding our theories on our analytical intuitions which, in turn, are nothing but products of our implicit psychological faculties. In the context of the philosophy of mind, this tends to lapse concept research into speculative or introspective psychology because the method employs the exact conceptual capacities that it is probing. While this does not necessarily imply vicious circularity, it makes a bad division of scientific labor, in that philosophers first should figure out what concepts are and then psychologists should try to find out how or to what extent we can possess such things. For one, empirical findings change the way we think and discuss concepts—that is the point of research after all, and this arguably changes our analytical intuitions. That, I gather, is the philosophical impact of dispensing with the analytic/synthetic distinction; that is, giving up on a clear distinction between what we know about things and how we conceptualize things.

Moreover, analytical intuitions are notoriously unreliable and unstable across cultures, individuals, and contexts. To understand semantic intuitions, it would help us to know where these discrepancies come from. Even if intuitions were stable, we should ask what gives them epistemic authority, especially in peculiar and wildly counterfactual thought experiments. If the source of this (in)stability is individually learned psychological dispositions, empirical research can be useful if not indispensable tool in understanding methodolog-

ically important aspects of philosophical inquiry.[66] Some instability of intuitions should be expected even among experts, if philosophical intuitions stem from the same psychological mechanisms as any judgment making process.[67] Moreover, as James Beebe (2012) remarks, intersubjective *stability* of intuitions among experts may be a sign of socialization rather than of mutual enlightenment. However, even if that is the case, it does not mean that shared intuitions are always incorrect or untrustworthy but that examining their source and factual character should be on the agenda of philosophers who wish to be careful with their methodology.

Lastly, if it turns out that our philosophical analyses of concepts are fundamentally incompatible with psychological results, the apparent implication would be that human thinking, reasoning, and linguistic understanding are not strictly speaking conceptual. If you accept this, you probably should also stop grounding conceptual facts in analytical intuitions. This conclusion, of course, presumes that analytical intuitions are products of our conceptual cognition and that human mind learns or constructs meanings rather than has access to the realm of objective semantic facts. The previous chapter aimed to show precisely that: Intentional content is a product of the activity of a cognitive agent and that the ontological foundations of discursive meaning are in social and linguistic behavior. The evidential basis of many of the claims was left more or less open, and in this chapter and the following ones, I present empirical results that back those claims and examine these empirical results from the standpoint of pragmatist inferentialism.

3.1.3 Section summary

Following the reasoning laid out above, I try to motivate the remarriage of philosophical and psychological concept research by questioning the rationale of strictly demarcating concepts from conceptions. The theory required to account for non-linguistic intentionality is somewhat hard to formulate. There are attempts to produce generic theories of content that should apply to non-human agents also. Except for instrumentalist theories such as neobehaviorism, these generic theories often need to bootstrap intentional content from biolog-

[66] See e.g. Knobe & Nichols (2008) for an overview of this point. A body of empirical research on philosophical intuitions have been done, e.g., on folk conceptions of knowledge, which shows how social and cultural background, context, the framing of questions, and other factors affect intuitive responses to thought experiments (Beebe, 2012).

[67] See e.g Kahneman (2011) and especially Gendler (2010, Chapter 5 & 6), which address directly the nature of intuitions and philosophical methodology.

ical or other teleological functions, the notion of normal conditions, and the like (e.g. Millikan, 1989a,b). I have found these attempts rather unsatisfactory; however, I have no intention of criticizing them here. Eventually, some of these theories may prove to be fruitful, but to me it seems they are doomed to lapse into underdetermination problems discussed in the previous chapter. In general, there seems to be no theoretically privileged unique descriptions of *why* some organism does what it does and *what* it is trying to achieve—i.e. what are its ends and the rationale of its means. Hence, with non-human agents, it is often unclear if propositional attitudes can be even properly applied.

Recall from the last chapter that with humans we generally do not have this problem. We have a sort of epistemologically privileged (even if not infallible) position to understand human intentions and their derivatives, such as artifact, institutions, and intentional descriptions. This does not mean that there is anything inherently special about being human. Only that we have a privileged intuitions about human affairs simply because our conceptual understanding is a product of our sense-making activity, and our intentional terms get their meaning from our own human capacities and practices. Bear in mind that this does not mean that we always have explicit knowledge of what constitutes the proper use of propositional attitude terms. For example philosophers were content for two millennia that "knowledge" means justified true belief. However, what "justified" amounts to is hard to define and after Gettier (1963), we have been less sure whether this is correct analysis to begin with. The point is only that the foundation of our conceptual understanding rests on our local practices, and living a human life in human cultures arms us with experience that makes us intuitively understand human intentionality even if we sometimes find it hard to explicate it analytically. Concerning other lifeforms, we simply do not enjoy such luxury. Therefore, generally the least controversial way to understand the behavior of non-human agents is either to treat them in merely causal terms or treat their intentional ascriptions as instrumental, because we do not have better intentional vocabulary than commonsense propositional attitudes, which receive their intuitive contents from our own human lifeworlds.

It is worth noting again that our insights into other people's intentional realm are also somewhat limited by intercultural and intersubjective differences; that is why underdetermination problem of content is not strictly limited to human/non-human difference. This is also obvious when one considers conceptual and methodological problems in developmental psychology. Now, because I do not believe in conceptual universality, it serves my purposes well

if we can formulate our theory of non-linguistic content as an empirically adequate theory of *human* intentionality. This helps a lot, because we are licensed to resort to actual (or hypothetical) mechanisms in human cognition while the in-principle counterarguments based on content underdetermination should not be cause for insurmountable concerns. At least, we should be able to tolerate underdetermination as far as we can tolerate miscommunication, intersubjective instability of meaning, and indeterminacy of cross-cultural translation.

The enactivist framework relieves us from the burden of devising philosophically deep conditions for operative intentionality. Some kind of goals and intentional behavior are already presupposed in the notion of an active organism which is, I believe, philosophically quite non-problematic. An active organism does not need an external authority to define its goals and conditions of success. The situation is somewhat different with social behavior where shared goals and procedures are expected by others, which provide grounds for norms and hence the conditions for communally shared intentional descriptions. Because non-linguistic intentionality may be inexplicable and lack intersubjective normativity, some might prefer to refer to related cognitive content as sub- or non-conceptual. Although this choice of terminology might respect some reasonable pretheoretical intuitions, I do not believe that there is a clear distinction of putative subconceptual and bona fide conceptual content. Linguistic meaning and norms of commonsense rationality are vague enough to make the implicit/explicit demarcation somewhat a moot criterion, and I think there is no sharp boundary of discursive and other practices. Intentional contents, including linguistic concepts, are primarily devices for action whether they are explicable by reflection or not.

This observation does not require any deeper insight than appreciation of practical syllogism. If paradigm cases of intentional contents are putative referents of the terms appearing in practical syllogisms, then generally contents are somehow related to action because they often appear as conclusions in the syllogisms. To pragmatists, it is clear that through action norms, discourses, and eventually referential use of language are engendered. Explicit referential use of language does not represent the core but a limiting case of concept use. The etiology of explicit mental representation and the putative subconceptual content that it carries should be ultimately built into any theory of concepts and linguistic meaning. Therefore, it is at times justifiable to speak of discursive contents (or discursive concepts); however, it unwise to consider other mental contents as categorically distinct and non-conceptual. After all, if the

theory advanced here is correct, even discursive contents are brought about by mostly implicit practical abilities.

To deter misunderstanding, a note on my use of the word "discursive" is in order: Linguistic and other communicative acts are perhaps the standard way to understand the term "discursive"; however, my use of the term is more encompassing. Any activity that connects to the nexus of cultural practices, especially practices of giving and asking for reasons, can be taken as discursive. For example, gathering and interpreting evidence for scientific inquiry, are clear cases of discursive acts in my terms. The notion of discursive practice does not presuppose the presence of other people. If you are studying mathematics alone, you are engaging in discursive activity.

I presume that the notion of successful action is not philosophically problematic and hence require dedicated analysis. Still, the account of pragmatic knowledge that I need for my purposed should be able to account for failure and also misunderstanding. This requires kind of rational analysis of behavior; however, we need not formulate its details prior to research mainly because the analysis pertains to the interaction of the organism and its environment, which requires an ecological understanding of the empirical constraints of the cognitive system and the structure of its practical environment. Hence, any *a priori* analysis should be considered tentative rather than authoritative. While the remarks in this and the two previous paragraphs primarily concern individual non-discursive behavior, the desiderata for pragmatic knowledge does not strictly cut the linguistic/non-linguistic nor social/individual demarcations because there are no such strict divisions. Our biology, bodily constitution, individual dispositions, and communal practices are all genuine determinants of meaning. The implications of this will be elaborated later. For now, it suffices to remember that linguistic practices are rooted in non-linguistic practices, and our sense-making process is guided by interaction with both social and non-social environment. The social environment is constituted by real causal and material processes and hence nature and culture do not constitute two fundamentally different realms. In conclusion, we better avoid needlessly wasting effort in seeking distinct cognitive basis for reflective linguistic thinking and for non-linguistic practical cognition.

3.2 Category theories in cognitive psychology

Concept research in cognitive psychology covers a diverse range of phenomena such as categorization, induction, cognitive taxonomies, word meaning, com-

munication, conceptual combination, and cognitive development (e.g. Murphy, 2002). This variety reflects the centrality of the notion of concept in almost all higher cognitive processes. Still, the core of the research is heavily focused on categorization perhaps mainly because of the idea that we do not have direct access to reality but concepts act like filters that bridge the mind–world gap. To interpret, that is to see something *as* something, is to categorize, and the fundamental constituents of (propositional) thought are category representations. This is the heir to representationalism. Some authors have even suggested that cognition fundamentally *is* categorization (e.g. Harnad, 2005). This insight can be construed as misguided, trivial, or profound; depending on how categorization, concepts, and even the notions of the mind and world are conceived. Either way, categorization allows us to treat distinct entities as equivalent in some sense, enabling inductive generalizations and accessing knowledge, and ultimately systematic reasoning and behavior. These jobs are often attributed to concepts and hence, wisely or not, categorization and concepts are often treated as synonymous in cognitive psychology.

During the late 20th century, two major shifts occurred in concept research (Medin, 1989). The first was the abandonment of classical definitional theory in the late 1970s in favor of similarity-based prototype and exemplar theories. Although the two are often pitted together, seen as differing mostly in detail, they make very different assumptions about memory and cognitive processing, and hence are discussed separately below. The second shift happened about a decade later and focused on how concepts are organized as theories.[68] The account might look like a step back towards classical ideas in philosophy of language, for it can be construed as a species of inferentialism and amenable to classical view; however, the framework is not committed to symbolic representations or logical computations and it focuses on causal rather than analytical knowledge in concept representation.

As things stand, no canonical version of the knowledge account exists. While there are detailed quantitative models for both prototype and exemplar theories, none of them explain all the data. The three theories combined accommodate known categorization phenomena extensively, even though it is often unclear which theory fares the best. As diagnosed by Machery (2009), the

[68] After Gregory Murphy (2002), I will use the term "knowledge account" to denote this account of concepts instead of the more common "theory–theory". The terminological choice is not merely stylistic but reflects the conviction that not all causal, taxonomic, and other ostensibly declarative know-that need to be mentally represented in (quasi)linquistic format; see also Murphy & Medin (1985), a seminal paper on the subject.

reason seems to be that these theories are all adequate but limited in scope; they cover limited aspects of human concept representation and processing, and hence they are perhaps best viewed as complementary rather than rivals. Another factor which may limit their applicability is that they are intended as general accounts of concepts; however, factually, they are specifically developed to explain specifically categorization. There are indications that category structure is partly determined by concept use (Barsalou, 1987, 1991; Sloman & Malt, 2003; Markman & Ross, 2003; Christensen et al., 2016), and therefore the data we have might reflect more faithfully our experimental paradigms than the intrinsic nature of conceptual cognition. In other words, the limited empirical adequacy of similarity-based theories, in particular, might emanate from reliance on an ecologically invalid methodology that overemphasizes categorization, often with highly artificial stimulus sets while dismissing the actual *raison d'être* of concepts, which is contextual reasoning in real-world situations. Given this, certain philosophers' reservations about the adequacy of these theories as proper accounts of concepts might look slightly less reactionary.

Besides these three accounts, there actually is a fourth major, albeit still emerging, branch of concept theories, often called neo-empiricism[69], that takes concepts to be constructed from of modal components, namely motor schemata and especially sensory representations instead of symbolic or other *a*modal constituents. While rather radical (and quite compatible with enactivism), these developments can be seen as refinements of older accounts rather than completely new stand-alone theories. Also, there may be an emerging consensus on the canonical expression of knowledge account, namely causal models with Bayesian inference (Holyoak & Cheng, 2011). In the next chapter, I will discuss selected neo-empiricist and Bayesian theories to produce an account of knowledge representation, which is influenced by both of them and incorporates elements from similarity-based theories. The following brief review of the three major research programs aims for conciseness rather than completeness with a focus on details relevant to my argument. For more extensive discussion the reader is advised to consult an excellent (albeit slightly outdated) book length summary by Gregory Murphy (2002) or any standard textbook on cognitive psychology.

[69] Perhaps the best-known references are Barsalou (1999) and Prinz (2002); see Machery (2007) for a critical review.

3.2.1 Prototype models

There are several prototype models, but for our purposes it is enough to understand their general outline. The central postulate is that a category is mentally represented by a *prototype*, which contains summary information about the category. The prototype can be thought of like a caricature that need not represent any actual member but rather a combination of characteristic features associated with category members. For example, the category *bird* might be encoded as a set of features such as *has a beak*, *can fly*, *is a non-mammal animal*. Often the features are assumed to be somewhat vague and may consist of shapes, behaviors, affordances, and other qualities that can be more or less clearly present in its various instances (Rosch et al., 1976). Importantly, the theory does not postulate that there must be any specific set of features that all the members share.

A focal feature of category representation in prototype theory is a *graded structure*, which determines the distance of each instance from the prototype. The more features an instance shares with the prototype, the closer to the prototype it is. The graded structure also pertains to non-members: Chairs are better examples than butterflies as *non*-members of the bird category (Barsalou, 1987). Categorization is based on the assessment of proximity to the prototype: If a target falls too far from the prototype along this similarity gradient, it is judged to be either a poor member or a non-member; if it is close enough, it is judged to be in the same category as the prototype. Very similar members to the prototype are generally judged to be more typical of the class than the more distant ones. Hence, the distance to the prototype is often conceptualized as subjective typicality of the target as an exemplar of the category. (Rosch & Mervis, 1975)

While closely related studies were conducted already in the 1960s (e.g. Posner & Keele, 1968), the prototype theory was refined to its current form largely by Eleanor Rosch and her colleagues during the 1970s (e.g. Rosch, 1973, 1975, 1978; Rosch & Mervis, 1975). In her early studies, Rosch (1975) asked 200 subjects to rate the typicality of 60 items in several categories on a 7-point scale. For example, in the *furniture* category *chair* received an average score of 1.5 (most typical), *telephone* scored 6.68 (most untypical), and *lamp* 2.94 (intermediate). In the early analyses, very high (above 0.9) within subject correlations were reported (Armstrong et al., 1983; Rosch, 1975); however, according to Barsalou (1987) this resulted from a flawed methodology, and a corrected analysis gives more moderate agreement averaging around 0.45. This

has little significance, though, because it does not put into a dispute the existence of graded typicality structure. The reason that this *might* matter is that if there is zero correlation of typicality rankings between subjects, one might legitimately question if typicality plays any role in categorization, given that there is some (and often high) intersubjective agreement on category membership. However, what matters is that these rankings were later shown to predict several behavioral patterns, such as reaction times, error rates, and ease of learning in classification tasks.

In prior studies, it was already established that subjective typicality rankings correlate with reaction times in classification (Rips et al., 1973). However, Rosch could show that the graded structure of superordinate categories (e.g., *furniture*) also affects in predictable ways how its subordinate instances (e.g., *chair* vs. *lamp*) are processed in perception. The study suggested that category representations contain more information about instances rated as good rather than bad examples of the category, and, therefore, category processing has a *central tendency* that makes the category processing more efficient with instances that are more similar to its prototype. Moreover, the study concluded that the effects were not solely due to perceptual processing. Nevertheless, because perceptible features play an important role in typicality estimates, category representations may be somewhat more like images than symbols.

The central tendency is also seen in category learning. People tend to learn typical instances sooner than atypical ones and learn faster to discriminate categories with fewer overlapping features (Rosch & Mervis, 1975). Therefore, the central tendency does not only reflect the subjective conception of already acquired concepts but also influences the very category formation. In addition to learning and recognition, typicality influences inductive inferences to the extent that it can result in violations of extensional logic (Tversky & Kahneman, 1983; Osherson et al., 1990; Shafir et al., 1990). For example, American college students generally find robins a more typical bird than ostriches, and they are more willing to infer that *all* birds have an ulnar artery than that ostriches have an ulnar artery, given the premise that robins have one. This means that typicality is not only a categorization phenomenon but it also affects concept use. These studies show that categorization and category-based inference is not based on knowledge of defining features but on a mostly automatic assessment of similarity to an implicit prototype. Categorization principles are often opaque to the subject, and categories may be idiosyncratic and vague. Often when subjects are asked to define a concept, they produce generic descriptions that are generally, but not always, true of its members. (Hampton, 2006)

All in all, the impact of typicality on category processing is extensive and robust; even to the extent that, according to Gregory Murphy, if an experiment fails to reveal any effects of typicality when comparing category members, it is a good reason to suspect that there is something wrong with the experiment (Murphy, 2002, 23). This is quite devastating for the classical view of concepts in psychology. Arguably, only similarity-based theories can naturally explain central tendency, sensitivity to family resemblance, and related typicality effects. The core of the classical view is that conceptual content is determined by a defining set of features, and that by instantiating those features, all members are equivalent as category instances. However, if characteristic features really are constitutive of conceptual contents, it turns out that meaning (or intension) is continuous rather than discrete and not all category members are equivalent.

The following two examples show how this works.

1° The *body mass index* is defined by $weight_{kg}/height_{m}^{2}$ and classifies people as *normal weight* if the value of the index is in the range of 18.5–25. Thus, the category is defined by two features with continuous values. The index is not supposed to be a psychological model and does not contain a singular prototype. Instead, it defines the concept with a numerical range. It clearly has a graded structure, however, and the prototype could be defined as an (idealized) person with an average height and ideal weight defined by the index. The distance of actual persons from the prototype increases as the function of both height and weight, but as long as the combination of values stays in the predefined range, the person gets classified as normal weight. This is an example of a category that is defined by a two-dimensional continuous feature space with a strict category boundary. The classification is based on a graded structure along these two dimensions.

2° DSM-5 defines the *generalized anxiety disorder* (GAD) as excessive anxiety that is associated with three or more from the list of a total of six symptoms, f_1 =*restlessness*, f_2 =*fatigue*, f_3 =*difficulty concentrating*, etc. (American Psychiatric Association, 2013, 222). None of the symptoms are necessary for diagnosis, while any three of them is sufficient. Therefore, two persons can be classified into having GAD while satisfying no common features besides the underlying anxiety. Hence, there is no underlying "essence" of the disorder but a cluster of related symptoms. The prototype of GAD can be defined as the presence of all the six symptoms, and we get the distance measure by subtracting the number of symptoms from the total of six in the prototype. A person gets classified as having GAD if there are at least three of the symp-

toms present, and anyone with fewer symptoms is too far from the prototype to satisfy the classification.

These two examples are not based on psychological models but artificial definitions for medical classification purposes. Hence, we have to stipulate the prototype for both. This is as it should be, however, because the prototype is not supposed to be an intrinsic property of a category but its implicit mental representation. Yet, the category definitions in these examples arguably track important properties of our conception of normal weight and pathological anxiety. Of particular significance here is the graded similarity underlying classification.

The category structure in example 2° is called *family resemblance* after Ludwig Wittgenstein, who remarked that ordinary concepts are often structured this way. The general idea is that the extension of these kinds of concepts can be formed by chains of the close resemblances between instances (Wittgenstein, 1953, §66 onwards). The name comes from the idea that in a big family we can supposedly find a closely resembling member y for any member x, and again a third member z that resembles y but not that much x. Eventually, we may find two individuals who share no common features at all but who are adjoined by a chain of resembling relatives. Along the way, some features are adjusted, some added, and some removed.

If the prototype is defined as a collection of all the features that are important in determining category membership, it may differ from *statistically* typical category member. It may actually present a rather strange case. For example, Wittgenstein used the concept of *game* as an example of a family resemblance. Bridge is like solitaire to the extent that both are played with cards; bridge is like football because both games are played in opposing teams, and so on. But what sort of a game would have all the characteristic features of all possible games? The theory does not claim that prototypes always need to be formed in precisely this way, however. It allows an abstraction hierarchy where a concept is split into several subdomains with each having their own prototype. The higher-order prototype may not have all or any attributes common to its subordinate categories (Rosch et al., 1976). In other words, games (like many other categories) form subordinate conceptual clusters where instances resemble each other more than any other instance from a different cluster (e.g., ball games, card games, board games, etc.); and the superordinate *game* prototype does not necessarily need to contain attributes such as "played with cards," and so on. In any case, the crucial point is that the prototype does not generally represent an average or the subjectively most familiar member

of a category but an aggregate of features that are characteristic of the whole category. The more abstract you go in the taxonomic hierarchy, the less you will find commonly shared features. (Mervis et al., 1976; Barsalou, 1985).

In prototype theory, the family resemblance structure is taken to be a generally inherent property of categories.[70] To understand the formal properties of prototype representations, it often helps to think of category instances as vectors; i.e. basically as lists of numbers. Binary values can be used to represent the absence and presence of features. Take the definition of GAD for example. We can present list of the six symptoms as a vector of features $[f_1, f_2, f_3, \ldots, f_6]$ where f_1 =*restlessness*, f_2 =*fatigue*, f_3 =*difficulty concentrating*, etc. For example, someone with the three mentioned symptoms and no others would be described with the vector $[1, 1, 1, 0, 0, 0]$; 0 marks the absence and 1 the presence of a feature. The prototype is a vector with all the symptoms present (i.e., $[1, 1, 1, 1, 1, 1]$). Instances $[0, 0, 0, 1, 1, 1]$ and $[1, 0, 1, 1, 1, 0]$ are also in the category while $[0, 0, 0, 0, 1, 1]$ is not because it has only two symptoms present and, therefore, lies too far from the prototype. The feature space is six-dimensional and discrete. It is straightforward to compute the distance measure: subtract any vector from the prototype and take the sum of the remaining values. If the sum is at most at the threshold (which is 3 in this case), the instance represented by the vector counts as a category member. This construction explains the psychological propensity to recognize family resemblance structures not as an ability to construct plausible similarity chains from member to member but as a comparison of targets to the underlying prototype (Rosch & Mervis, 1975). This makes categorization an essentially associative pattern recognition and enables inductive inference without explicit deduction: If the values (i.e., presence or absence) of some features are unknown, their default values can be retrieved directly from the prototype.

Now, one might question if a strict threshold is psychologically or conceptually sound. Consider the examples 1° and 2° above. Do we take normal weight or pathological anxiety to have strict boundaries? Also, what justifies placing it exactly at three symptoms in GAD, for example? Technically, this is not necessary. Think of the decision boundary as a binary function c that maps instance x to a category member; i.e., $c(x) = 1$, if the distance to prototype is less than a predetermined threshold θ. If x is too far from the prototype, then

[70] To avoid confusion, note that the family resemblance structure is supposed to be an intrinsic property of categories while the prototype is not. The prototype is an internal mental representation of the former which, together with a similarity sensitive pattern matching mechanism, enables us to recognize the external graded or family resemblance structure.

x is a non-member (i.e., $c(x) = 0$). For psychological modeling, we may c to be linear instead of binary, for example. Then there is no strict decision rule, e.g., for anxiety but a six-point scale where the score of 6 definitely indicates the presence and 0 the absence of GAD and in total there are 64 combinations of features that represent cases with varying degrees of category membership. We might also want to allow the features themselves to take graded values (like real values in the normal weight example), say, from the normalized range $[0, 1]$ (from definitely not present to definitely present) and add the values to get the instance's distance from the prototype in a real-valued 6-dimensional feature space. The characteristic function may take forms other than linear or binary. The only restriction is that its value always decreases as a function of distance to the prototype.[71]

We can generalize these remarks further. It seems reasonable to assume that some features are more important than others. We can implement this by associating each feature with a coefficient called *weight*. Take a category that has three characteristic features f_1, f_2 and f_3, and assign a weight w_i to each. Thus, we get a vector $[(w_1 \times f_1), (w_2 \times f_2), (w_3 \times f_3)]$. We allow weights to have negative values, meaning that the presence of a feature is a contraindication for category membership. Now, let θ be a *threshold value* determining a decision boundary: If $(w_1 \times f_1) + (w_2 \times f_2) + (w_3 \times f_3) \geq \theta$ then $[f_1, f_2, f_3]$ is a category member. That is, a target counts as a member if the sum of the weighted features exceeds the set threshold, which determines the critical distance from the prototype. This is perhaps the best-known prototype model, called the *independent cue model* (Medin & Schaffer, 1978; Hampton, 1993). The name comes from a classification scheme, which is sensitive only to the presence of features. It detects if there are enough categorization cues present and disregards their possible structure (Rosch & Mervis, 1975; Tversky, 1977).[72] For example, the *bird* classification might be sensitive to the presence of *wings* and the *ability to fly* but not to the causal relation of wings and flying. This kind of mechanism is a paradigm example of what *associative processing* means in modern cognitive science.

The readers familiar with artificial neural networks should readily notice that the independent cue model is formally equivalent to the simple perceptron (see Rosenblatt, 1958, 1962). More generally, the related systems are called *linear classifiers* because they classify instances based on the linear combination

[71] See Osherson & Smith (1981) for a similar but more detailed formal definition.

[72] But see Hampton (1995) and Smith & Minda (2000) for arguments that this simplest feature addition models do not fit well to all categorization data.

(i.e., basically an addition) of features. Simple linear classifiers have severe limitations; however, they are used extensively for pattern recognition and other tasks in machine learning because, in practice, they are fast to execute, often easy to train, fairly effective in classifying data in rich dimensional feature space, and can learn to recognize features and categories that defy clear definitions. All these properties are also commonly attributed to human intuitive judgment.

While linear classifiers treat features independently, the prototype theory does not claim that property correlations are unimportant. Quite the contrary, the prototype theory assumes that the *environment* is highly structured in the way that certain properties combine frequently (e.g., wings and flying) while others do not (e.g., fins and walking), and this empirical fact makes linear classifiers highly effective despite their inherent theoretical limitations. The human categorization system picks out these stable environmental combinations and makes the basic category cuts to maximize the *cue validity* of features. Cue validity of feature f is its diagnostic value in respect of a given category c. It is defined as the conditional probability of x belonging to a category c, given it has a feature f; i.e. $P(c(x)|f(x))$. If $P(c(x)|f_1(x)) > P(c(x)|f_2(x))$, then feature f_1 is more diagnostic for category c and hence bears more weight. For example, *feathers* is highly diagnostic for birds because all birds have them and especially because other animals do not. *Flying* is somewhat less cue valid. The features that the category prototypes consist of are those with high cue validity. As a result, typical items share many properties with other category members and fewer properties with non-members.

Note that cue validity measure is contextual because it depends on the assumed contrast classes (Tversky, 1977), meaning that cue validity is not an intrinsic property of a category or a feature but depends on how features are distributed in the environment. The category partitions supposedly reflect the environmental correlations of properties, and classes that do not have salient central clusters make poor categories. (Rosch & Mervis, 1975; Rosch, 1978) Again, means that instance typicality is not a function of instance familiarity or statistical typicality. Instead, psychological typicality tracks *central tendencies* of property clusters present in the environment as perceived by the organism.

Behind these technicalities hides an important point. If we think of prototype mechanisms as linear classifiers, the actual prototype or category boundary does not need to be represented in the system. The system is only sensitive to an array of features, and their combination does not represent the features. The summation generally wipes out all the information about the

input. Hence, the system can perform classification in the absence of any sort of representational capacity traditionally understood.

While the allure of the prototype theory rests mostly on its ability to handle inherently vague concepts, linear classifiers can also accommodate many concepts with clear definitions. For example, GAD has a family resemblance structure but it can be captured by a logical rule. The complete definition, according to DSM-5, can be spelled as "A: A presence of an excessive anxiety that is difficult to control, with B: at least three of the aforementioned six symptoms (call these features $b_1, \ldots, b_6$), and C: the disturbance is not better explained by substance abuse or by other mental or physical disorder." To see how an independent cue model can capture this definition, we define a feature vector: $[A, b_1, b_2, b_3, b_4, b_5, b_6, C]$, associate each feature with respective weights $[7, 1, 1, 1, 1, 1, 1, -4]$, and set the threshold $\theta = 10$. Now, if the presence of a feature is represented by 1 and the absence by 0, we can easily verify that the model exactly computes the definition of GAD (here C stands for the *presence* of substance abuse or alternative disorder). An instance cannot exceed a score of 10 if feature A is absent (a necessary condition) or if less than three of the features $b_1, \ldots, b_6$ are present. These constitute a jointly sufficient condition for classification, *given that* C has the value 0. With value 1, the coefficient -4 of C prevents the total score from reaching the threshold.

Regardless, many classification problems need recursive rules or other methods that can transcend the inherent limitations of linear classifiers (see Minsky & Papert, 1988), and hence these systems do not work as a straightforward generalization of definitional account. For example the DSM-5 (p. 160) criterion of major depression consists of a linearly nonseparable set of symptoms. In fact, linear classifiers cannot compute as simple operation as the exclusive disjunction; i.e. either x *or* y, but not both x *and* y. Formally, the problem can be surpassed by chaining linear classifiers to form what are essentially multi-layered neural networks; however, these models are too unconstrained. The resultant models can approximate any computable function (Hornik, 1991; Siegelmann & Sontag, 1995) and hence as such they hardly constitute a very informative theory of concept structure. Nevertheless, mental representation of linearly nonseparable categories may not require linguistically structured and rule-based representations; however, they do necessitate systems that can do more than mere feature additions (e.g. Marcus, 1998).

3.2.2 Exemplar theories

Prototype comparison is not the only way to implement similarity-based categorization. Perhaps the most obvious way is to introduce a set of examples and use them as a standard for category judgments. Category theories that take this approach are called *exemplar theories*. The general idea is that whenever we classify an object, we tacitly memorize it as an *exemplar* of that category. Category representation is based on stored exemplars instead of abstraction of generic information, such as definitions or prototypes. Categorization is done by comparing novel stimuli to exemplars in each category. In the context of exemplar theories, we can retain the essentials of the above discussion about similarity-based categories, and exemplar and prototype theories make mostly similar behavioral predictions. Hence, they are often discussed together as an alternative frameworks to account for the same categorization phenomena. Still, they make substantially different claims about the underlying cognitive psychology, and their predictions differ in some critical points.

As with the prototype account, the exemplar framework has several detailed models. They all share the core idea that category representation consists of a set of concrete exemplars that are grouped under the same label. As in the prototype theory, instances are associated with a set of n features, and any stimuli can be thought of as occupying a point in n-dimensional feature space. However, instead of having a central prototype, category representations consist of a set of points with none having any special status. A novel stimulus is mapped onto this feature space and the categorization process compares it to stored exemplars in its neighborhood. There are many ways to do this. For example, the stimulus can be classified based on the nearest neighbor, accommodated to the category with most matching exemplars (inside a predefined distance), and so on.

Perhaps the most famous exemplar model is the so-called *context model* (Medin & Schaffer, 1978) and its generalization (Nosofsky, 1986). Below, I will mostly limit the discussion to the original context model. Whatever account one chooses to employ, it is important to appreciate how radical this approach is. Prototype theory attacks the assumptions of necessary and sufficient conditions, the equivalence of category members, and bivalent extensional logic underlying conceptual structure. However, the prototype theory does not question the idea that category representations are summary descriptions of their extensions. Exemplar theory goes further and dismisses this assumption. It claims that all the information of a category is encoded in the representa-

tion of its specific instances. It should be clear that exemplar representations generally differ from prototypes. Prototypes are lists of independent features, but, in exemplars, features are always correlated because they are correlated in actual objects. Therefore, in contrast to independent cue models, these systems are often called *interactive cue models* because they are sensitive to the co-occurrence of features.

On the theoretical side, one might wonder what good exemplar-based category encoding could do in principle for the whole point of categorization is to abstract away from individual instances. But consider concept learning: If you do not have any previous understanding of a category, pretty much all you can do to categorize novel stimuli is to rely on reminding them them of prior examples. After seeing a few instances, you are in a better position to evaluate what the common characteristics of the category are based on their shared features (Posner & Keele, 1968; Medin & Schaffer, 1978; Ross et al., 1990). If an understanding of the important properties of a category is lacking, it is reasonable to store the instances in rich detail because features that may prove important later can then be readily retrieved. This strategy reduces the risk of making premature, biased commitments on the nature of the category (Brooks, 1978).

Moreover, exemplar-based systems readily lend themselves to analogical reasoning. If a novel entity x is grouped with familiar exemplar y, then knowledge about y can be used to reason about x. This sort of interchangeability of instances is the hallmark of conceptual thought. With instance-based reasoning the problem then comes to select the relevant information about y to support category based inference, because exemplars are encoded without highlighting what information is relevant to the category. This might appear as shortcoming because selecting the appropriate information for further inference is one of the key problems that categorization is supposed to solve, but assessment of relevance is a problem for any reasoning or category theory anyway, and it is at least not evident if the contextually correct categorization of an entity is different problem from choosing the relevant information about it for the task at hand. In short, much can be done with instance-based category representation, at least in principle. Hardly anyone deny that people remember and recognize previously encountered things; nonetheless, the question is, are there any reasons to support the idea that category judgments are based on exemplars? A considerable body of data suggests that there are.

Oft-cited exemplar research comes from Lee Brooks, Norman Allen, and collaborators. In one study they investigated medical residents and experienced

practitioners and showed that their ability to visually classify skin diseases were affected by the examples of lesions that they had seen previously. Considering that medical diagnostics is a learned expert skill, it should appear evident that it is affected by learning; however, the study revealed several phenomena that are significant to the psychology of categorization. The effects of similarity were compulsory and lasting. Even though subjects had diagnostic rules available, they still classified on the basis of similarity to previous exemplars, which led to misclassification with clinically significant error rate. Moreover, the classification was sensitive to the *correlation* instead of mere addition of features, contradicting the independent cue model. Indeed, dermatological diagnostic rules are often based on a summation of evidence, constituting a family resemblance structure like with GAD. (Brooks et al., 1991) Similar results were discovered earlier in a related study that used artificial categories based on simple additive classification rule and clearly identifiable features (Allen & Brooks, 1991). Hence, the misclassifications imply a categorization strategy that is different from both explicit rule following and prototype matching, and the effect was not merely due to the complexity of the stimuli or vagueness of the features. In dermatologist study, the effect was not limited to novices but was also apparent with experienced practitioners. It was still present two weeks after the introduction of the learning stimuli, demonstrating that the exemplar effect is relatively durable. Also, an earlier study found imperfect but relatively impressive retention of dot pattern stimuli after 10 weeks of delay (Homa & Vorsburgh, 1976), and according to Thomas Palmeri (1997, 324), the effect lasts for years for highly similar stimuli.

An interesting feature of the dermatology study was that the target categories were natural skin diseases and not artificial stimulus sets. One problem with the comparison between prototype and exemplar models is that while they make substantially different assumptions about category structure, their predictions are often empirically almost identical in ecologically valid situations. Rosch and Mervis' (1975) findings show that typical members of natural categories generally share more common features than atypical ones. Therefore, most exemplars tend to cluster around category prototypes. This is because if the characteristic features of common kinds are highly correlated, it simply means that most instances of these categories exhibit many of the prototype features. Therefore, the point in the feature space occupied by the prototype is close to many exemplars while there are fewer exemplars close to the category boundary. The exemplar theory explains the typicality effects as reflecting the amount of exemplars in targets' proximity. With categories that have high

family resemblance scores the density of exemplars is basically a function of distance to the prototype. Hence, exemplar and prototype theories imply effectively identical predictions. This is why critical tests of the theory generally employ artificial, hard-to-learn categories with low family resemblance scores. (Murphy, 2002, 95–96) The methodology may compromise the ecological validity of the research, but the categorization data we have considering artificial categories, provide good support for exemplar models.

The tables below depict the so-called 5–4 category structure which was used extensively in the studies that led to the wide acceptance of the exemplar theory. It was introduced in a seminal paper by Medin and Schaffer in 1978 and it was employed at least in 30 experiments in the following two decades (Smith & Minda, 2000).

Category A

stimuli	dimensions			
	f_1	f_2	f_3	f_4
a_1	1	1	1	0
a_2	1	0	1	0
a_3	1	0	1	1
a_4	1	1	0	1
a_5	0	1	1	1

Category B

stimuli	dimensions			
	f_1	f_2	f_3	f_4
b_1	1	1	0	0
b_2	0	1	1	0
b_3	0	0	0	1
b_4	0	0	0	0

What you see here are two abstractly depicted stimulus sets from categories A and B. "Dimensions" refers to four binary features in each individual stimulus $a_1, \ldots, a_5$ and $b_1, \ldots, b_4$. In various studies, the categories have been given different specific contents, such as colored geometrical shapes, Brunswik faces (Medin & Schaffer, 1978), or yearbook photos (Medin et al., 1983). Thus, the binary values stand for the absence or presence of perceptible features in each study, such as colors or facial features. The four binary features makes 16 different instances in total. In the learning phase, the participants are exposed to these $5 + 4$ stimuli, and the remaining 7 are used in a subsequent transfer test that probes how the subjects project the category knowledge to novel instances. Note that category A is derived from prototype $[1, 1, 1, 1]$ and B from prototype $[0, 0, 0, 0]$. These categories are quite ill-structured and thus hard to learn, as predicted by the prototype theory. Even if you knew the underlying prototypes you could not infer that $[1, 0, 1, 0]$ is in A and $[0, 1, 1, 0]$ is in B because they are equally far from both prototypes. Initially, subjects just have

to guess and go through several runs of learning trials to memorize the correct classification for each item and extract weights for each feature dimension to guide categorization for new stimuli in the transfer task. See (Smith & Minda, 2000) for an extensive critical review of these studies. I will skip the critical part for now and discuss three phenomena that Smith and Minda considers to be the key evidence for exemplar models cited in the literature.

The first notable effect is the preference for old items during the transfer test. Prototype models assume that information about specific stimuli is not stored. Therefore, subjects should show no better categorization of training exemplars in comparison to transfer items; however, this is not the case. The further the target is from a prototype or the less there is a difference in the number of similar exemplars in competitive categories, the less sure the subjects are about to which category the target belongs to. Since they cannot use a simple classification rule to decide this, they sometimes assign vague training exemplars to the wrong category in forced choice situations. Therefore, the graded structure in each theory implies predictable errors in category judgments. Broadly speaking, both theories make qualitatively identical predictions about the error profiles in the transfer task. However, a detailed analysis will reveal that the predictions differ quantitatively. The prototype theory predicts that in the transfer task, the error rates for the old items are similar to novel items—i.e., a linear function of their distance from the prototype. However, the exemplar model predicts substantially better classification for the old items encountered in the training set. In in comparison to the prototype theory, this is especially true for items far from the category center because targets are matched against stored exemplars in category memory instead of the central prototype; and this is what we actually see in human categorization studies. Unlike the prototype models, the exemplar theory readily explains this processing advantage for old items.

Note that the processing advantage does not make classification perfect for the old items. The exemplar theory explains the emergence of a graded structure partly by a gradual forgetting of the details of stored exemplars. This means that there generally is uncertainty in category judgments and, therefore, the possibility of misclassification, even with familiar stimuli. If the memorization was perfect, subjects should simply use the already familiar exemplar to decide the category. However, if the memory trace is degraded, a familiar exemplar serves just as one more highly similar instance in exemplar memory. Without postulating this sort of forgetting, exemplar theory should predict perfect categorization for familiar stimuli, which is an empirically in-

valid prediction. Forgetting also contributes to the explanation of prototype effects in exemplar theory: In many natural categories, there are lots of exemplars clustered around the prototype. If their specifics are lost, then highly similar exemplars begin to look pretty much the same, engendering a cluster that works effectively like a prototype and causes pronounced central tendency in category processing. (Murphy, 2002, 54–56) In this sense forgetting can be actually adaptive. It may reduce overfitting in category learning and thus promote focusing on relevant dimensions. This sort of forgetting may appear as somewhat an *ad hoc* postulate, but, of course, we often do forget the details of past encounters, hence, it is at least *prima facie* plausible and supported by empirical evidence (see Homa & Vorsburgh, 1976).

The second important effect is more subtle but theoretically very suggestive. Consider items $a_1 = [1, 1, 1, 0]$ and $a_2 = [1, 0, 1, 0]$. Clearly, item a_1 is closer to prototype $A = [1, 1, 1, 1]$, and hence the prototype theory predicts that a_1 is learned more easily and categorized with fewer errors than a_2. However, it turns out that whether this is the case depends on other items in the learning task. Although the 5–4 stimulus sets depicted above may appear quite random, they are carefully selected. Notice that $a_2 = [1, 0, 1, 0]$ is close to two other exemplars in category A: $a_1 = [1, 1, 1, 0]$ and $a_3 = [1, 0, 1, 1]$ but it differs at least by two features from any exemplar in B. Item a_1 is highly similar to only one exemplar in A, namely a_2, but equally similar (i.e., differing in only one feature) from two exemplars in set B: $b_1 = [1, 1, 0, 0]$ and $b_2 = [0, 1, 1, 0]$. Hence, with this particular set of learning items, the exemplar theory predicts a processing advantage of a_2 over a_1. This prediction was verified in (Medin & Schaffer, 1978) and in several subsequent studies, and has been taken as highly important evidence for exemplar theory against prototype models. However, it should be noted that a meta-analysis (Smith & Minda, 2000) revealed that the effect is not robust. The effect appears in several studies but seems to disappear in the more encompassing analysis. Then again, the meta-analysis revealed no sign of a_1 advantage over a_2, either. The exemplar theory at least has theoretical resources to explain a_2 to a_1 advantage when we see it and why it should be sometimes absent. Still, even if the analyzed data sets were chosen in a way that at least ostensibly favored exemplar theory, no predicted effect was reliably found. Naturally, the mentioned theoretical resources are pretty void if the true explanation happens to be simple statistics.

Third, independent cue models predict that linearly separable categories are easier to learn than linearly non-separable ones. This is because linear separability simply means that the extension can be partitioned by a weighted

additive combination of independent feature component information about instances. The prediction, however, was found not to be always true of human category judgments (Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981). Unfortunately there are, again, some caveats with this finding. These studies were conducted with two categories that have very weak structure and were hard to learn. Learning was very difficult despite the fact that the training exemplars were sparse and their dimensionality was small. A subsequent study used four categories with either three or nine learning items and with either linearly separable or non-separable structure (Blair & Homa, 2001). The study looked into the performance of individual subjects and found that they were, in fact, often operating under linear separability constraint. When the complexity of the task was increased, subjects found it easier in the linearly separable condition. With the larger category using nine learning items, the proportion of subjects displaying classification strategy based on linear separation increased dramatically from 9% to 81%. Other results also illustrate that large, high dimensional, and well structured categories promote prototype rather than exemplar learning (Smith et al., 1997; Smith & Minda, 1998, 2000).

For the reader not too fond of the methodological niceties of cognitive psychology, the take-home message is this: Even if prototype and exemplar theories make very different assumptions concerning processing and representation, they are difficult to compare empirically. The differences are subtle, and the critical results are often not very robust. It is clear that the exemplar effects are real and significant. Still, their *theoretical* significance is not completely clear, and explaining human categorization data seems to require prototype theory also. Now, if we are about to resolve the debate between prototype and exemplar models we, should probably seek evidence from other phenomena besides mere category learning. Converging evidence from closely related areas could conceivably tip the scale in either direction. However, we may wish to preserve both frameworks because their straightforward combination seems to explain artificial category task performance quite accurately.[73] In that case, we need to find their proper place in conceptual cognition and specially to explain what the exemplar effects actually mean, given that summary information about categories is already contained in prototypes.

One should note that the exemplar effects are most pronounced with artificial categories that have undifferentiated and difficult-to-learn structures. These categories are not artificial in the sense that they consisting of artifacts but in the sense that they have no apparent content. They are not used for

[73] For more detailed discussion see Machery (2009), and Smith & Minda (1998, 2000).

anything, save for their role as reaction probes in forced choice tasks. The categories employed in these experiments are very unnatural in all respects, which has clear impact on their cognitive processing. Often we get the point of common natural categories pretty swiftly, but with artificial categories it may take tens of runs through sparse learning items to obtain about a half of the subjects to classify correctly (Murphy, 2002, 103ff.). Compared to natural concepts, these are special sort of learning episodes altogether because *there is absolutely no point to get* with artificial categories. I explained earlier why I think it is a misguided worry that psychologists working with similarity-based categories are not actually studying concepts; however, this methodological aspect of exemplar research should raise some red flags. The categories in these studies are hardly concepts since they do not have any content or sense. The best evidence we have for exemplar theory is thus derived from unnatural hard-to-learn categories where the list of their extensions is pretty much their simplest description.

Usually we categorize to reach some other goal. That is the whole point of having concepts in the first place. Making a correct category judgment can be an end itself (e.g., in some games, exams, and psychology experiments), but categorization as such is only one particular and austere way to interact with category members. Category knowledge and other associated cognitive content are shown to be determined by interactions with category instances which, in turn, affect further categorization—as we will see in the next section and the following chapters. Generally, category identity is learned as a *by-product* of the ways concepts participate in these interactions and category learning is not the actual goal of the actions unlike in laboratory learning tasks (Markman & Ross, 2003). The upshot is that artificial categorization studies largely dismiss the actual point of conceptual cognition and the methodology produces settings where category use happens in very unnatural tasks with unnatural category structures. Hence, the data from the experiments may mostly mirror the experimental design rather than the cognitive psychology of conceptual representation.

In conclusion, the exemplar effects are there and they do affect categorization; however, this does not mean that they necessarily have a lot to do with how categories are represented in general. If you look at these studies, it appears very much like as if the subjects are gradually learning very specific, exemplar-based skills for responding appropriately to ill-structured stimulus environments. In fact, mainstream cognitive psychologist think that skills are learned precisely this way, that is by accumulating a vast, implicit knowledge

base about the specific instances of the domain. This is basically what training is supposed to do.[74] However, if the amalgamation of pragmatist inferentialism with enactivism is the correct theory of human conceptual understanding, then skills are fundamental in making sense of the world. That is, if exemplar effects reflect skill learning and concept use is a kind of a cognitive skill, then exemplars do have a profound place in conceptual understanding. However, to appreciate their significance to conceptual cognition, we need to consider them in a wider context than pure categorization with meaningless stimuli.

3.2.3 The knowledge account

The second major shift in concept research began in the late 1980s. As remarked earlier, this third paradigm lacks a canonical formulation to date, but its core claim is that concepts are organized as theories. As such, this may sound like a step back toward traditional ideas prevalent in analytic philosophy. For example, conceptual role semantics is one way to explain what it means for concepts to be structured this way. We do not need discuss this in-depth again, but recall that the problems with CRS were related to identifying intentional content under the assumptions that conceptual representation is veridical and referential in the sense of being universal and respecting bivalent extensional logic. The psychologists' claim here is substantially weaker, in that knowledge account only claims that conceptual contents carry inherent inferential potential, which often dictates performance in category learning and use. No claim is made that conceptual content in full can be reduced to this aspect of category representation. It is pretty clear that raw similarity matters; however, it is equally clear that sometimes the similarity constraint is violated.

A casual observation seems to confirm this. We count whales as mammals, although they ostensibly are more similar to fish. This as such is not very damaging to similarity-based theories. It is possible to adjust features so that some of them bear more weight than the others, rendering some of them even necessary for category membership.[75] Hence, beyond surface appearances, there might be more important properties that make whales actually more similar to mammals than to fish. However, his solution risks being vacuous or psychologically unrealistic. If similarity means simply a predicate comparison, we could assign the property *mammal* to whales and postulate that this property

[74] See e.g. Smith & Minda (2000); Johansen & Palmeri (2002) and Murphy (2002, 103–114) on this interpretation of the exemplar effects in relation to skill learning.

[75] For example, we saw this with the linear classifier definition of GAD on page 121.

is necessary and sufficient for something to count as a mammal. However, this is pretty uninformative, and similarity-based accounts should explain how we decide something to be a mammal on the basis of an array of other properties. Then, we may associate whale prototype with properties that are empirically important for being a mammal, such as *live birth* and *lactation*. But, realistically, does the folk really categorize whales as mammals because they know that whales lactate, etc., or simply because they have been told that whales are mammals? (In which case, psychologically, lactation is more likely taken to be an implication rather than a conceptually constitutive property of the belief that whales are mammals.) I personally find the latter option more plausible, and in any case, a good deal of psychological literature indicates that such pieces of explicit knowledge often does dominate category judgments.

One of the most famous experiments demonstrating the dissociation of knowledge and similarity in categorization was Lance Rips' (1989) experiment where he asked subjects to estimate the size of the largest U.S. quarter and the smallest pizza they have ever seen. Then he read his participants a description of a circular object that was halfway in size between these estimates. Thus, if a subject told that the largest quarter she had ever seen was 1 in. in diameter and the smallest pizza was 5 in., the description was "a circular object 3-in. in diameter." Then he asked them to categorize the mystery object as either a quarter or a pizza; or alternatively, he asked them to estimate either the similarity of the object to a pizza or its typicality as a pizza. The study included 36 similar problems where one category was always fixed and the other variable in size. The results demonstrated a dissociation of categorization, typicality, and similarity judgments. For example, about 2/3 classified the item as a pizza, while the same number of participants considered the item to be more similar to a quarter. The typicality estimates were about 50/50. This result does not mean that similarity is unimportant; however, it shows that categorization is not *solely* a matter of similarity, because it can be outstripped by world knowledge.

Obviously, we can keep track of exceptions like we can keep track of, say, one's belongings. Perhaps exceptions, such as that whales are actually mammals, work *psychologically* similarly to any trivial conventional knowledge—i.e., as culturally transmitted explicit factoids. In Rips' study this seems to be the case. Subjects reported that even though the mystery object was more similar to fixed size category, it still could not be its member, reflecting the conventional understanding that coins do not come in arbitrary sizes. In some category judgments similarity still dominates, and this discrepancy may stem

from the arrangement where automatic similarity matching is the default mechanism but reflective processing may intervene and override its output. There is evidence for this kind of dual-processing in categorization. For example, a subsequent study did not find the reported dissociation when the subjects were not required to think out loud or justify their decisions (Smith & Sloman, 1994). The effect was absent even though the participants were working under the instructions which were designed to encourage rule-based inference. Moreover, the subjects classified the object as a coin when it was described as 3 in. in size *and* silver colored, thus apparently switching (or lapsing) from necessary-feature-based analytic processing to similarity- or characteristic-based categorization. This suggests that the assumed similarity-based default mechanism might be quite hard to suppress.

This looks like pretty much what generic dual-process theories claim about cognitive architecture: an associative default mechanism, which is sensitive to concrete perceptible features, augmented with analytic rule based reasoning. However, the real story is not as simple as *prototypes by default + rules for exceptions*. The subjects were clearly doing something else than simple similarity matching, but their responses also differed from deductive inference with explicit definitions. Instead, the participants often seemed to engage in abductive material inference. The following subject report displays this lucidly. The target object was 4.75 high and it was either a stop sign or a cereal box: "It would probably be one huge cereal box, because no one would see the stop sign if it were that small [...] while I wouldn't expect to see a cereal box that big, it wouldn't make sense to have such a small sign." (Rips, 1989, 37)

In one of the seminal papers on knowledge account, Gregory Murphy and Douglas Medin (1985) used a vivid example of abductive categorization: Jumping into a swimming pool in a party while dressed might be strongly associated with with *intoxication*. The event is probably quite rare and hence not very similar to drunken behavior in general; however, the association can be drawn from a plausible abductive inference. Additional information such as there is someone drowning in the pool quite likely changes the conclusion from drunkenness to heroism. In that case, the information about the person in question remains the same, but the classification from a drunk to a hero changes because added contextual factors influence the inference to the best explanation of the behavior. One might protest that this is not an example of taxonomic but event classification or perhaps property induction. Hence, we probably should expect the underlying cognitive processes to differ from taxonomic object categorization tasks. However, the point of taxonomic classification *is*

often property induction, and (in agreement with the objection) the example highlights that categorization going beyond taxonomic classification of *things* is not be reducible to considerations about similarity alone but we need to resort to causal and contextual inference.

The above example shows that world knowledge or lay theories are important for classifying events or behavior but it does not clearly demonstrate that they are essential for taxonomic classification of things. However, there are conceptual domains that are both taxonomic and theoretically constituted. Folk biology is an extensively studied example. Preschool children seem quite susceptible to classify species on the basis of surface features but this tendency changes around the age of 6 or 7.[76] When children mature, they acquire many interrelated beliefs about biological kinds that guide their category judgments, for example, that biological things reproduce transmitting many of their important properties to their offspring, that biological kinds have a complex internal structure, and that biological kinds grow and undergo irreversible changes but it is the less mutable internal structure that determines the organism's kind. Based on these principles, children (like adults) judge kind category of living things as remaining the same even though they undergo superficial changes through surgery, exposure to fictional toxic substances, or mutation, which make them look like another kind of animal. (Rips, 1989; Keil, 1994) What is going on is much more sophisticated than the application of a simple rule. It does not matter whether the inferences are deductive, abductive, or inductive; what matters is that these categorization judgments apparently deploy a *system* of rules, principles, or something similar that have causal or explanatory relevance in determining the target's kind (Murphy & Medin, 1985).

These sorts of *knowledge effects* were decisive in shifting theoretical focus from similarity to knowledge in the psychology of categorization. If similarity refers to the tabulation of observable raw properties, it is too weak a notion to explain all the relevant categorization data. On the other hand, similarity can accommodate functional or other non-obvious properties; however, if it is extended to cover inferential relations in complex knowledge structures, the notion becomes too unconstrained and vacuous to be informative. That would in any case necessitate fundamental changes in the core of the theory to the extent that it would be rather misleading to use the term "similarity" anymore. Still, these results do not entirely invalidate similarity-based theories. The

[76] But see Gelman & Wellman (1991); apparently *something* changes around that age; however, even 3- to 4-years olds can appreciate that surface features are often less relevant for species identity than hidden internals.

emerging picture rather looks like that we have a fundamental tendency to categorize things on the basis of similarity but that gives way to more or less explicit and/or analytic knowledge-based categorization. On the face of it, this makes good sense. Superficial appearances can be misleading but you need to start with something in order to get causal learning going—something to lock causal knowledge onto—and generalizations formed by raw similarity is a good candidate for that something. But the clear-cut two-process story that treats similarity and causal knowledge as distinct layers is misleading. Actually if it were correct, the knowledge account would be somewhat superfluous as a theory of categories. Knowledge would be just a sort of an add-on to already established ontology (apart from perhaps purely formal concepts if there are any).[77] To avoid making this simplification, it is relevant to understand the impact of knowledge that people bring to category-learning situations.

Earlier already Rosch (1978) noticed that the analysis of objects into attributes is often a cognitively sophisticated activity that transcends the mere extraction of features from the environment. For example, the functional attribute "you eat on it" requires knowledge about humans, their activities, and the real world to be understood. Later research with made-up artifact categories confirmed that subjects' knowledge about object functions affects the perception of object properties. For example, one study (Lin & Murphy, 1997) used a rod with a loop and two other gadgets attached to it as a learning stimulus. Half of the subjects were told that the instrument was for catching animals with the loop, and the other half was told that the instrument was for fertilizing the ground. For the latter group, the loop was explained to be for hanging the tool for storage. For the test stimuli either the loop or one of the other features was missing, and the participants were asked to judge whether the novel object was in the same category as the original tool. For the first group, the presence of the loop was critical for judging that it was, while for the latter group it was irrelevant. This, of course, is not very surprising, but the finding shows that the object's function and its parts are not coded as separate features. The subjects deployed their knowledge of the causal relation of an object part and its use to make category judgments. This happened even in speeded test conditions that were designed to suppress explicit reasoning, implying that knowledge affects categorization early in the processing, making functionally relevant features more salient in perception.

Knowledge also influences similarity judgments of category attributes. In Douglas Medin and Edward Shoben's (1988) study, subjects rated *white* and

[77] Recall our earlier discussion about similar points in the context of two-factor CRS.

gray to be more similar than *gray* and *black* in the context of *hair* but reverse judgment was found in the context of *clouds*. Presumably, this was because of the subjects' previously established inductive generalizations that associate gray and white hair with aging and gray and black clouds with bad weather. This effect might be better construed as pertaining to similarity of feature *values* rather than the features themselves. That, anyway, would be an equally interesting interpretation, demonstrating essentially the same point: similarity metric is affected by causal properties associated with the category. Later, Evan Heit and Joshua Rubinstein (1994) tested directly whether similarity-based reasoning is guided by feature- rather than category-level similarity. The underlying assumption was that if category A has property P, then B is considered to have that property also to the extent that A and B are perceived to share other features. They found that no single formal measure of category similarity existed to support such inferences but the subjective understanding of the nature of property P itself determines which other features of A and B are deemed as important. For example, the participants were more prone to make anatomical inferences from bears to whales but behavioral inferences from tunas to whales (Rehder, 2009).

Lawrence Barsalou's research on *ad hoc* and goal derived categories is another important demonstrations of these points (Barsalou, 1983, 1985). He defined the category *ideal* as a feature that category members should have to ideally serve their purpose as category members. For example an ideal *food to eat on a diet* (a goal-derived category) has *zero calories*, and the fewer calories an instance has, the better it serves the goal associated with the category. His main finding was that ideals determine a graded structure that drives categorization independently of family resemblance. In a related study, he showed that people are readily able to form a coherent, graded category out of ostensibly disjoint objects. For example, on how could things like *children*, *a blanket*, *a dog*, and *stereos* form a good category? They are all good candidates for *things to take out in case of a fire*. That is also a readily projectible concept: a wallet is a better exemplar of the category than a cheap and bulky kitchen table. The lesson is that the correlational structure of category features does not always determine the goodness-of-example measure. At least for goal-derived categories, it is determined how sensibly the combination of features satisfies the category goal. In sensible, practical contexts, ill-structured and disjoint clusters may well make good and coherent categories.

The discussion above illustrates that the psychological contents of our object concepts is not determined by a tabulation of features alone but also how

they are embedded in larger knowledge structures that include human goals and practices. These results do not deny that feature-based similarity is important; in fact, most studies clearly confirm that it is. However, it is also crucial that often the features themselves are encoded in a way that highlights their interactions. For example, it is not only relevant to our *bird* concept that birds fly and have wings but also that it is the wings that make them fly (see Ahn et al., 2000b). Therefore, causal knowledge and similarity-based processing hang together quite tightly. In category learning, it does not matter whether categories are phrased in terms of meaningless or meaningful features if they do not form a coherent whole. What matters is that these features are meaningfully related in a way that makes sense to the cognitive agent. For example, formally identical categories phrased in terms of meaningless formal symbols $+, \{, >, \$, [$ are as hard to learn as the same category structure with meaningful features that lack conceptual coherence: *lives alone*, *made in Africa*, *fish kept there as a pet*, *has barbed tail*, *thick heavy walls*. However, if the features form a sensible whole, for example *made in Africa*, *lightly insulated*, *green*, *drives in jungles*, *has wheels*, the category is substantially easier to learn (Murphy & Allopenna, 1994).[78] In the next, chapter we will see that a comparable phenomenon is observed in a conditional reasoning task; that is, subjects are not sensitive to formal versus thematic content in rules but whether they can interpret the rules as expressing a familiar or meaningful relation between variables.

Knowledge effects are prevalent throughout the psychology of concepts ranging from categorization to concept learning, conceptual combination, induction, and word learning. Formal properties (such as linear separability) of category structures might be important in initial learning of a foreign domain,for example, but background knowledge seems to have a significant impact in every aspect of conceptual cognition whenever it is available. Theoretically, this is expected. One main impetus to research in knowledge effects was the question of category coherence (Murphy & Medin, 1985), i.e., what makes some collection of objects or features intuitively sensible wholes and others less so. Similarity as such is logically too unconstrained a notion to be very useful here. The Eiffel Tower and pet fish are similar in many respects. For example, they both are smaller than the moon. For a functional organism,

[78] See also Wattenmaker et al. (1986). The study shows that whether it is easier to learn linearly separable on non-separable natural categories does not depend on formal category structure as such but on category content. Some properties combine more sensibly in additive and some in interactive fashion having implications on the ease of category learning regardless of the linear separability.

something more is required to limit the logical space of possible partitions of the world. Importantly, as Barsalou's work shows, this something is not fully object centered. We impose our own goals on our environment, and once we get some causal learning going it comes to affect our conceptual understanding profoundly. Artificial category research misses these critical points because the experiments are intentionally designed to suppress the effects of background knowledge. The question of how the agent determines the relevant features does not rise because the stimuli is often manufactured in the way that makes the critical features (e.g., colors and geometrical shapes) highly salient.[79] The insufficiency of the methodology in similarity centered tradition, particularly in exemplar research, lies in its frequent dismissal of the role of content, and human intentional content can not be understood purely by formal means, in isolation of human interests and practices.

In summary, we should expect similarity effects to dominate in artificial categorization tasks unless the stimuli is contextualized in a way that makes intuitive sense to participants, which means the subjects can connect the description of the stimuli with their already established knowledge structures. Knowledge effects should be prevalent in tasks employing natural categories. The word "natural" here should be understood in a subjective or relative sense of "familiar." The distinction between "natural" and "artificial" in this context has little to do with the distinction between natural kinds and artifacts. What appears "natural" is determined by the task and subjects' background knowledge. For example very young children who lack biological knowledge rely on similarity-based categorization, even though there clearly is nothing unnatural about biology. Likewise, in a sense, there is hardly anything more natural than elementary particles; however, at least to the untutored, their nature is rather alien making perhaps properties such as "spin" to appear psychologically and phenomenologically similar to arbitrary features in artificial categorization tasks. The more acquainted, on the other hand, might develop a sense of familiarity with the concepts of fundamental physics, which makes its odd properties appear utterly natural in light of other strange features, such as the appearance of the interference pattern in double slit experiments as a natural consequence of the wave–particle duality.

If the hypothesis developed here is correct, the above remarks should concern knowledge use extensively. The exemplar effects in artificial categorization tasks are supposed to reflect a limiting case of knowledge effects: They emanate

[79] Even though early researchers *did* mention the theoretical importance of context in attribute perception, e.g. Tversky (1977), Rosch (1978), and Medin & Schaffer (1978).

from selective activation of contextually relevant pragmatic knowledge in case no rules, or summary or causal information about the category is available. Then the only (pragmatic) knowledge that is associated with the test stimuli is its proper categorization response in the context of the experiment. Hence, categorization in these tasks reflects a rather superficial skill for selecting the correct categorization response for a test stimulus that is utterly meaningless outside that very task. This can be seen as one of the simplest purely formal goal-directed behaviors, where the sense of the stimulus is solely constituted by its role in more or less trivial stimulus–response tasks. If this is true of thinking in general, albeit with more complex means-ends patterns and inferential structures, we should also find exemplar effects in reasoning, especially when reasoning employs concepts that do not form prototypes and are relatively isolated from other conceptual domains (e.g., abstract discursive concepts where cross-domain background knowledge bears little impact on reasoning.) Thus expert reasoning, especially with theoretical and formal hard-to-learn concepts, should be an excellent place to find these effects. This is because the psychological contents of these concepts should mostly consist of knowledge about their discursive use, and the exemplars should contain information mostly about their previous applications in public and private reasoning. Because exemplars consist of very specific information, this should result in a limited ability to apply conceptual resources outside familiar tasks and situations. Indeed, if exemplars work as the intuitive gateways to contextually relevant information, the result should be even the dismissal of existing relevant knowledge in unfamiliar contexts. That is discussed further in Chapter 5 in connection with the analogical transfer.

Unfortunately there is little consensus on how the relevant causal knowledge is represented and how concepts are accordingly organized. The basic idea is that grasping a concept is an ability to explain its instantiation by resorting to some of its causal properties; hence the often used name "theory–theory." However, the notion of "theory" seems to be rather loose here if not entirely misleading. Traditionally, theories are taken to be explicit linguistic constructs that allow prediction and explanation. But knowledge account is not committed to conflating concepts with language. Instead, the idea is more general: that concepts are represented by properties that cluster around causal attributes. Some classification strategies are explicit, conventional, and culturally transmitted (e.g., in the pizza/coin case and quite likely in the folk classification of whales as mammals), but often conceptual thinking is implicit

and idiosyncratic, making the role of public discourse less relevant. (Murphy & Medin, 1985)

The primacy of language and explicit theorizing are also questionable in light of developmental psychological evidence. Even prelinguistic infants are able to pair a toy dog with a toy bird instead of a toy airplane, which looks much more like the bird, displaying categorization capacity that transcends mere observable surface similarity (Mandler, 2004, Chapter 8). Moreover, the effect of knowledge about the causal structure of a category to typicality ratings is evident in 750 msec time frame, implying that explicit thinking is not necessary to bring about these effects but the impact of knowledge can be produced by fast intuitive processing (Luhmann et al., 2006). It should be clear that similarity- and knowledge-based processing do not form cognitive competencies that align neatly with the implicit/explicit or intuitive/reflective distinctions as they are generally conceived in the dual-process theories.

Therefore, we perhaps should conceive the knowledge effects as arising from general sense-making capacities. Explicit theoretical or rule-based reasoning is one but perhaps a rather marginal way of making sense of the world. Some sort of mental models or other less representationalistic ways of understanding how things hang together should suffice to bring about these effects, at least in principle. However, in principle, nothing reviewed thus far emphatically rules out the possibility of some kind of innate language of thought for representing causal and taxonomic knowledge. For now, I leave the discussion at that. The next two chapter will introduce the main empirical hypothesis and theoretical contribution of this work. It revolves around a pragmatistically oriented intuitive knowledge account that puts some enactivist and neopragmatist ideas to work. The next chapter focuses mostly on non-linguistic competencies, and Chapter 5 concentrates on explicit, formal, and theoretical discursive reasoning and its roots in implicit cognition.

4 Concrete, situated, and pragmatic intuition

The core idea of the hypothesis elaborated below is that while causal knowledge and affordances concern things, they are also about situations. Pragmatic knowledge, i.e. the procedural and causal knowledge about things and situations, stems from the summary encoding of experiences of particular situations that the agent has encountered and yields a capacity to generate immediate expectations on how events unfold and how the actions of the agent affect the outcomes. Here the notion of *situation* is construed subjectively from the point of view of the agent. Situations are roughly defined by the goals of the agent with a material context composed of variables that are causally relevant to it to reach those goals—generally surrounding things and their properties, including actors and their dispositions.

It is a remarkable and ill-understood property of both common sense and expert reasoning that we are capable of focusing selectively on the relevant variables in complex task environments while automatically disregarding the irrelevant. We all have a massive amount of more or less trivial knowledge about the world, and hence the problem ensues that which from all these facts we need to take into consideration when we perform an action or make an inference. Somehow we manage to exploit that vast knowledge base very fluently, and this vital aspect of human cognition is part of pragmatic know-how—both in the common non-technical sense and the specific sense expounded here. Commonsense has proven to be very difficult to model, especially with general logic-based algorithms, and the problem to date lacks an effective general solution (e.g, Davis & Marcus, 2015).

Relevance assessment happens immediately when interpreting a passage or an event. Consider these sentences: "I stuck a pin in a carrot; when I pulled the pin out, it had a hole." and "I stuck a letter in a spindle; when I pulled the letter out, it had a hole." Now, to what does "it" refer to in each sentence? The answer is immediately obvious; however, it is not determined by syntax. To resolve the referent, you need some basic causal knowledge about how pins, carrots, letters, and spindles may interact, and what it means to stick things in each sentence. Of particular importance here is that once you have attained this knowledge you do not use it explicitly to infer the referent. The meaning is completely transparent, and the alternative reading most likely does not even cross your mind. Recall from the introduction the example of a note to a friend saying "I'm making some bigos tonight. Could you check what's in the fridge and drop by the store? Some cash and the

car keys are on the kitchen table." One immediately grasps how making a specific dish, inspecting the fridge, visiting a grocery store, cash, and the car keys relate to each other and understands what needs to be done to fulfill the request. This is quite a feat considering the extent of cross-domain knowledge that might, in principle, be relevant and, therefore, how much information is outright dismissed from consideration. We are so good at connecting the dots that most people presumably do not even realize that all the sentences in the above example are logically unrelated. We just automatically see a meaningful whole.

The issue was noticed very early on in artificial intelligence (McCarthy & Hayes, 1969; Minsky, 1974), and it has also drawn considerable attention from philosophers. This, of course, is no wonder because of its importance to human reasoning and knowledge representation.[80] The problem of relevance is likely not resolved here; however, developing new ways of thinking about knowledge employment and representation is a necessary step toward that goal. I have come to take it for granted that relevance is not a formal but essentially a pragmatic matter. Intuitive common sense results from a continuous sense-making process we participate throughout our lives. This means engaging in all kinds of activity, ranging from play and random exploration to planning and problem solving which endows us with steadily accumulating experience on how things in our environment work and what can be done with them. Through our activity, we create the ontology of our lifeworlds, and making sense is finding out what makes a meaningful difference to what. This is how we begin to understand the situations and the nature of things we encounter.

Note that the construction of an ontology here does not refer to how we change our environment through our labor. Of course, we do that too, both individually and on a massive scale as a cultural species. We are niche constructors. However, we also adapt to existing environments, and this "ontology of the lifeworld" is a subjective partitioning of the environment into meaningful phenomena, which happens gradually through this adaptation as a product of goal attainment. It is an active and constructive process, but I mostly mean this in the psychological sense.

In what follows, I will use overtly representationalist language in describing human conceptual cognition. Still, I maintain that the fundamental meaning engendering cognitive phenomena involve action-relevant selective sensitivity to the environment. Whether or not the sensory contents, associated actions, and anticipated outcomes are represented somehow internally (clearly they

[80] For often cited sources discussing the issue see e.g. Dennett (1984); Dreyfus (1965, 1992).

are at least associated with systematic occurrences of neural or other internal states), none of these components need to correspond to anything that *has* meanings in the representationalistic sense. Rather, they constitute systems of capacities that *are* meanings in phenomenological sense (see McIntryre 1986). Nonetheless, these capacities are employed in explicit thought, and hence I find it difficult to drop the representationalistic vocabulary altogether. This is because thoughts presumably take some internal states as objects that, in turn, receive their contents from these meanings. Since I assume these meanings are mostly tacit, they are not direct objects of thought; that is, unlike classical (computational) representationalists and semantic externalists, I take that the meanings are not mind independent referents but our intuitions about these putative referents. Hence, the constitution of meaning can be unaccessible to consciousness and it sometimes manifests itself only by observing our own activity. I think that the psychology of conceptual analysis should be conceived mostly as mental simulation of concept use situations rather than as introspective access to mental representation. This idea is discussed further in the following sections.

These anti-representationalistic remarks may eventually boil down just to semantics. I am not sure whether I am rejecting or reforming the notion of representation here. Nonetheless, the bottom line is that the basics building blocks of mental content are not symbols that get their meanings from their representational character (like classical computationalist envisioned) but instead it is these underlying tacit systems of meanings that should ultimately explain the symbolic and representational character of thought.

Here are the main features of the proposed intuitive concept producing and consuming system:

a) Inductive (Bayesian) search for causally predictive and pragmatically relevant regularities in concrete situations:
   - Category representations are (prototype) feature clusters structured by causal relations between constituent features.
   - Basic ontology is formed around affordances and features relevant to event–event causation.
   - Basic situation representations are quasi-perceptual model type structures.

b) Instance-specific encoding of event/action/outcome exemplars:
   - Context- and goal-dependent causal expectations are generated by associative (pattern matching) memory retrieval or analogical mapping from exemplar-based situation representations.
   - Without valid expectations, control shifts to exploration or explicit reasoning.

c) Basic reasoning mechanisms are forward causal inference, simulation of situated action, and abstraction by exemplar-based analogical transfer:
   - Surface features are the primary retrieval cues.
   - Valid analogies bind different tasks under shared pragmatic schemata.

First, we the item a). The basic idea follows the common tenet in psychology that knowledge representation incorporates an ontology of entities and causal relations that hold among them.[81] I explain how basic object and event categories may be composed of specific sensory and causal information without a separate symbolic core. I explain the rough outlines of recent application of causal graph models of knowledge representation in cognitive psychology. The bulk of the work happens when addressing of items b) and c) where we see how schematic abstractions can be built bottom-up from concrete event representations via procedural and causal knowledge. This last plank is discussed in the next chapter in addition to how this pragmatic knowledge framework can be applied to yield a skill-based inferential theory of top-down learned discursive and theoretical conceptual understanding.

4.1 Object representation

I have characterized the intuitive system as a learning device that tracks affordances and other causal properties that have pragmatic relevance to us. Binding that knowledge to something requires that the system partitions the environment into categories or, at the very least, be sensitive to certain features and their correlations. I assume that intuitive cognition encodes object and event categories and associated structural information, such as how object features relate to their causal properties. The information encoded is rather concrete and action and outcome centered. Although nothing very crucial hangs on this, I assume the basic category representations take the form of prototypes. The reason to make this assumption is empirical. Clearly, some similarity-based mechanism is needed to explain human categorization, and

[81] See e.g. Griffiths & Tenenbaum (2009, 670). I presume that this is approximately the standard view also in (analytical) philosophy, although philosophers, in general, may have some reservations as conceiving the fundamental relations as being necessarily causal.

it seems that prototype account generally fares more naturally with the data than exemplar models (Murphy, 2016). The latter has a distinct advantage in artificial learning tasks only, and we saw above that it is questionable if those studies tell much about the mental representation of natural categories. Apart from that, prototype research also connects to relevant research concerning the basic representational level and hierarchical structure of taxonomic categories.

4.1.1 Prototypes and basic level categories

It appears a universal tendency for us to group things to categories which form taxonomic hierarchies. For example, a bulldog is likely to be conceived alternatively as *bulldog*, *dog*, or *animal* instead of belonging to non-nested alternative classes such as *bulldog*, *drooling animal*, or *something to rescue in case of a fire*. I hope the reader at this point finds it obvious how prototypes can implement taxonomic hierarchies: A subordinate category has all the features of its superordinate class with some added details. Then subordinates consist of richer and more specific feature sets, and the inheritance of important properties from higher levels of taxonomic hierarchy is guaranteed. This works like a basic set-theoretical taxonomy. We do not need to discuss this in detail here but exemplar models have fared poorly with these sorts of hierarchical structures. (Murphy, 2002, Chapter 7) However, by assuming prototype representation for concrete categories we, in theory, get *taxonomic abstraction* pretty much for free.

Unfortunately, in the human conceptual system, nothing seems to be that straightforward in reality. For example, people are likely to accept that car seats are chairs and that chairs are furniture, but still think that car seats are not furniture (Hampton, 1982). Steven Sloman (1998) found that related violations of extensional logic in taxonomic inference are surprisingly common. This implies that a tree-like hierarchy, based on set inclusion, is very likely not an inherent constraint in the human conceptual system. This does not pose a problem for the prototype view, however. Both studies found that typicality effects may explain the data. Car seats are perhaps seen too untypical as furniture while they share some (rather obvious) critical features with chairs. People perhaps make these category judgments by direct comparison of these features rather than by locating the target category in an existing taxonomic system. This means that we do not necessarily have a precompiled conceptual hierarchy stored in our long term memory but sufficiently abstract superordinate concepts may be usually computed on the fly as needed, and taxonomic

inference may be feature-based rather than category-based. This is likely (if not necessary by definition) with *ad hoc* categories reviewed earlier. See (Murphy, 2002, 204–210) why this is probably true of object categories in general.

The psychological details of the implementation of taxonomic hierarchies are not highly relevant for now, but certain related findings are. According to several authors, most notably the pioneers of the prototype theory (Rosch et al., 1976; Rosch, 1978), a level in conceptual hierarchy, particularly with object concepts, is the most basic in human cognition. This basic level can be observed in people's preferences for naming things at a certain level of detail. For example, a bulldog is much more likely to be called "a dog" than "a bulldog" or "an animal." Words at this intermediate level are also learned first, and basic level categories enjoy other processing advantages, similar to categories with high family resemblance scores. When asked to list features associated with categories in different levels of hierarchy, the basic level stands out as the most informative. Subordinate listings usually contain a few additional adjectives while superordinate level descriptions are sparse and contain mostly schematic (i.e., non-specific) functional features, such as *keeps you warm* and *you wear it* in the case of *clothes*.

If the pragmatist theory that I am trying to advance is correct, then the most critical features in basic level representations should be rather concrete, action centered, and causally relevant. Basic level (as per domain) should also be malleable across individuals and cultures, and this indeed seems to be the case. According to Rosch et al. (1976, 382), in taxonomies of common concrete nouns in English:

> [B]asic objects are the most inclusive categories whose members: (a) possess significant number of attributes in common, (b) have motor programs which are similar to one another, (c) have similar shapes, and (d) can be identified from averaged shapes of members of the class.

Perceptual cues certainly are important for object classification, but here and in (Rosch, 1978) the authors claim that primary object concepts concern things we can interact with, and that the categories are partly determined by their affordances, associated movements, or "motor programs". These are closely involved with perception but not strictly perceptible properties of objects.[82] This might look like a chiefly sensorimotor theory of concepts, but

[82] Although if perception activates contextually meaningful action knowledge associated with the percept, then it can be argued that affordances actually are perceptible properties, at

the status of attributes in plank (a) is left open and may include higher level schematic attributes. Moreover (d) concerns identification, not representation, and hence the characterization leaves room for more abstract information to be incorporated into category representation.

According to developmental psychologist Jean Mandler (2004), the basic building blocks of concepts are structural and functional properties. She calls these structural features such as *path*, *support*, *containment*, and relations such as *up/down* and *above/below* image-schematic since they are not symbolic or linguistic but not strictly perceptual either. Already during the first year, infants also possess conceptual structures about intentional agency and mechanics of inanimate objects, for example (Carey, 2009). Whatever the representational format, the proposed mental contents in this line of theorizing do not consider intellectual essences nor strictly perceptual or motor properties but functioning and structure of things in the environment. According to Mandler (2004, 86–87):

> [T]he meaning of objects for human beings ultimately depends on what they do or what is done to them, as Katherine Nelson pointed out many years ago (Nelson, 1974). If the world stood still, there would not be no conceptual mind—it is events that demand interpretation: What is happening? What is going on? Although it is possible that it is solely the attraction to motion that sets up perceptual meaning analysis in terms of paths and their characteristics, it seems more likely that what infants attend to and analyze for meaning is one of those indirect innate factors, determined perhaps by the need of our species. [. . .] What things do is the core of their meaning, and for some time in infancy it is the only meaning that is available.

least in the phenomenological sense. This particular sense means that in perception features are always interpreted, and hence it is somewhat moot to separate the subjective import from objective observables, even if this difference can be meaningfully made between the physical signal the organism's sensory organs pick up and the resulting cognitive processing. I understand that this is approximately the original idea of affordances in perceptual psychology introduced by James J. Gibson (1979, Chapter 8). In artificial learning tasks, it has been demonstrated that it is possible to associate even arbitrary movements to perception so that executing learned movements consistent with the percept facilitate object recognition. This suggests that action information is, in effect, incorporated into object representation (Ross et al., 2007).

Cognitive capabilities enrich and proliferate when growing up; however, the core capabilities of an infant's mind remain active for life (Carey, 2009). If all this is correct, our object concepts deal in large part with functional information. The more appropriate term would actually be *pragmatic* information because what is represented is functional information with a point of view, and categories contain information about object affordances and how they behave as a result of our actions. So basically here is the basis for my claim that basic object representation is rooted to observable affordances and causal properties.[83]

Despite strong empirical support, the whole notion of basic levels has been challenged. Mandler has complained that there is considerable variability in basic levels across cultures, individuals, and conceptual domains, casts doubt on empirical adequacy of the concept (Murphy, 2002, 234). Second, she and her colleagues have found that the early conceptual system of prelinguistic infants is quite undifferentiated, and the concepts are less specific compared to how Rosch et al. characterize the basic level. For example, in their studies, 14-month-olds differentiated between birds and mammals but not between different kinds of mammals, apparently treating *land animal* as something like a basic category rather than, say, *dog* (Mandler, 2004). This is not a domain-specific effect. Infants generally tend to group things on a leve which corresponds to superordinate rather than basic in adult category hierarchy.

This criticism would be well-founded if the basic level should be universally fixed with the environmental correlation of features as the only determinant. This was suggested by Rosch (1975); however, later Rosch et al. (1976) considered that basic level is an *interactive* notion where the other determinant is the agent with its needs and capacities. They found out, for example, that their subjects' basic level of biological taxonomy was higher than what has been reported in ethnobiological literature. Their explanation for this discrepancy

[83] As far as I can tell, unfortunately, little or no comparative research done with disabled persons who are blind or severely paralyzed, for example. I take that all the talk about visual features and movements are generalizable to other perceptual modalities and ways to interact with the environment. Hence, one should not assume that it is precisely *visual* features that play such a crucial role in basic level determination—even if they might be particularly important for most people with unimpaired vision. The general point, I gather, is that our capacities which allow interaction with the environment play a constitutive part in determining basic intentional content. With different people with different abilities the specific contents are likely variable to a degree, and this might or might not affect the resultant subjective ontology. However, these factor are anyway malleable across individuals (regardless of any disabilities) as a result of our individual life experiences (see the following paragraphs).

was that the ethnobiological studies were carried out with people living in rural, non-industrial societies, while their subjects were Western urban dwellers. For example, for folk in non-industrial societies the genus level of trees (e.g., pine, elm) is basic while in urban industrial societies the more general life form (tree) is basic. This stems from different individual experiences and different practical and social needs in different cultures. At an individual level, this discrepancy translates into a difference in expertise with a particular domain of interest, which also should be manifest in subcultural levels. In fact, the referred ethnobiological work mentions that for American bird watchers the genus level seems to be psychologically unique (in contrast to the general population) for displaying signs of basic level processing in perception (Berlin et al., 1973, 238).

In addition to biological categories, Rosch et al. (1976) also used vehicles as stimuli. One of their subjects happened to be a former airplane mechanic. An interview was conducted that confirmed that the subject had a different canonical view of airplanes than typical subjects (undersides and engines instead of top and side view). His motor programs and list of attributes associated with airplanes was different and taxonomy affected: His conceptual differentiation of aircrafts was not unlimited but his basic level airplane category was more accurate than the norm. (Rosch et al., 1976, 430) The effect of expertise on taxonomy was experimentally confirmed later (Tanaka & Taylor, 1991) with the main result that experts tend to exemplify basic level processing in their area of expertise on the level of taxonomic hierarchy that would correspond to subordinate for novices, effectively blurring the line between these levels and implying that experts categorize selected domains or smaller idiosyncratic niches in more detail. The effect is not marely due to the volume of exposure the experts have with particular things but also due to how they actively pay attention to certain features and their values. Simple instructions make subjects pay attention to specific features which trains them to be more sensitive to specific value ranges (e.g., specific brightness or size), thus enabling them to make more fine-grained classification compared to untutored subjects exposed to the same learning material (Goldstone, 1994). Expertise also affects category structure, in that it tends to shift goodness-of-example measure from the central tendency to goal derived ideals in the sense defined by Barsalou (see page 135) (Lynch et al., 2000).

In summary, the prelinguistic infants studied by Mandler and her colleagues may be in the process of learning cognitive capacities to cope with their environment. Their undifferentiated concepts, in comparison to adults,

might be akin to the undifferentiated biological taxa of urban residents in comparison to people of agrarian societies—or more to the point, that of the basic level of novices in comparison to experts. The basic level found by Rosch et al. is perhaps not the starting point for individual ontology but the generic end-point (shared as a norm in one culture because of widely shared practices), where we stop our conceptual dissection of the environment at the fineness of grain that happens to satisfy our practical needs. After all, the basic level is not the most accurate level of classification of which we are capable. The basic level is instead a compromise between detail and generality that maximizes the cognitive economy of category representation following our needs (Rosch et al., 1976, 384). More inclusive concepts are not accurate enough, and more accurate concepts compromise the economy with irrelevant detail. Only experts find practical needs to adapt to a more refined taxonomy. With infants, there presumably are developmental factors involved also; however, it is a reasonable possibility that not all factors observed in conceptual development be age-related but at least partly reflect general trajectory of concept learning.

The fact that the basic level is malleable across individuals, cultures, and conceptual domains is vital to my claim that not only discursive but also concrete concepts are acquired constructively, i.e., they are forged locally and interactively. Part of the content determination stems from our sensory and motor apparatus and other bodily and cognitive capacities. Contents are also partly determined by the structure of our environment and partly by the practices and the way of life in which we are engaged. All these factors interact, making content determination complex, idiosyncratic, and local but constrained in ways that are, at points, widely shared among humans. I hope these considerations convince the reader that chances for informative *a priori* analysis of fundamental conceptual content, which disregards the variety of human condition, are slim at best.

4.1.2 Causal cores with surface miscellanea?

Concerns can be raised that not all subjectively important attributes of perfectly coherent categories are causally relevant. When one thinks of cats or stars or works of Monet, for example, she probably thinks primarily of sensory features and not what these things do and how they work. The tension between the relative importance of causal and sensory attributes may suggest that categories perhaps are encoded by two independent sorts of information: (a) causal cores, which form the actual conceptual content, and (b) surface

miscellanea, which serve only to identify the category (e.g. Osherson & Smith, 1981; Armstrong et al., 1983). Almost certainly there is some truth to this idea. If such strict demarcation can be made, it could serve as a last line of defense for the classical theory of concepts, retaining the possibility of some more or less hidden metaphysical or intellectual essences. Agreed, at least some categories have clear identity conditions and therefore essential properties, at least in some sense. But the prime examples belong to domains that are epistemologically constituted by theoretical, conventional, and other explicit discursive knowledge. When discussing the knowledge account, we already saw that with natural object categories causal knowledge and similarity cues are often closely linked.

From an early age, we clearly causal over non-causal features when information about the causal structure of a category is available (Ahn et al., 2000a,b), but these priorities are not clear cut. Adults use different knowledge in comparison to kindergartners in assessing what causal connections are plausible and consequently stress different features and make different category-based inferences (Keil, 1992). In laboratory experiments, subjects tend to rank cause features as more important than effect features for category identity: If subjects know that a category features are structured in the way that X causes Y and Y causes Z, then they usually rank X as the most important feature and Z as the least important for determining the category membership.[84] This *causal status effect* means that, all else being equal (e.g., cue validity and perceptual salience), cause features are weighted more than effect features in similarity computations.

Thus, in brief, surface and causal properties do not constitute mutually independent sets. Instead, there is a continuum ranging from a more to a less peripheral properties, and this continuum at least partly depends on general world knowledge rather than solely on a category-specific semantic knowledge. That is, category representation is for understanding how things hang together in general and not for intellectual apprehension of intrinsic essences. Causal attributes do not work like classical defining conditions, but their status is gradual and malleable both in determining and identifying a category. Indeed, not even the causal status effect is robust. More important than the status

[84] For example, if animals called "roobans" tend to eat fruits, have sticky feet, and build nests on trees, then subjects rank the first feature as the most important and the last as the least important if they are told that fruit eating makes roobans' feet sticky, and consequently they are able to climb trees to build nests there. (Ahn et al., 2000a)

of particular features is the coherence of their combinations under the known causal structure of the category.

The notion of coherence here means roughly this: Assume a category is (partly) characterized by a feature vector $[x, y, z]$, and that there are strong causal links $x \rightarrow y$ and $y \rightarrow z$. Given this information, people tend to rate exemplar $[0, 0, 0]$ as a more likely member of this category than $[0, 1, 0]$. The reason is presumably that the presence of y in the absence of x means that y is produced by a cause not intrinsic to the category and hence its presence is then not a very relevant indicator of membership. On the other hand, because $y \rightarrow z$ is strong, the absence of z in $[0, 1, 0]$ is evidence that the exemplar is a non-member because it violates a causal expectation associated with the category members ($[0, 1, 0]$ instead of the expected $[0, 1, 1]$). Instance $[0, 0, 0]$ does not exemplify any important properties, and therefore it is probably a non-member; however, it does not violate any causal laws of the category, either. Hence, it is more coherent and therefore a more likely category member than $[0, 1, 0]$. Apart from structure, this inference depends on the assumed parameters of the category: If y is prevalent regardless of x, and $y \rightarrow z$ is weak, the coherence effect in found to diminish, like it should. (Rehder & Kim, 2010) In conclusion, representation of natural categories should be construed as causal models and not just as unstructured weighted feature lists.

Perhaps apart from explicitly defined concepts, the supposed cores simply have no clear utility in explaining human categorization performance (Rips et al., 2012; Murphy, 2002, 24–28). Almost everything we currently know about human conceptual cognition rather points out that, by default, the information deemed relevant in categorization and category-based inference depends on personal knowledge and specific tasks we are posed or questions we ask. So the works of Monet *as* the works of Monet are, naturally enough, intrinsically dependent on the fact that they are produced by Monet. What makes them fabulous and unique is perhaps a matter of their intrinsic combination of perceptual features that, again, are causally related to the fact that they are painted by Monet. What makes them instances of *art* (superordinate category) is perhaps dependent on some unspecific functional or causal properties, for example, that they are *effects* of artist intentions, *causes* of aesthetic apprehension, the *role they occupy in* history and *praxis* of the social institution called ”art”, or some combination of these. Then, what kinds of things cats are depends on whether we are asking a philosophical, biological, or more casual question which, in turn, may suggest different answers depending on whether cats are considered as pets or utility animals, why are we posing this question

in the first place, and so on. There is no unambiguous conceptual core that would provide the correct answer in every conceivable context.

The point is not to claim that the distinction between accidental and conceptually constitutive attributes is entirely useless, but that seeking them without specifying a context that is, and taking these limiting notions as the fundamentals of category representation is misguided. The second point is that properties that may look like superficial may (a) be complex products of causal factors, or (b) afford various human practices. For example, specific pieces of art are perhaps not best conceived by their raw perceptual attributes but rather via the expressive patterns they exemplify, which are unique products of certain artists and embedded in wider cultural practices. The ability to recognize and appreciate these patters, as the ability to recognize cats, dogs, and their general characteristics, may be a requirement for us to be competent members of our society. It depends on our cultural norms and practices (including personal status and subcultural niches) which things we need to be able to recognize and cope with. In this way, perceptual properties may be important *affordances* that enable us to take part in these practices.

This brings us to the second leg of the answer to the worry that some categories may essentially consist of superficial attributes. The less apparent but better substantiated answer is related to the plank (a) in the previous paragraph: Even if causally vacuous attributes are prominently featured in mental representations of objects, it is not necessarily a problem for the idea that concrete concepts are formed around causal attributes because observable features might be coded as *effects* of an underlying causal structure, which is also responsible for behavioral and other non-obvious properties. Indeed, people often assume that there is an underlying causal factor responsible for making things the way they are, which is treated as a placeholder for unobservable and unknown causes and effects; for example, that there is a causal factor that all cats, and only cats, share which makes them cats (that perhaps would be the cat DNA). This tendency is called *psychological essentialism*.

A sign of essentialism is that even one year old children prefer category-based inferences over surface similarity if they have a reason to believe that apparent similarity is not diagnostic of category membership. For example, they infer that beetles share properties with bugs that look like leaves, but not that beetles share properties with leaves. Hence, even as infants, we prefer to treat things as alike on the basis of presumed unobservable similarities over evident observable similarities. The second indication of essentialism is that we tend to assume that if we change the observable features of organ-

isms (and some artifacts), leaving their internal structure intact, the organism retains its original category identity because of the critical unobservable structure. (Gelman, 2004) Since the diagnostic features that point to the assumed essences are malleable, and the domains where essentialist reasoning is manifest changes with age and experience, this tendency might be partly learned. However, because of the prevalence, cross-cultural stability, and the early onset of essentialist thinking, it seems more likely that these differences reflect how acquired domain-specific knowledge modulates our innate tendency to essentialist reasoning (Ahn et al., 2000a, B41).

Principal domains of essentialist thinking are (assumed) natural kinds, such as chemical elements, organisms, and diseases (Ahn et al., 2013). However, it depends on the specific categories and tasks if and how essentialism is manifests (Gelman, 2003, Chapter 6). Moreover, if subjects have other plausible and relevant causal knowledge available, it often overrides both category- and similarity-based inferences (Medin et al., 2003). For example, people are more prone to infer that sharks are more susceptible than deer to a disease that trouts have, because sharks are more like trouts; however, many tend to think that bears are more susceptible to the disease than sharks because bears eat trouts, and hence the disease might be transmitted to bears more easily. Also, people readily understand that some observable properties are caused by hidden variables, while some are not, and the resultant inferences depend on the subjective plausibility of the causal connections. For example, greenness is more probably considered an accidental feature of a car than of a pine needle. Research conducted into artifact categorization shows that whether artifacts are considered to have essential features or not depends on the task as well as the assumed nature of the essence, for example the intended function or the intended kind membership by the creator (Sloman & Malt, 2003). Children weight more surface similarity instead of assumed essences in artifact than in biological domain, reflecting the reasonable idea that artifact identity is determined by its function, which can often be inferred from its appearance (Gelman & Markman, 1986). Sometimes, it is precisely the observable surface features that are deemed essential (e.g. in case of a pizza slicer).

In summary, the research on psychological essentialism underscores the recurring theme in contemporary concept research: there is no single unitary process for categorization, and different conceptual domains and different goals the categorization serves uses different information (Gelman, 2004). Basically, essentialism boils down to a default assumption that there is an intrinsic, yet often unknown, reason why the things are the way they are and which warrants

taxonomic inductive inferences. However, this default stance can be dismissed in the face of any other known relevant information.

Essentialism shows that category representation with all its surface miscellanea is causal to the core: Observable features are often tacitly encoded as diagnostic effects of an assumed hidden structure. It is instructive to note that this interpretation of psychological essentialism supports the idea that the conceptual system is tracking epistemologically indeterminate causal structures. Causal mechanisms and relations are not directly observable but need to be inferred, and it is generally reasonable to assume that observable feature correlations in categories are not accidental but an effect of an underlying structure, which makes the members the way they are. In the causal learning literature, it is a common observation that people spontaneously assume hidden causes and track their effects (e.g. Cheng, 1997; Gopnik et al., 2004; Griffiths & Tenenbaum, 2009; Kushnir et al., 2010). Hence, a natural consequence is that we treat the causal constitution of kinds potentially open and indeterminate like any causal structures. As Douglas Medin remarked, essentialism may be bad metaphysics but sound practical epistemology: Similarity guides learners toward causal structures and works as a constraint on the search of causal knowledge (Medin, 1989). In the absence of better knowledge, it is a sound practice to discern the environment as clusters by appearances and take the dissimilarities as a sign of potential relevant functional differences. However, appearances are often misleading, and hence it is also reasonable to treat the observable structure not as an aggregate of causal loci but a sign of one and restructure the category representation around better causal knowledge when it is available.

Psychological essentialism is not competitive to similarity- and knowledge-based views but rather an additional component that works as a unifying bridge between them. This reading of psychological essentialism is highlighted, e.g., by Medin (1989), Gelman (1994), Rehder (2007), and Rehder & Kim (2010).[85] It also shows that however concrete our intuitive thinking might be, simple empirical associationism is out of the question. Mental representations of con-

[85] Note that Rehder (2007) aims to show that essentialism combined with the causal model approach to categories can salvage the distinction between accidental and essential properties in mental representation; so we apparently have quite a different drift here. However, I think we are just emphasizing different messages. As a psychologist, Rehder tells that the traditional distinction in philosophy can be partly reconstructed by modern cognitive psychology, and I agree (see below in the main text). As a philosopher, I try to tell that this distinction is not (psychologically) neither fundamental nor all present, and if I am not mistaken, Rehder would agree.

crete entities are not just aggregates of distinct obvious features, but we expect the world to be structured as a causal nexus where categories are construed as nodes even when the sources of causal power are not observable. In other words, we do not track mere regular patterns of sensory features but also infer causal regularities and make tacit assumptions that transcend mere observation. Categories are not represented as a static collection of knowledge but as open structures that also serve a placeholder function for potentially unknown properties (Gelman, 2004). This, of course, is mandatory if category content is learned and potentially always revisable.

Psychological essentialism is also my final answer (promised on pg. 107) to the claim that Kripke–Putnam concept externalism implies the dissociation of psychology and philosophy of concepts. Regardless of whether we can (or should) replace philosophical analyses of concepts with psychology or not, psychological essentialism helps us understand the nature and the etiology of some of our analytical intuitions without implying metaphysical or analytical essentialism. As I see it, Kripke and Putnam are exercising introspective psychology or explicating the commonsense phenomenology of kind concepts, which emanates from our psychological disposition to essentialist thinking. There is absolutely nothing wrong with that, and I take that psychological essentialism corroborates some of their key insights. What I oppose is the claim that their account makes empirical and philosophical concept research incompatible. If we are to understand how concept essentialism works, we perhaps better concentrate on empirical research rather than *a priori* metaphysics. However, "empirical research" here is not confined to cognitive psychology.

Chemical elements are prime examples of natural kinds, but do different isotopes have the same or different essences? If common people think that ibuprofen cures headaches, does their ibuprofen concept refer to the ibuprofen molecule *per se* or specifically to its biologically active *S*-enantiomer? What if the people in question happen to be medicinal chemists? (The question is, do medical chemists and common folk track different but equally real kinds, or does the folk fail to track anything or perhaps a disjunctive kind, instead?) Is the essence of water a substance that conducts electricity or H_2O, which is, in fact, an insulator? Is the essence of a disease its underlying cause or the symptoms it causes? (Think, does "flu" refer to a virus or a condition?) Do superordinate categories such as "pet" have intrinsic essences?

I think these are not perfectly good questions to ask. The right questions are how these concepts are deployed in intentional action (including thinking) and what determines their use. These are not metaphysical but practical,

psychological, contextual, and cultural issues. I find it evident that it is not strictly an empirical question if, for example, isotopes of iron or spin-isomers of water are the same *kind* or not. If these sort of questions are supposed to make sense at all, they need to be asked in specific times in specific contexts that delineate what difference it makes, which, in turn, means that these are not purely metaphysical or *a priori* questions, either (i.e., independent of contexts and human interest). And indeed, the philosophers who discuss these matters generally introduce concrete examples and scenarios to prompt our intuitions about them, and these intuitions, I maintain, are products of the interplay of human practices and cognition.

The assumption that our analytical intuitions track extensions determined by stable essences is problematic also because we are disposed to track essences when there are none. Biological species are a good example. They are often taken to be representative examples of natural classes, determined by hidden essences, namely genes. However, we know that there is no strict demarcation if two organisms are different species (especially along the phylogenetic lineage). Common (mis)conceptions about the nature of race, gender, and various social groups are also good examples of where genetic essentialism goes wrong (Dar-Nimrod & Heine, 2011). With artifacts, people tend to be inconsistent and unsure whether they are dealing with things with essences or not (Sloman & Malt, 2003). Now, apparently "gold" means the physical stuff, whatever it is, and the modern science tells us that it is an element with atomic number 79, as Kripke proclaims. But this observation only tells us something about our scientific epistemic practices, and if it really contributes to anything, it only underlines that the assumed essences are multifaceted and fluid entities—products of our cultural practices and psychological dispositions, affected by context, and *also* by critical thinking and theoretical knowledge. In any case, concepts are psychologically complex and essentialist thinking is only one piece of the larger puzzle.

Lastly, one reasonable interpretation of essentialism is that it eventually vindicates symbolism in mental presentation. Clearly, the defining feature of essentialist thought is a sort of referential character: Since assumed essences are often unknown, they cannot be defined by specific contents but by something like a formal character of category representations to involve a placeholder that points to some unspecified thing. I have been deliberately avoiding any reference to linguistic thought, but it is particularly interesting how language relates to essentialist thinking. Category labels are processed in a different manner than other category properties, not as features but a reference to the category

as a whole (Markman & Ross, 2003). Children tend to group things based on thematic associations (e.g. pairing *poodle* and *dog food*); however but labeling things by nouns promotes taxonomic classification, instead (pairing *poodle* and *German shepherd*) (Markman & Hutchinson, 1984). Labeling entities with generic names or describing them with noun phrases invites children to regard things as belonging to a kind category and also to essentialize them accordingly. Hence, *words* (or more specifically nouns) *in natural language* have a symbolic character. They are not merely associated with already established category knowledge but also point to a potentially unknown category essence and foster certain ontological assumptions. If one thinks of language and mental representation as somehow isomorphic (a core assumption in the Language of Thought tradition associated to classical computationalism, e.g., Fodor 1975), then one perhaps cannot avoid the conclusion that conceptual mental representation also has this symbolic character. However, as we have seen, category representation and essentialism do not require linguistic representation.[86] This does not mean that all concepts can (or always are) represented independent of language, but that any category knowledge, such as an object function or a set of its observable features, can serve as cues that point to the assumed essence. The relation of this knowledge to its associated category is not symbolic (in the sense of arbitrary and referential) but constitutive of category knowledge and identity.

As far as I can tell, essences are not always mentally represented (e.g., with amodal symbols). Essentialism is a processing characteristic or an attitude associated to a category rather than a distinct representational feature. Essentialize something apparently requires taking it as a kind, which means that essentialist tendencies are not fundamental to intentionality but necessitate specific conceptually structured processing.[87] It is still important and

[86] See Gelman (2003, Chapter 8) for an overview of the relation between language and essentialist thinking, especially in children.

[87] Remember my criticism of Fodor on page 43, that a "settled policy of using [e.g.] 'water' as a kind term" requires intentional thought and, hence, cannot be part of reductive basis of intentionality. In somewhat similar fashion, essentialism requires that the agent considers something as a kind and, hence, essentialism is a product and not a foundation of conceptually structured thought. Recall from our earlier discussion that although young children essentialize, they often treat observable properties as more critical for category membership compared to older children (p. 133). Therefore, while the domain-general capacity to essentialize may be innate, it is likely that domain-specific dispositions to do so are learned. The point is that essentialism is not a universal constraint on thought, which is relaxed when the child learns that it is sometimes inappropriate, but more likely essentialization happens after learning some causal principles about a domain.

interesting that apparently a mere word token is representation enough to be taken as a sort of mental sign of an essentialized category (and perhaps word use as a piece of knowledge about the category). More specifically, (a) words as such may work as mental symbols, and an additional symbolic representation for the associated linguistic concept may not be required, and (b) words are associated with non-linguistic category representations in non-trivial ways.

The analogy between essentialist thinking and the referential character of language might not be accidental. This is quite speculative, but perhaps the cognitive capacities that enable observable attributes to "point to" essences also enable nouns to point to unknown absent objects; that is, both are founded on a capacity that enables us to anticipate that there is something real beyond manifest signs. If that is the case, and if the mental representation of categories as causal models explains essentialism (as Rehder (2007) claims), then one may suspect that causal learning likely has something to do with referential aspect of natural language. This is relevant to neopragmatist language theories but this specific aspect should not be overstated. Quite likely the coordination of joint attention and other communicative pragmatics is also part of the explanation of the referential capacity of language and thought, especially for present objects.[88] Generally, neopragmatists are more interested in causal regularities that are not inherent in things but brought about by agents.

Before moving on, I should clarify that the "non-trivial ways" in which language is associated with non-linguistic knowledge means that labels are not cognitively processed as category features or simple name tags.[89] As discussed above, language might constitute a partly independent representational device and linguistic knowledge may be encoded somewhat separately from non-linguistic category knowledge. For example, explicit linguistically mediated theoretical knowledge is perhaps (at least partly) dissociable from knowledge encoded in implicit prototypes. Note that this does not mean that feature-based representation is independent of knowledge in a sense meant in knowledge account. Above, I argued that it is not. Instead, the claim here is that there might be different kinds of knowledge—linguistic and non-linguistic, or *de dicto* and *de re*—which are somewhat separate in the sense that discursive use of language is not simply an overt expression of covert mental representations.

[88] In short, the picture should look like this: Count nouns promote both (a) taxonomic classification, which probably has an important communicative function for underwriting the intent communicated is about an object, and (b) essentialization, which likely is brought about by causal representation of objects and may have the communicative function that the reference is not sensory data but a thing with generic, inductive supporting identity.

[89] This was proposed by Anderson (1991b) and refuted by Markman & Ross (2003).

For example, the fact that whales are mammals may not be part of the whale prototype while it may be part of whale knowledge. This would explain why people seem to construe "fruit" differently in biological and culinary contexts (the former being discursive *de dicto* type and the latter pragmatic *de re*). I am not arguing that this necessarily implies two concepts, linguistic and non-linguistic. Clearly, biological and culinary notions of fruit are closely related and largely co-referential. Instead, I have in mind that different contexts such as biology class and cooking activate different knowledge about fruits. Does this mean that the human conceptual system is inconsistent? Not necessarily. Later, I will explain that superordinate categories (e.g., *fruit*) are not always stable precompiled information structures but constructed (or interpreted) on a contextual basis as needed. Even basic object concepts have different structures and functions and the kinds of information that we learn and employ vary according to context. There is an interplay between perception, language, and knowledge, which mutually affect task performance, and it is needless to pit these determinants against each other.[90] Before that, we will have a brief overview of how (language independent) causal knowledge is represented and acquired.

4.2 Causal knowledge representation with Bayesian networks

We are now moving into a less charted territory, but in the last decade or so it has become increasingly popular to model causal inference and learning by Bayesian causal networks.[91] Although Bayesian methods have a long history in machine learning, it was not until the 1990s when Bayesian statistics emerged as a *normative* model of reasoning in cognitive psychology. Jonathan Evans (1991, 493) has summarized the Piagetian argument for logic-based approach in psychology as follows: "[P]eople are intelligent, intelligence requires accurate reasoning, logic describes how correct deductions are made, so people must reason by logic." This is abductive inference where a conclusion is partly derived from the normative status of deductive logic.

[90] Cf. e.g. Jones & Smith (1993) and Gelman & Medin (1993) for discussion about these points specifically within the framework of psychological essentialism.

[91] The major contributions were the development of Bayesian networks for probabilistic knowledge representation and reasoning (Pearl, 1985, 2000), a sort of Kantian *a priori* framework for causal induction (Cheng, 1997), and their amalgamation for psychological modeling (Gopnik et al., 2004; Griffiths & Tenenbaum, 2005, 2009; Lu et al., 2008b). See Lagnado et al. (2007) and Holyoak & Cheng (2011) for review and Sloman (2005) for an accessible book length introduction.

On reflection, it is unsurprising that the logical model soon faced mounting counterevidence. Classical logic is bivalent and monotonic, roughly meaning that it is suited for decontextualized inference when information is certain, static, and consistent. With mathematics these assumptions are perfectly fine but most definitely they do not characterize everyday reasoning in real environments. Bayesian probabilistic inference, in turn, is precisely the theory about how you should update your beliefs in the face of new, uncertain, changing, and conflicting information. So the new argument goes that people are intelligent, intelligence requires reasoning under uncertainty, Bayesian probability calculus describes how to reason rationally in the absence of complete information, so in order to understand human reasoning and rationality, psychologists must interpret behavioral data in a Bayesian framework.

Note that the conclusion this time was is that people must reason by Bayesian, methods but that if psychologists use formal theories of inference to interpret their data and observed error patterns as evidence for cognitive theories (they often do), then they need to understand the goals and capacities of the organism and the demands of the environment to which it has adapted, to understand what counts as a lapse of rationality and what may actually be an adaptive response in ecologically valid tasks (to minimize uncertainty, for example). The adequate normative model is probably not logic, and the issue cannot be settled *a priori* on the formal basis alone but in concert with hypotheses concerning the goals, practical challenges posed by the environment, and representation and processing capacities of the agent. This methodology is an actual realization of the critical points that I have been arguing along the way. Since this approach focuses explicitly on intentional characterization of the agent—or *rational analysis* as it is called—it directly connects to our previous discussion in Section 3.1.3 and elsewhere where I argued that intentional interpretation of cognitive agents (i.e., interpreting their means and ends and, by implication, their putative mental contents) has an inherent empirical aspect.

In the classical computationalist tradition, it was already well acknowledged that because of computational constraints people cannot be expected to perform perfectly logically (Simon, 1955); however, most of the focus was on mathematical analysis of algorithmic complexity. Nevertheless, because the deviations from classical logic and the impact of logically irrelevant content in human thinking are apparent and systematic even in trivial tasks, it is unlikely that these findings can be attributed to computational constraints or random performance errors, such as lapses of attention (Stanovich, 1999). Instead,

the rational analysis paradigm focuses on how to interpret the systematic deviations from formal rationality in the laboratory as reflections of adaptive reasoning strategies in agents' normal environment. As such, it does not make strong commitments concerning cognitive representation and processing.[92]

4.2.1 Overview of the basic concepts

The standard χ^2 statistical inference is indirect. Generally, we are interested to know if there is a causal relationship between events or variables C and E. For, example hypothesis H might be that a gene (C) causes a specific syndrome (E). In order to find this out, we collect evidence D about the correlation of E and C and then compute the probability of those data under a *foil* hypothesis H_0, which is generally that C and E are not dependent, i.e., that the incidences of E are distributed evenly in samples having C and not having C. This computation is an elementary statistical test, which tells us how likely the observed correlation in the data is given the foil hypothesis. If the probability is low, this gives us grounds for rejecting the null hypothesis. So what the standard statics gives you is $P(D|H_0)$, which is the probability of the data given the foil hypothesis. However, what you probably want is $P(H|D)$, which is the probability of your hypothesis given the data. But that is something you can not get because of the usual considerations of the problems of induction, empirical underdetermination of theories, etc. Statistical tests tell you what data are not likely (given some boundary conditions) and not what hypothesis is likely. As a point of logic then, the test does not directly upvote your actual hypothesis but you need some independent considerations why the hypothesis is a good idea and a likely alternative in case the null hypothesis is suspect.

Fortunately, if you can give some prior estimate of the likelihood of H you can calculate how you should change that estimate after the data. This is what *Bayes' theorem* gives you:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}$$

[92] See Anderson (1991a) about the fundamentals, Oaksford & Chater (2007) for further developments, and Jones & Love (2011) for an overview of the relevant research and especially for critical remarks on how mitigating the relevance of cognitive mechanism risks producing vacuous theories that end up taking any behavior as rational by default. This discussion is philosophically interesting if one notices how it parallels the standard philosophical critique of purely empirical inferentialism, i.e., if meaning is simply factual use, then all use is correct by definition.

The problem, of course, is that you generally cannot tell the *objective a priori* probability of the hypothesis $P(H)$; however, the point of the formula is that if you have some *subjective* estimate, the formula will tell how you should change the degree of your faith in it *a posteriori*.

Here is a standard example of how this works: Say, you are concerned if you have a genetic disposition to a certain disease. Your hypothesis H is that you have it. You take a test, which is established to give a false positive one time in a thousand, and the result comes out positive. This is your data D, which may now look alarming. Previous research has found that in your reference group the prevalence of the gene is only 1/10000, so this is the prior likelihood of H. For simplicity, let us disregard any false negatives and assume that the probability of the positive test result is 1 if you have the gene, meaning that $P(D|H) = 1$. Then by Bayes' formula the *a posteriori* probability $P(H|D)$ becomes simply $P(H)/P(D) = 0.00001/0.001 = 0.1$. So you start with the prior probability 1/10000, and after the data end up with posterior probability 1/10. Because of the positive test result your subjective assessment of your risk should have been changed but you cannot estimate from the test alone how much to change that estimate. The test as such seems reliable; however, the possibility of false alarm is still much more higher than the *a priori* likelihood that you have the gene because the of low base rate of its occurrence. If you continue your investigations, the $P(H|D) = 0.1$ that you got is then the new prior you adjust after the evidence.

Thus, in Bayesian framework the key notions are *conditional probabilities given data* and *prior probabilities* of hypotheses. The notion of probability is construed as a subjective degree of belief (rather than objective frequency of an event, for example), and inference is direct and proceeds by computing how to change those degrees after observations. The above example also shows how posterior probabilities can be heavily affected just by a single event. In that example, all the relevant empirical facts and quantities were already in place, reducing our task to simple calculations. Things are slightly more complex without such information, as we shall see soon. As everyone remembers, David Hume assured us that causality in itself is not observable. Generally, we need to use prior knowledge to assess the plausibility of a candidate for the observed covariations and events (Lagnado et al., 2007; Griffiths & Tenenbaum, 2009). One might think that bringing prior knowledge to causal judgments is irrational or at least unscientific and means to contaminate evidence with one's biases and prejudices. While the concern is valid, generally that is far from the truth. When variables grow in number, the number of possible causal relations grow

factorially. Because of this combinatorial explosion, it is impossible to keep causal induction tractable if you refuse any tentative judgment on what are plausible and what unlikely causal links.

Imagine walking into a dark room. You see two switches on the wall next to the doorway. You flip one, and nothing happens. You try the next one and lights go on. You immediately grasp how the switches and lights connect causally. No statistical test, of course, would grant you the conclusion. Simple mechanisms such as light switches tend to work near deterministically; however, most causal relations in both science and everyday life are stochastic. Without previous knowledge that associates switches to strong expectations about lighting, these two events (first switch, no light; second switch, light on) would never imply statistical significance and causation. Moreover, there is no logical reason to assume that the switches are relevant; some lights react to sound, some to movement, and *in principle* possibilities are limitless. Yet, even infants seem to infer causality from a handful of observations and this is the norm with adults (Gopnik et al., 2004; Sobel & Kirkham, 2007). Temporal order and interventions are powerful cues to causality (Lagnado et al., 2007); however, they are not always available. Recall the time-worn example that in a hot summer we find a correlation between ice cream consumption and drowning incidents. By intervening on each of the variables (e.g., banning ice cream sales), it should become evident that the correlation is spurious. The point of the example, though, is that even if such an intervention were possible, it would not be necessary. People know that ice cream consumption does not cause drowning and hence readily infer that the correlation must be due to how the weather affects people's behavior. As clichéd as the example is, note that it is rarely mentioned that, regardless of the correlation, drowning does not cause ice cream consumption. We tend to dismiss such a possibility without even noticing.[93] Of course, we get things often wrong, but the point is that we are disposed to find causation in our environment, and specific knowledge helps a lot in narrowing down where to expect causal links, enabling us to draw conclusions from very sparse data (Griffiths & Tenenbaum, 2009). That is the

[93] And note that in the example ice cream sales and drowning rates are measured according to the population level. If you happen to live in a culture where ice cream is widely eaten in funerals and almost never otherwise, the idea that drownings actually cause increases in ice cream sales would probably not appear odd at all but rather obvious. It is just not the dead people eating the ice cream. The morale is that intuitions of both obvious and obviously wrong hypotheses are generally learned rather than given *a priori*.

upside of having a conceptual system with concepts as aggregates of causal knowledge.

Unfortunately, specific background knowledge is not always available, and the fundamental epistemic problem is to find a causal structure in observed data to start with. Fortunately, however, some causal structures can be inferred from raw data by relying on generic assumptions. To use Judea Pearl's (2000, 43) example, imagine someone is tossing two coins, and a bell rings whenever either coin comes up tails. The outcomes of coin tosses (call them B and C) are mutually independent and the operation of the bell (E) is correlated with both. These variables constitute the following causal structure: $B \rightarrow E \leftarrow C$. The entry *b*) below is an example of the data that the system produces. Bit strings B and C represent sequence of simultaneous coin tosses in each column (1 = tails, 0 = heads) and E whether the bell rings 1 or not 0.

a) deterministic system

B C E	*odds*
[0, 0, 0]	0.25
[0, 0, 1]	0
[0, 1, 0]	0
[0, 1, 1]	0.25
[1, 0, 0]	0
[1, 0, 1]	0.25
[1, 1, 0]	0
[1, 1, 1]	0.25

b) B : 011011101000000111001
C : 101000101110001111101
E : 111011101110001111101

Given enough of observations but no knowledge about the causal structure of the system, standard statistical analysis shows the independence of B and C and their respective correlations with E. Statistics only discover correlations, not causation; however, the causal structure of the system can be recovered from the data. This necessitates a reasonable, albeit potentially fallible, assumption that the data is produced by a stable and a minimal, unique mechanism (i.e. no unnecessary hidden variables; see Pearl, 2000, Chapter 2); in essence, a combination of Occam's razor with the supposition that the world has a stable causal structure that is responsible for most recurring events.

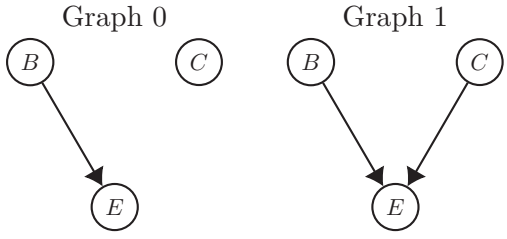
c) stochastic system

B C E	*odds*
[0, 0, 0]	0.2469
[0, 0, 1]	0.0031
[0, 1, 0]	0.0250
[0, 1, 1]	0.2250
[1, 0, 0]	0.0250
[1, 0, 1]	0.2250
[1, 1, 0]	0.0003
[1, 1, 1]	0.2497

d) $P(x = 1) = P(x = 0) = 0.5$
$P(y = 1) = P(y = 0) = 0.5$
$P(z = 1|x = 0, y = 0) = 0.1$
$P(z = 1|x = 1) = 0.9$
$P(z = 1|y = 1) = 0.9$

Standard statistical models can be completely defined by a *full joint distribution*, which specifies the probabilities of all atomic events (i.e. combinations of the values that the variables can take). For convenience, I will restrict the discussion to binary variables. In the above table, *a)* depicts the joint probability distribution for the coin–bell-system. It is basically a truth table of logical inclusive-OR with equal probabilities for all the events. However, both in science and in everyday life, interesting causal relations are rarely deterministic. Entry *c)* represents the same system when events B and C make the bell ring with 90% of certainty, and the bell rings for some unspecified reason 10% of the time. The system constitutes a *noisy-OR* gate, which is simply a probabilistic version of the logical OR. As you perhaps notice, these tables are often difficult to read. A more critical problem, however, is that they grow exponentially: Any added variable doubles the number of rows.

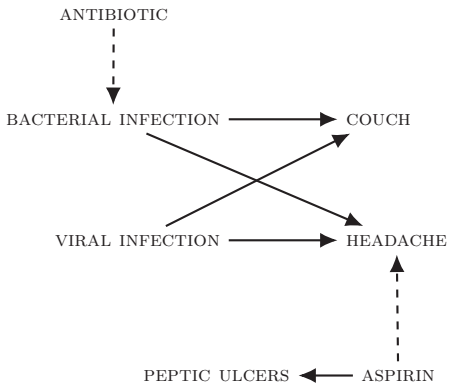
Fortunately, the model can be stated by declaring the base rates and conditional probabilities between events as in entry *d)*. These systems of dependencies are even more intuitive to represent with graph notation, depicting variables as nodes joined by directed edges. Below are two elementary causal graphs. In Graph 0, variable C is completely independent of the others. Graph 1 depicts the (noisy-)OR.



But why *elementary* graphs, if they are not the simplest structures that depict a link (or lack thereof) between single cause and an effect? Graph 0 represents the null hypothesis that C is ineffectual and E is brought about solely by background causes. In

Graph 1, cause C influences E in addition to the background causes B, which are responsible for the base rate of the effect. Note that C can be preventive, decreasing the probability of E. In that case, Graph 1 does not constitute noisy-OR but *noisy-AND-NOT* function, meaning that effect comes about when background causes B are present, AND preventive cause C is NOT present. Just as noisy generative causes do not work deterministically, preventive C generally decreases the occurrence of E and absolutely prevents it only at the limit. Elementary causal judgment is often taken to be a selection between these two graphs. Constant (and undefined) background causes B are assumed to be there independently of the candidate cause C, and B are observed only indirectly by the occurrences of E in the absence of C.

It is possible to omit the background nodes from explicitly appearing in the graphs and implicitly factor them in the models by setting a corresponding base rate for E. Then the elementary graphs depicting causal link reduce simply to two linked nodes, which can be used to build more complex networks such as the one below.[94] As in the rest of the text, I will use dashed lines here to represent preventive causes.



The advantage of graphical models is that they make all the links and independences explicit while keeping the model complexity in check. They also enable qualitative causal judgments without quantitative parametrization: It is not necessary to accurately define the strength of the links or to specify how multiple causes interact to use the bare structural model to infer the influence of interventions and assess the independence of variables. This is psychologically relevant because generally people are taken to implicitly follow the qualitative version of these models rather than (explicitly) use or even understand the

[94] Adapted with modifications from Holyoak & Cheng (2011, 138).

exact mathematical analysis (e.g. Cheng, 1997). Parametrization establishes the exact connection between the graph and the data by specifying how the model exactly generates the data. This gives a fully defined causal net.

A common parametrization in psychological modeling is to use the noisy-OR and noisy-AND-NOT functions to represent generative and preventive causes, respectively. As will be explained below, this basically means to assume that causes operate independently. This specific parametrization is obviously inadequate to model continuous variables that combine linearly, for example; however, it is strongly motivated by psychological research on causal induction from binary contingency data. If the evidence or task so dictates, people are readily able to adopt other functions to integrate causes.[95] In any case, using noisy logical functions helps express intractably complex full distributions effectively (Russell & Norvig, 2010, 520). After the network is set, typical inference proceeds as setting (or observing) specific values for certain nodes and querying the resultant probability distributions in others. That means computing posterior probabilities for query nodes after evidence, which is to carry out basic Bayesian inference. Formally, the inference in the networks is equivalent to the probabilistic extension of propositional logic (Russell & Norvig, 2010, Chapter 14). This is quite evident if the models are chains of generalized OR and AND-NOT functions. From a processing perspective, Bayes nets readily translate to parallel computing systems, aking to neural networks, which pass signal from node to node (Pearl, 1985). Indeed, nets with noisy logical functions are highly reminiscent of the McCulloch–Pitts neural network model (McCulloch & Pitts, 1943).

In the previous section I brought up the idea—after Rehder (2003; 2007) and Kim (Rehder & Kim, 2010)—that concepts should be conceived as causal models. The meaning of the proposal is that concepts are structured like the causal networks described here (parametrized or not). At the limit, all nodes are unconnected, categorization is solely similarity-based, and category knowledge associative. However, representations of coherent and informative natural categories tend to have causal links between features, which makes category judgments knowledge-based and category knowledge structured. If a category is essentialized, the essence can be represented as a hidden/unobserved background cause node. Just like in general causal induction, category learning should be affected by available background knowledge, enabling inference of

[95] See e.g. Novick & Chend (2004); Beckers et al. (2005) and Lu et al. (2008a) for Bayesian model of the latter results; indeed, even rats are capable of learning the conjunction and exclusive disjunction of causes (Fast & Blaisdell, 2011).

unobserved properties in unfamiliar categories and rapid acquisition from few or even one example.[96] Categorization judgment, then, can be seen as both an inferential and associative process: you observe features of a candidate member and (tacitly) use the model to compute how likely your causal model of the category produces that assignment of features. In case interfeature causal knowledge is not available, causal models are unconnected graphs—i.e., basically unstructured prototype representations. Then the classifier simply matches the observed features against the assumed default values of the category which is the limiting case of the same general mechanism. Causal status and coherence effects (discussed on page 151) can be derived from this general account under different model parameters (Rehder & Kim, 2010).

4.2.2 Parameter and structure learning

As a first approximation one might assume that probabilistic causal relations can be stated simply as $P(e^+|c^+)$—i.e., probability of E occurring when C occurs—and the larger the value is the more there is evidence of causal link $C \rightarrow E$.[97] This, however, conflates causal structure learning to parameter estimation which means determining causal strength under the assumption that the link is there. Clearly, this is conceptually different from estimating the probability that the link exists in the first place. As we shall see shortly, these assessments are also psychologically distinct. Moreover, $P(e^+|c^+)$ is the correct estimation of strength only if C is the *only* cause of E. To see this, assume that we would like to know if chemical injection to laboratory rats (C) causes the expression of a particular gene (E). One hundred rats in the test group are injected and 40/100 express the gene while none does in the uninjected control group. In a second experiment, different chemicals and genes are studied, and the results come up as 53/100 and 46/100, respectively. Clearly, in the second experiment, $P(e^+|c^+)$ is higher than in the first (0.53 vs. 0.40) but people judge the first experiment to provide stronger evidence for causality. The gist is that causality is not just about the association of cause and effect but also what would happen in the absence of the cause. So what we want is something like the difference $P(e^+|c^+) - P(e^+|c^-)$, which is 0.4 in the first experiment and 0.07 in the second. This measure is called the

[96] This ability has been long recognized in humans, and recently Bayesian methods have been used to implement it in visual classification (e.g. Lake et al., 2015).

[97] I use the standard notation where variables are denoted with upper-case letters C, E, etc., and their binary values with lower-case with superscript: c^+ marks the presence of the (candidate) cause, e^- the absence of the effect, etc.

ΔP. However, people also judge the result 7/100 (test) vs. 0/100 (control) to give better support for causation than the second experiment (53/46/100), although the ΔP is exactly the same 0.07. (Tenenbaum & Griffiths, 2001)

Patricia Cheng (1997) proposed a correction with the ΔP model by insisting that *causal power* estimates should be sensitive to occurrences of E in the absence of any other cause than C, giving $\Delta P/[1 - P(e^+|c^-)]$ as the correct estimate. Cheng's model is a combination of the Humean idea that causal judgments are sensitive to observed covariation of cause and effect, and the Kantian view that people bring prior assumptions about causality into learning situations. Specifically, these assumptions are as follows:

1. Any event E is always caused and hence in the absence of C it is brought about by some (unknown/unobserved) background causes B.
2. Causes (C, B, etc.) influence E independently (unless there is a reason to suppose otherwise).
3. The causative powers of C and B (expressed as weights w_c and w_b) are independent of the frequency of their occurrences, meaning that rarely occurring events may be strongly causative and frequent causes weak.

Given these assumptions, the combined effect of B and C on E constitutes the noisy-OR: $P(e^+|b, c; w_b, w_c) = w_b b + w_c c - w_b w_c bc$. Then $\Delta P/[1-P(e^+|c^-)]$ gives the maximum likelihood point estimate of the causal power of candidate cause C, that is, value w_c that would maximize the likelihood of the contingency data used to compute that estimate. If C is preventive, the interaction of B and C takes the form of noisy-AND-NOT: $P(e^+|b, c; w_b, w_c) = w_b b - w_b w_c bc$. This is the theoretical explanation of why elementary causal induction conforms to noisy logical parametrization.

While Cheng's causal power model and χ^2 statistics predict some learning data quite well, they both fail to explain causal judgments when the ΔP stays constant at 0 while both $P(e^+|c^+)$ and $P(e^+|c^-)$ decrease equally. Think of test result 8/8/8, that is, 8/8 positives when the candidate cause is present and 8/8 positives when the candidate is not present. The effect is at the ceiling, so a statistical test can not differentiate the experiments and hence provide any grounds for rejecting the null hypothesis of no causation. The ΔP gives 0 difference, and causal power is actually not defined but we can take it take it to be 0 (for it is always 0 when $\Delta P = 0$ and $1 - P(e^+|c^-) > 0$). Hence, all these models give 0 evidence for causality. The same goes for results 4/4/8 and 0/0/8. Human subjects, however, are quite unsure on 8/8/8 and presume

moderate support for causality. This decreases on 4/4/8 and goes pretty much to zero on 0/0/8.

Joshua Tenenbaum and Thomas Griffiths conjectured that this pattern of judgments could be explained by separating structure learning from parameter estimation. After representing the contingency data, the question "What is the probability that C causes E?" can be interpreted as either querying $P(E|C)$ or $P(E \rightarrow C|D)$, and it is the latter question that the subjects are answering in the above experiments (Tenenbaum & Griffiths, 2001; Griffiths & Tenenbaum, 2005). Recall the two elementary graphs, Graph 1 representing a connection $C \rightarrow E$ and Graph 0 missing that link. If we are not interested in the strength of the link but only if it exists, and moreover we assume that the possible background causes of E are the same in both models, the hypothesis space is strictly bivalent: either Graph 0 or Graph 1. Given data D about the correlation of E and C, we can compute the so-called *Bayes factor* from the ratio:

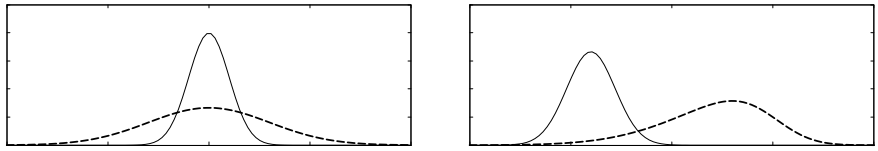
$$\frac{P(Graph\ 1|D)}{P(Graph\ 0|D)} = \frac{P(D|Graph\ 1)}{P(D|Graph\ 0)} \times \frac{P(Graph\ 1)}{P(Graph\ 0)}$$

Assuming uniform priors, i.e. $P(Graph\ 0) = P(Graph\ 1)$, the factor is simply $P(D|Graph\ 1)/P(D|Graph0)$. It is often expressed in the common logarithm form, which gives a direct comparative measure of how strongly the data support the hypothesis. For example, if the $\log_{10}$ value is 5, then Graph 1 is five times more probable, given the data; the $\log_{10}$ value -3 means that Graph 0 is three times more probable, and on 0, there is no support for either side. Hence the factor is a Bayesian measure for model selection. Tenenbaum and Griffiths used the log factor to define *causal support*, a measure that is proportional to the subjective strength of belief in the existence of the causal connection. The actual posterior probabilities can be computed from the factor after the priors are set.

Generally, χ^2 statistics is used likewise to discover causal structures rather than strength. However, it fails to model psychological causal induction because it is more conservative than human judgment. Causal support model predicts, inter alia, the learning data mentioned above. The explanation is that on 8/8/8, the effect is already at the ceiling, so no influence of C can be seen from the data but there can be no evidence against causation either. Therefore, if you presume equal prior probabilities for and against the existence of the causal link, you should stay at that after the data. On 4/4/8, there are

chances for the influence of C to show up. We don't see it, but because of he high base rate and small sample size, this does not mean that it is not there. On 0/0/8, we can be more confident that the effect is either weak or lacking. Therefore, here the Bayesian search for causal structure works differently from χ^2 statistics. Unlike χ^2, it is also applicable with scant and qualitatively restricted data, for example, when information about c^- events is completely missing. (Griffiths & Tenenbaum, 2005)

What comes to learning causal strength, one problem in Cheng's theory is its insensitivity to sample size. It does not make a difference whether the data are 2/0/4 or 25/0/50. Both suggest the same 0.5 estimate for causal power, but clearly you should be more confident of your estimate in the latter case. Things are different with χ^2 statistics which is sensitive to both sample and effect size. Statistical significance test conflates them into a single measure finding weak effects from large samples and large effects from small samples. An interesting aspect of causal support model is that it can combine the representation of both strength and associated uncertainty by using statistical distributions that integrate over all possible estimates of w_c. Below are depicted two illustrative distribution curves. The peak of the curve gives the best estimate of strength, and if noisy logical parametrization is used, the peak corresponds to Cheng's causal power. Hence, "... this results in a relationship between causal support and causal power. Speaking loosely, causal support is the Bayesian hypothesis test for which causal power is an effect size measure: it evaluates whether causal power is significantly different from zero" (Griffiths & Tenenbaum, 2005, 350).



Posterior probability distributions of w_1 ranging from 0 to 1. On the left two data sets results to the same power estimate 0.5. The solid line represents a larger sample size leading to more certainty and support. On the right, we have two distributions with the solid curve showing more certainty but smaller strength. Causal support is a point estimate that is basically determined by how much distributions place their mass away from zero, and hence it is not directly proportional to the peak height: In no cause condition, a high curve peaks at 0 and implies no causal support.

The last learning model that I discuss builds on these conceptual foundations. Causal support theory with noisy logical parametrization is basically a Bayesian extension of Cheng's causal power with uniform priors over causal

structures. Lu et al. (2008b) introduced the so-called *SS power* model which, utilizes largely the same conceptual framework but assumes that human causal induction has a marked tendency to expect dependencies to be *strong* and, by implication, *sparse* (hence the "SS"). The theory leans towards Cheng's proposal than to causal support model because it places more emphasis on effect strength than sample size, and especially because priors are defined over causal strength estimates rather than graph structures. The model utilizes directly the strength distributions depicted above, and computes the likelihood of data under the assumption that C and B influence E strongly. This means that posterior probability distributions are by default elevated near the extreme values (0 and 1), and a causal hypothesis is considered likely only if the weight distribution is spread mostly near 1. This does not make it impossible to learn weak dependencies, but they require more data than in the causal support model with uniform priors. The SS model seems to have an empirical edge over the support model (Lu et al., 2008b; Holyoak & Cheng, 2011; Powell et al., 2013), although the differences are mostly subtle. At this point, I will save the reader from the empirical details, and conclude this section with a discussion why SS priors are an excellent to have to guide commonsense causal learning.

Recall the earlier remarks on how strong expectations lead to fast learning. The obvious implication of the SS model is that by assuming strong default priors, learning is similarly rapid in generic situations where no background knowledge is available. Hence, people are expected to jump to conclusions by witnessing small samples of confirming evidence. That is an obvious source of errors, but people should also drop their false beliefs quickly if induced expectations are not subsequently met reliably. In pathological environments, this might lead to the rapid adoption of false beliefs; however, the more obvious problem with SS priors is that they tend to dismiss genuine but weak causes. Therefore, causal induction with strong and sparse priors looks almost like an antithesis of scientific investigation, which stresses reliability and often seeks non-obvious causal links. Looking back to the beginning of this section, where the Bayesian framework was advertised as a rational model of causal reasoning, suspicions may arise that hence the SS model veers off from that tradition. But quite the contrary is the case. The causal induction theories discussed here are designed to model our innate everyday capacity to form causal beliefs from observing or producing events, without the help of explicitly learned mathematical tools. The practical demands in controlled scientific studies and everyday causal reasoning are often quite different, and we use statistical methods precisely because weak dependencies are difficult, if not

impossible, to spot casually without the aid of careful experimentation and analysis.

Recall that reasoning in Bayesian nets contains propositional logic as a special case. This means that, in general, computations are intractable because, for example, resolving whether a given assignment of three variables is possible in an arbitrary network is computationally NP-hard. That is, the time for the required computations grows at least exponentially with the size of the network in the worst case. This is what makes the modeling of commonsense reasoning intractably hard with logic. Any two variables or events may be indirectly connected in complex ways and verifying these links can be a daunting task in large networks, and the networks can get large pretty swiftly. Writing a complete truth table or a full joint distribution requires 2^n entries, given n variables. Thus, specifying a model with 30 variables takes 2^{30} numbers, which a bit more than a billion. In a fully connected network, the number of links is the same. Luckily, these worst-case scenarios haunt us only if the models are unrestricted. If we require that every variable most depend on some constant number of c other variables, that exponential growth becomes linear: $n2^c$. For example, if $c = 5$ we only need to specify at most 960 dependency relations, a millionfold decrease in complexity (Russell & Norvig, 2010, Chapter 14).

These kind of restrictions are conceptually problematic with logical relations because they are abstract to the point that basically every thing is related to every other in some way. Causal relations, in turn, are factual dependencies between specific variables and hence naturally restricted in occurrence. They are also local in the sense that every effect is strictly dependent only on its immediate causes. Moreover, logical relations lack any intrinsic way to specify relevance. With causal relations, the effect size (i.e. causal strength) can be taken to be an approximate measure of significance. Some *ceteris paribus* reservations are needed here, however. You may want to avoid severe but improbable outcomes, which makes weak causes sometimes relevant. Also, base rates and control matter. Strongly causative but uncontrolled and extremely unlikely events (such as being killed by an asteroid impact) are generally irrelevant to everyday decision-making. So are generally also vital but highly prevalent causes (such as the presence of oxygen as a necessary cause of a fire). In general, it is reasonable to focus only on the most predictive causes if you are interested in controlling or predicting a specific effect. Now, it might be that the world actually is an extremely complex nexus of small but genuine causal events, and hence the causal maps that emphasize strong and sparse causes yield a distorting representation of reality. This is undoubtedly true,

if we consider the physical microstructure of the world; however, think only of a casual walk through a crowded street and imagine all the macroscopic events that are happening all around. In principle, almost an endless variety of unobserved and remote events may interfere with our everyday plans. Still, everyday world continues to be highly structured and predictable, and all the underlying chaos mostly results in only minor events of no practical significance to our daily life.

Given these considerations, causal induction with SS priors should start to appear as a rather rational choice for tracking pragmatically relevant dependencies in an open and complex world. These priors reduce the number of causal relations by pruning all except the most significant dependencies making the subjective representation of the world streamlined and more manageable. Regardless of your sentiments toward the information-processing theory of the mind, complexity issues should always be something to concern about. Because of our inherent cognitive limits, increasingly accurate representation of reality should at some point turn to hindrance rather than an advantage.

Classical computationalist approaches have mostly resorted to some sort of satisficing strategies to manage complexity, such as heuristic algorithms that give suboptimal but good enough answer to complex problems (e.g. Newell & Simon, 1976). The causal learning model that is prejudiced to favor strong and sparse causes mitigates the problem right at the representational level by preferring a causal topology with the smallest set of weights that are sufficiently strong to predict relevant effects. Adding a link to a graph increases its accuracy; however, it is always a trade-off with growing complexity. A single link does no harm but if you are about to model a domain with numerous weak causes, you probably want to consider whether saturating the graph with these is a sound general policy if there are also more predictive and pragmatically relevant cues available. This would mean that our causal maps of the world are geared toward subjective and pragmatic rather than objective veridical representation, and it is the nature of probabilistic modeling that makes this possible without expressly distorting the reality. Discarding a causal hypothesis does not necessarily mean to accept a belief that the candidate cause is completely ineffectual but to downgrade its status to a part of the constant background noise. All the background causes are implicitly factored in the model as the base rates of the effects and the probability distributions of explicit links.

Indeed, the SS priors are defined over causal strengths, which does not directly advocate sparse structures. The principle strictly advocates only structural representations where causal strengths tend to be close to 0 or 1. The

effect of priors is maximal early in the learning, and the accumulating data can swamp their impact in the long run. Hence, the SS priors do not preclude the possibility of learning the actual strengths eventually. Neither is learning with default SS priors contradictory to Griffiths and Tenenbaum's (2009) theory of how specific priors guide learning after some causal knowledge has been established. SS model only makes unique predictions about how learning proceeds initially in the absence of knowledge when the sample size is small, and learners have no obvious reason for having any other specific priors about weights. (Lu et al., 2008b, 961) In time, more nuanced and accurate networks, employing more variables and more fine-tuned weight estimates, may be extracted when more experience is acquired in a specific domain. Although if this is the case, then the information about those causal events that are not incorporated into the causal belief system need to be retained in some way.

Most of the research reviewed above use static summary tables to introduce contingency data about C and E to subjects, but dynamic or sequential models of causal learning have also been developed. Usually they adjust the distribution parameters after singular observations and then discard the data (Holyoak & Cheng, 2011). It is somewhat hard to see how such models could retain the information about events attributed to unspecified background causes and learn about hidden variables.[98] An obvious way to implement sequential learning while retaining information about specific events is simply to keep a running tally of the contingency data, and to update the posterior probabilities over structures and strengths as you go. This involves counting the number of actual events and not only proportional frequencies of C and E, since the record of the sample size is needed. This technique was employed by Lu et al. (2008b) as a preliminary investigation with no particularly striking results. At least, the model should also account for gradual forgetting and keep track of the order of the events because there are asymmetries in learning due to order of observed events.[99]

[98] Different methods used for this purpose are mainly variations of non-Bayesian *Expectation Maximization* and related constraint based algorithms; see (Pearl, 2000; Luhmann & Ahn, 2007) and (Russell & Norvig, 2010, Ch. 12).

[99] For example, if cues A and B are paired with effect E, and subsequently subjects witness only A to produce E, they lower their rating about the causal strength of B and elevate A. If subjects first lean causal rule $A \rightarrow E$ and then they witness $AB \rightarrow E$, the previous knowledge about A blocks the learning about the possible influence of B. These *blocking effects* are interesting phenomena in their own right, demonstrating that human causal learning is not (at least straightforwardly) associative: In both learning situations, B is only positively correlated with E but people still disregard it in the forward condition (first $A \rightarrow E$, then $AB \rightarrow E$), and even make negative causal judgments about its relation

Assuming that keeping track of the order of events implies memory for separate traces, that would require exemplar or comparable encoding, but more investigation in this area is needed. Logically speaking, you could just keep track of posterior distributions of all the possible causal relations in case some weak link eventually shows up. However, that would be tantamount to tracking the fully connected network. In order to contain the combinatorial explosion, then, you would need domain-knowledge to inform what dependencies are meaningful to track (Tenenbaum et al., 2007; Penn & Povinelli, 2007; Holyoak & Cheng, 2011). However, you need the SS priors the most when you *do not* have such information. In the next chapter, I will show how that sort of knowledge may be engendered by exploiting analogical reasoning over exemplar-based event representations.

4.2.3 Section summary

Bayesian modeling of causal learning is a highly developing field with a host of open issues, alternative accounts, and not much consensus on details beyond the basics covered here. This is mostly true of causal cognition in general which, currently lacks a canonical presentation in introductory textbooks of cognitive psychology.[100] Therefore, more in-depth comparison of the various models of causal learning, reasoning, and representation would not be relevant for the purposes of this work, which is not supposed to be an exhaustive survey of the state-of-the-art. Anything beyond what has been said thus far is mostly theoretical speculation, but let's sum up what we have learned.

I have argued that category representations are feature clusters structured by causal knowledge—basically graphical models where, at the limit, there are no (known) connections between constituent features. Generally, though, representations of natural categories tend to contain causal links. This enables us to get a hold on the notion of category coherence, i.e., why some collections of features make sensible wholes. To grasp something is to understand how its properties hang together. Moreover, often the relevant causal properties are not strictly intrinsic to the entity, and critical part of category knowledge is to track how things participate in events. For example, *pain medication*

to E when it is absent in the backwards case (first $AB \rightarrow E$ then $A \rightarrow E$). This is quite incompatible with the core idea of associationism, which is that mental processes are sensitive only to co-occurrence of events; see (Beckers et al., 2005; Penn & Povinelli, 2007; Holyoak & Cheng, 2011).

[100] See the introduction in Waldmann (2017), perhaps the first entry level handbook of the subject, published just recently.

cures headaches, and this is a conceptually constitutive relational property that links pain killers to events. I have argued that, in general, categories are agent related, that is many important properties are determined by our goals and capacities, even to the point where specific motor programs may be encoded as a part of category representation. Theories discussed in this section are not specifically designed to address these matters but causal knowledge and learning in general. My aim in this section was to give some idea what the Bayesian inductive search for causally predictive regularities means, and to show that existing models provide viable, albeit imperfect, theories of human causal induction.

Epistemologically, it is significant that the causal structure of the world can be extracted from covariation of events, but that process is (naturally) fallible and requires generic presuppositions about the structure of reality. This is somewhat Kantian in spirit, in that the theory does not fit neatly into empiricist/associative tradition or the rational/inferentialist tradition in cognitive science but includes aspects of both. Often we can not directly observe these causes at work but they are there as part of the noise of unspecified weak background causes or unknown specific variables, which may strongly affect the observed events. Initial learning, in particular, is often guided by temporal order and more significantly by observing and producing interventions. Interaction is particularly important for producing variety and control over the data, and for discovering independencies and hidden variables (Steyvers et al., 2003; Lagnado et al., 2007). Hence, elementary causal knowledge is often produced by active participation in concrete situations, and therefore it is practically warranted that much of it focuses on pragmatically relevant events and regularities. After some generic knowledge of event types is gathered by these general processes, an increasingly efficient grasp of (domain-)specific novel events is attained. Domain general learning mechanisms are apparently guided by a default policy to extract only a crude but sufficiently predictive causal map of reality, shaping our causal understanding of the world as primarily pragmatically oriented instead of toward veridical objective representation—although these two are not necessarily in any way contradictory. The next step then is to explain how increasingly abstract domain knowledge is produced from these concrete and specific representations.

In outlining the main empirical hypothesis of this work, I postulated that situations are mentally represented as quasi-perceptual model type structures, but I have not thus far said much about that. The reader may also begin to wonder what was all that fuss about skills and conceptual understanding,

because so far we have seen mostly standard cognitive psychology of knowledge and category representation. In the next chapter, I will explain how these aspects of cognition are related. Cognitive skills and understanding are both about how all this knowledge actually gets used. It is important to understand that the Bayesian framework is no silver bullet to the complexity issues posed by commonsense and expert reasoning. Learning and inference with complex causal networks can be computationally very hard, and Bayesian methods are often considered to pose unrealistic processing demands on human cognition (Gopnik et al., 2004; Knill & Pouget, 2004; Lagnado et al., 2007, e.g.). Computing the integrals of posterior probabilities is one thing, but the more pressing issue is that when we move beyond elementary causal induction to wide domains where the number of variables grows, the complexity of the resulting hypothesis spaces may undergo combinatorial explosion. That is, we are no longer comparing two graphs but posterior probabilities of all the possible graphs, and suddenly that can be computationally infeasible. This is also not what people seem to do behaviorally. When subjects can produce data by interventions, they seem first to arrive at a preferred hypothesis through initial observations and then produce data to optimally investigate that particular hypothesis rather than the whole hypothesis space (Steyvers et al., 2003). Sparse causal models and the guide of established knowledge conceivably helps to manage complexity, but commonsense knowledge is still very intricate. The problem gets worse when we need to deal also with abstract level of regularities that are not directly associated with basic-level categories and elementary causal knowledge.

Note that everyone in the Bayesian modeling business believes that causal induction happens as an implicit process. Very few adults have any knowledge of Bayesian methods, and probably no one can do all the necessary computations explicitly in their minds. Implicit cognition is generally taken to have a high capacity parallel architecture, and it is also a realistic possibility that probability density functions can be computed quite straightforwardly by using population encoding over groups of neurons (Knill & Pouget, 2004). Moreover, causal networks quite easily translate into parallel signal passing processing models where each node computes only sparse local information (i.e. joint distribution contingent only on immediate causes in Bayesian networks). This means that, conceivably, the brain may be optimized all the way down to "hardware" level for computing probabilistic networks. However, that capacity necessarily has limits, and it is a sound practice not to attribute intractably complex computations to any cognitive faculty. When the potential hypothesis

spaces grow, and we are dealing with flexible use of causal knowledge of more general level of abstraction than elementary causal relations, the issue still boils pretty much down to selecting relevant information to focus on—just like with logic based approaches. Bayesian inference is not a theory of cognitive processing and hence it does not readily specify how such information access and filtering might be implemented. Moreover, Bayesian nets are inherently not expressive enough to represent higher-order domain knowledge for narrowing down hypothesis spaces (like schematic principles contained in particular (folk) theories), since they are limited to state only specific relations between specific variables while sometimes you may need to express more general principles concerning event or variable types (Tenenbaum et al., 2007).

The reader should recall the discussion in the opening of this section, in that the Bayesian framework aims to provide a normative analysis of reasoning and behavior in a complex and uncertain world. The approach is not committed to any specific cognitive mechanisms that would produce this behavior but only to use Bayesian models of rational inference to conceptualize it. Regardless of whether that is an accurate description of the field, the paradigm is consistent with alternative suggestions about how causal learning and inference actually executed, perhaps by approximating unconstrained Bayesian inference. Note also that the research discussed here is mostly confined to elementary causal induction for quick and dirty causal maps of the world. It is possible that this elementary induction is implemented by exact Bayesian inference in the brain while slowly developing, more fine-grained, and complex capacities employ different methods.

In the next sections, I will propose something along these lines. The last technical issue we left open before this section summary was how the cognitive system retains rich contingency information about weak dependencies that are attributed to the constant background. I suggested that one way to do this is to retain exemplar representations about events. This is by no means the only viable solution, but there is independent evidence that exemplars play a significant role in conceptual cognition, and specifying their status in causal knowledge may help integrate various research programs such as causal induction, skill learning, categorization, concept learning, and analogical reasoning. The last item will explain how abstract knowledge is rooted in concrete event representations. This proposal, which is far more programmatic than the discussion so far, is pursued in the rest of this work. By now, we have seen how prototypes and causal knowledge are integrated, and what follows will outline a role of exemplars and specific situations in conceptual cognition.

4.3 Event representation, situations, and cognitive skills

Up to this point we have basically covered how human cognition represents information about concrete categories. Next, we need an account of event or situation representations and explain how more abstract schematic knowledge is attained. The latter is the primary goal of the next chapter. Here we start with the former but, roughly, the idea developed in the next chapter is that schematic understanding is an incremental product of acquired procedural knowledge. As explained earlier, procedural knowledge is simply a collection of long term-memories encoded as situation-action-outcome exemplars that are activated by a shallow associative memory search, cued by external factors and endogenous goals in specific situations. Hence, schematic knowledge is ultimately anchored to concrete knowledge, but more abstract schematic understanding can be produced by automatic activation of specific memories, which are mentally linked via procedural information. These recurring procedural patterns can become acknowledged, explicitly represented, lexicalized, and eventually form their own conceptual representation. Nevertheless, the intelligibility of the associated schematic concepts depends on the knowledge of the specific events constituting those patterns in the subject's memory and experience.

Although the idea is not tied to prototype theory, it is influenced by the research on basic levels of category representation discussed in the earlier sections. On the whole, the argument exploits insights from a wide variety of cognitive psychology, for example the early exemplar theory of categories (Medin & Schaffer, 1978; Brooks, 1978), instance theory of automation (Logan, 1988; Palmeri, 1997), theory of mental models (Johnson-Laird, 2008), neo-empiricist theories of concepts (Barsalou, 1999; Barsalou et al., 2003), research on reminding in problem-solving (Ross, 1984; Ross & Kennedy, 1990; Brooks et al., 1991), recognition theory of cognitive skills (Simon, 1992; Klein, 1998, 2008), implicit learning research (Reber, 1993; Sun et al., 2005), research on analogical reasoning (Gick & Holyoak, 1983; Novick, 1988; Holyoak et al., 1994; Gentner et al., 2003), and, to a degree, the representational redescription model of developmental psychology (Karmiloff-Smith, 1992). The core assumptions about fundamental cognitive processes and control are essentially identical to the ones employed in recent research on prefrontal cortex function in learning and decision-making (e.g. Collins & Koechlin, 2012; Collins & Frank, 2013; Domenech & Koechlin, 2015).

4.3.1 Event representation and simulation

I assume that events can be represented as causal models, akin to graphs introduced in the previous section, which are composed of any variables that are relevant to the agent in terms of goal achievement. Often these variables are objects and actors, and most superficially, events can be characterized simply by itemizing present things and their properties. However, usually events are specified by what is happening rather than by the entities present. We can transform such lists to causal models by specifying how the items and their properties interact. This is implicit in such descriptions anyway, as far as entity types are characterized by causal attributes. In any case, events can be abstractly represented as causal models, which can take category causal models as variables. This definition is supposed to be as general as possible. As a limiting case, categories can be represented by simple unstructured feature vectors, and vectors can consist only of a single variable. Formally, this characterization allows event representations to have anything from a single-valued variables to complex structures as constituents. Generally, we tend to factor our environment in causally relevant elements, and since we are born to track what is happening, the salient causal properties tend to be informative for event–event causation. Full event descriptions then point to other subsequent events, which are the possible outcomes of how things unfold. I refer to these causal models with event–event structure as *situations*.

While this representation scheme may apply to an objective depiction of reality, "situation" often refers to subjective conditions. To translate the characterization into an empirically adequate notion of mental models—i.e., subjective representations of states of affairs, we need some qualifications. First, situations in this sense are partly determined by the environment and partly by the agent. While external physical conditions obviously are constitutive of events in general, situations are properly construed from a point of view. In other words, basic building blocks of situations are not objects as such but elements of the agent's conceptual system. Hence, the same stimulus environment can afford different models for different agents. Thus, in principle, representations of external conditions can even be incommensurable between agents. In practice, however, similar lifeforms tend to perceive many affordances similarly and shared cultural practices tend to produce congruent conceptual systems. In any case, different variables and event–event contingencies may be differently salient and relevant to different actors, and hence external conditions bear one-to-many relations to mental models. Second, we enter situations

with our transient needs and inclinations. Different goals make us focus on different affordances and expected outcomes, affecting the interpretation of external factors and even how we categorize objects. Hence, the same agent may produce different models of the same stimulus environment, depending on its current mindset. Lastly, both of these planks go similarly with all of our background knowledge that we bring into situations; in other words, what we know by our previous experiences modulates our anticipations and what we perceive as salient and relevant. In brief, how we represent external factors as situations is not straightforwardly determined by the physical or stimulus environment alone.

The above planks are familiar from our earlier discussion about concrete category representation. The point that I am after this time is somewhat different: It is wise to abstain from the firm *a priori* commitments to the ontology of situations. It is enough to think of situations as concrete events that are constituted by specific variables whatever they might be. Think, for example, planning a night out on Friday. Clearly, not all the relevant variables can be naturally construed as objects, actors, or their properties: the weather, your bank account balance, the length of your evening meeting, and so on. Bayes net formalism does not inherently differentiate between propositional, event, object, or property variables. Causal learning literature rarely specifies the general type of the variables in the mental representation of events and objects, and I think it is advisable to follow suit here.

It is theoretically motivated to choose such a policy of minimal commitment. Many categories are, to various degrees, characterized by how they participate in event causation, and earlier we discussed the inseparability of category identity from events (e.g., pp. 95–100). Thinking of event representations as causal models similar to category representations (i.e., as graphs with ontologically neutral variables) provides a natural means for such intermixed representations. It especially negates the need to postulate new representational devices for event representations. Second-order (event) models simply exploit the representational capacities of the first-order (category) models. Seeing things this way also enables us to import theoretical elements from category to situation processing. Similar to sparse causal maps and basic level category cuts that are produced by agents in accordance with their capacities and needs, situation representations are likely comparable coarse constructions where the elements and the level of grain are set by the agent on the pragmatic and contextual basis. Moreover, event categories may be processed similarly to category models: In case no causal information is available, they are character-

ized by correlations or family resemblance structure of characteristic elements. When causal information is available, it primarily guides event classification and events can be grouped with other events that share the common causal structure. If values of relevant variables are unknown, their default or expected values can be retrieved from the generic model or computed from the causal model based on other known values. Notions such as event category coherence can be derived from this scheme in an obvious way.

Nonetheless, there are apparent differences in object and event knowledge. Often one does not have the means to manipulate the causal factors that produce intrinsic property correlations in concrete categories. This is particularly true of non-artifact kinds. Hence, the category structure often needs to be inferred from passive observation in addition to background knowledge about mechanisms. On the other hand, we often intervene in event–event causation, and many important things in our environment are causally inert unless we work as causative agents. As discussed in the previous section, these interactions are especially important for discovering novel dependencies and hidden causal variables. Assuming that events are precisely individuated by such information contextualized—i.e. events are also characterized by intentional action with more or less specific goals—then event and object representations are mutually dependent, at least to a degree. It is still reasonable to assume that there is at least a quantitative difference in the types of knowledge that is coded in object and event representation. The former probably consists much of object-centered correlational information about property instantiations, while the latter is more characterized by agent-centered pragmatic knowledge about event–event contingencies. The relevant contingencies in novel situations may be obscure to novices and their experience with the variability of possible variables and values (e.g., the type of entities and their properties) limited. Therefore, we should expect novices to conceptualize situations on based on superficial features whereas experts to focus more on the functionally relevant deep structural features, which indeed is the case (Novick, 1988; Campitelli & Gobet, 2010).

Unfortunately, only limited research has been conducted to address the level of abstraction at which event variables are conceived, but the data we have indicate that people associate events labels with superordinate level categories (Rosch, 1978; Rifkin, 1985). For example, in one preliminary study, Rosch (1978) asked the subjects to describe the events of a particular day, and the participants consistently mentioned general activities such as getting dressed (with an implicit reference to superordinate categories—in this case

clothes) instead of more accurate activities such as putting on trousers (a basic level category). More encompassing events such as "all the morning chores" did not appear to have memory representation separate from these more specific procedures. Thus, it seems that events are primarily conceptualized on the level of schematic behaviors and other causal events, which can be broken down into more basic actions (e.g., *putting on pants*, *putting on shirt*,...) which do not further decompose naturally into more basic linguistic elements but idiosyncratic motor programs or something alike, which can perhaps be demonstrated but not easily described.

By insisting that event representations are quasi-perceptual, I mean that their default variables generally are basic level categories instantiated by specific entities. The basic level, remember, is supposed to consist of most inclusive categories that share common affordances, sensory properties, and motor programs. Apart from perhaps schematic affordances, superordinate categories tend to lack these attributes altogether. Think of *clothing*, for example. Different kinds of clothes (e.g., trousers and gloves) rarely share similar shapes or other common visual identifiers, and the same applies to specific affordances or actions associated with them. A central feature of clothing is that you put it on, but this property is multiply realizable: you put gloves differently from how you put trousers on. From a goal-derived perspective, you generally use trousers (at least partly) for different purposes than gloves. Hence, functional properties associated with superordinate categories are quite non-specific and abstract. The reason for this becomes clear below, to for my argument it is relatively important that functional and other not strictly perceptual knowledge hangs together with perceptual and motor representations. That basically is what I mean by that the fundamental level of learning and conceptual comprehension tracks concrete entities and events. "Quasi" here reminds us that not all information associated with concrete things are perceptual (e.g., causal ones) and neither situations need to be represented in full detail. This is evident in explicit thought for we tend to think and remember the general gist of events while dismissing many surface details. However, perhaps a bit surprisingly this seems to be true even *in situ*. When we perform tasks, we do not necessarily extract a detailed representation of the environment prior to acting but often, in the words of Rodney Brooks (1999, 81), use the world as its own model, and shift our focus back and forth to retrieve information online on a need-to-know basis to optimize working memory load and information retrieval time during execution (Clark, 2008, 11–13,118–122).

Now, my case may begin to sound slightly confusing. If the ultimate purpose of conceptual cognition is to track event–event contingencies to support intentional action which is, in turn, conceptualized at a schematic level, then why am I claiming that core event representations are quasi-perceptual models with basic level objects as default variables? In other words, if the basic building blocks of conceptual thought are structural and functional properties (as Mandler claims) and if superordinate categories are elliptic descriptions of functional attributes (as Rosch claims), then why I am using findings of the basic level categories to make my case? Would it not be more appropriate for me to expect the *superordinate* concepts to be fundamentally relevant and consider perceptual features and other such properties as mostly irrelevant residual knowledge? I gather that something like this is a common attitude amongst philosophers, especially the ones with essentialist inclinations. Indeed, I have apparently somewhat backed down from the concreteness or "quasi-perceptual" claim by insisting that we should make strong ontological commitments to the fundamental elements in situation representation.

It is easy to find counterexamples to the claim that pragmatically relevant situational variables are specific and concrete because as adults we have thoroughly learned how generic situations in our everyday environment work. Therefore, we have a fairly good grasp of these recurring patterns, which are not always characterized by specific contents but rather by functional features. For example, everyone understands the sentence "in order to get a permit for P, you need to fulfill an obligation O," even though P and O are nonspecific variables. In the next chapter, I will explain how these kinds of schemata are learned by clustering situation exemplars that constitute procedural knowledge. The role of specific exemplars becomes more obvious once we track the development of cognitive skills but the main point is similar to the one explained above: The comprehension of unfamiliar situations tends to depend on specific surface content while accumulating expertise provides a gradual understanding of a more structural kind. People tend to conceptualize their actions with a generic purpose in mind, but the level of abstraction depends on task difficulty. Novel tasks demand attention to specific actions; however, eventually such minute routines become processed unconsciously and chunked to more complex behavioral units while focus and control shift to a more generic "gist" level of understanding. If execution faces difficulties, comprehension relapses to more concrete details. (Vallacher & Wegner, 1987; Christensen et al., 2016) Thus, while event *types* are individuated by nonspecific functional knowledge, this conceptual capacity is grounded on processing episodes dealing with spe-

cific event *tokens*. Since the basic level is the default level of individuating things, with which we directly interact, one should expect that the default elements of these processing episodes are conceptualized accordingly.

The key claim here is more or less the core tenet of *situated reasoning*. However, I do not claim that every reasoning episode must happen in a context where the relevant set of variables are physically present in the current local environment. What I claim is that the human ability to reason on the higher level of abstraction does not result from logical knowledge about abstract task structures but from procedural knowledge that is initially tied to perceptually guided action. That capacity utilizes representational tokens that make possible for the agent to imagine specific situations and behaviors. Superordinate schematic expressions are understood by translating them into comprehensible concrete world knowledge, i.e. events involving specific acts with basic categories (Rosch, 1978). Children tend to learn novel basic level categories sooner than functionally determined superordinate categories (Rosch et al., 1976), and object concepts are learned before verbs that refer to structures and change (Gentner & Boroditsky, 1999); suggesting that schematic knowledge is extracted from concrete knowledge.[101]

I presume that causal properties and schematic structural attributes such as *above/below* or *containment* are not even comprehensible without reference to concrete objects or situations. If infants (or adults) extract structural and functional properties from the environment, they necessarily need to do this in specific contexts where the relations are implemented by concrete entities.[102] I personally find these relational notions semantically transparent yet impossible to imagine or describe without some specific concrete content—and so did the subjects of Rosch et al. (1976, 409). Well, inexplicability may be what being basic really means but, then, they *can* be explicated by examples referring to concrete situations. The importance of functional properties for mental

[101] Note that this does not mean that causal and other functional knowledge is learned later but that functional knowledge can be more easily dissociated from specific instances after learning. Infant basic levels may correspond to adult superordinate levels in *scope* but not necessarily in content. Therefore, the claim that *purely* functionally determined superordinate categories are learned later does not contradict the claim that basic levels for infants might correspond to superordinate classes in adult ontology. Infant conceptual taxonomy may be just more undifferentiated and their functional attributes more robust. But from infants' point of view, their basic level may be similarly specific and concrete and not determined solely by abstract function.

[102] Mandler's theory does not contradict this obvious point. The central part of her account is that basic contents are derived from sensory experience (Mandler, 2004).

representation is not in question but that they by themselves could engender intelligible content. An ability to track such functional attributes is a precondition rather than sufficient condition for making sense of the world. Here, I am iterating the point from page 142 that sensory-derived contents, associated actions, etc., do not need to correspond to anything that *have* conceptual meanings; instead, they are constituents in the systems that *are* meanings. The general point is that categories are aggregates of functional knowledge, but that knowledge needs to be anchored to concrete things. Merely procedural narratives are very difficult to comprehend. Here is an example (Bransford & Johnson, 1973, 400):

> The procedure is actually quite simple. First you arrange things into different groups. Of course, one pile might be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo things. That is, it is better to do too few things at once than too many. In the short run this may not seem important but complications can easily arise. A mistake can be expensive as well. At first the whole will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell. After the procedure is completed one arranges the materials into different groups again. Then they can be put into their appropriate places. Eventually they will be used once more and the whole cycle will then have to be repeated. However, that is part of life.

People tend to find it hard to grasp what is going on here unless one thinks of *washing clothes*, for example. After that, the initial confusion is perhaps difficult to evoke again. This results from our automatic ability to translate the narrative into comprehensible world knowledge—i.e. sensible practical acts with concrete things—and a sort of ”click of comprehension” is produced (Rosch, 1978, 20). In summary, whatever is the role of sensory and motor contents in conceptual content, certain specificity and concreteness are prerequisites for understanding events.

The other line of argument for quasi-perceptual processing in thinking comes from neo-empiricist theories of concepts and related model and simulation accounts of reasoning. These theories further claim that it is not only concreteness of content that matters but also how the content is represented

in thought often in the same way as in actual perception. For example, when people read a passage about a nail pounded either into a wall or the floor, subsequent processing about the nail is faster in the implied rather than unimplied orientation (Barsalou et al., 2003, 86). The idea that thinking and reasoning involves some kind of analogical model construction and manipulation is not new. It has roots in the 19th-century science and pragmatist philosophy.[103] The first clear articulation of mental models in the modern sense can be found in Kenneth Craik's *The Nature of Explanation*:

> If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to the future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which faces it. (Craik, 1943, 61)

According to Philip Johnson-Laird (2008, 428), perhaps the most prominent advocate of the theory, a mental model "[...] represents what is true in one possibility, and so far as possible has an iconic structure. Mental models are the end result of perception and of understanding a description." The iconic or image-like character of thought, "as far as possible," is repeated throughout Johnson-Laird's work. Unfortunately, he does not make much use of this idea nor of the internal structure of mental models; instead, he focuses on reasoning over sets of alternative models—usually represented by symbolic propositional variables. This not intended as a criticism of his work; however, although Johnson-Laird appears to share the spirit of knowledge representation advocated here, he is more interested in explicit logical inference than implicit situated reasoning.

To be sure, Craik's ideas about human cognition are closer to the standard logical cognitivism than the quotation above suggests. However, the quoted lines capture the essence of the mental models approach and the closely related notion that thinking is a sort of trying out—i.e. simulating—one's own action in imagined situations. This idea also has long roots in empirical psychology, e.g., in the works of behaviorist Edward C. Tolman (1932, Ch. XIII).

Recently, the notion of simulation has surfaced in connection with causal models and analogical reasoning (e.g., Hagmayer & Waldmann, 2000; Rehder

[103] See Johnson-Laird (2004) for a historical overview.

& Kim, 2010; Holyoak et al., 2010), but unfortunately it is rarely fleshed out in more detail beyond elliptical references to setting up and running a causal model in one's mind. The idea is, I presume, easy to understand as a forward inference in Bayes nets. However, that is not very restrictive characterization, since these networks can be used to implement, *inter alia*, logical inference. Indeed, since models are not supposed to bear isomorphic but one-to-many relation to their targets, language-like representation can be thought of as special cases of models and any (rule based) predictive inference as a sort of simulation (see e.g. Hegarty, 2004).[104] As mentioned above, we do not model the environment accurately even *in situ*, because selective attention to important features isolates selected information from the complete stimulus environment in perception and action (Barsalou, 1999, 358). Hence, it does not follow that rich and detailed representations are employed when we faithfully re-enact the mental representations invoked in the past events, and probably even less so when we imagine future ones based on our previous experience.

Therefore, the notion of simulation as such is not very informative and it needs to be made more precise to be theoretically useful. A common reference in this connection is Barsalou's theory of perceptual symbol systems and grounded cognition (Barsalou, 1999, 2008; Barsalou et al., 2003). The general idea is that category representations are "simulators" which, when activated, partly recreate the original sensory and motor experiences associated with the concepts. These simulators ground human conceptual cognition and support categorization and category-based inference. They can be combinatorially integrated to produce complex simulators (e.g., models). I will leave aside Barsalou's account of abstract thought; nonetheless, we share roughly the similar idea that schematic representations are understood by binding concrete (simulator) tokens to abstract type variables in schematic expressions (see Barsalou, 1999). In his words:

> Simulation is the reenactment of perceptual, motor, and introspective states acquired during experience with the world, body, and mind. As an experience occurs (e.g., easing into a chair), the brain captures states across the modalities and integrates them with a multimodal representation stored in memory (e.g., how a

[104] My reasoning is that generally (computer) simulations express aspects of the target phenomena in terms of differential equations, for example (which are symbolic expressions), and then use those equations to infer how the target system behaves under some specific conditions. Recall that also in the model theory of predicate logic, models are expressions in set theoretical language.

> chair looks and feels, the action of sitting, introspections of comfort and relaxation). Later, when knowledge is needed to represent a category (e.g., chair), multimodal representations captured during experiences with its instances are reactivated to simulate how the brain represented perception, action, and introspection associated with it.
>
> [...] From this [situated action] perspective, the cognitive system evolved to support action in specific situations, including social interaction. These accounts stress interactions between perception, action, the body, the environment, and other agents, typically during goal achievement. (Barsalou, 2008, 618–619)

We have discussed at length the constitutive role of perception, affordances, and action in category representation as well as the role of specific content in causal reasoning and event comprehension. However, the need for modal specific sensory or motor contents in conceptually structured thought is not strictly implied by these considerations. The pioneering studies about visual rotation (Shepard & Metzler, 1971) and scanning (Kosslyn, 1973) of mental images instigated a lengthy and well-known debate about quasi-perceptual thinking. We do not need to concern ourselves with that particular discussion or any issues considering introspection (or lack thereof) of mental imagery. Remember that here the notion of mental image does not exclusively refer to visual but sensorimotor representations and processes in general, and these images do not need to be conscious any more than any other cognitive phenomena.

While unconscious imagery may sound like a contradiction of terms to some, the sensorimotor system is a complex thing, and most of its processing happens outside the reach of conscious access. More to the point, if routinized details of perceptually guided action are processed unconsciously during execution, while conscious attention is focused mostly to the action gist (e.g. Vallacher & Wegner, 1987), it should be natural to expect that the same holds for the simulation of the event. Hence, mental simulation can be rather impoverished and more detailed elements may be brought to conscious attention only when necessary—just like in actual lived situations, where we can focus on selected details when needed.

The issue here is not conscious imagery but whether sensorimotor faculties have a constitutive role in thought. This would mean that thinking about concrete categories and situations exploit (at least partly) the same modal-specific cognitive resources as perception and action. A substantial amount of

neuroscientific evidence shows that it does.[105] Some theorists (e.g. Pylyshyn, 1981) argue that mental images are epiphenomenal and have no causal role in thinking. The same may hold true for the activation of sensory and motor areas associated with thinking. However, damage to sensory and motor areas cause selective deficiencies in conceptual processing, implying that these resources do significant work in category processing. For example, damage to visual areas disrupts the processing of categories that usually characterized by visual features (e.g. *birds*), and likewise damage to motor and somatosensory areas disrupt processing of *tools*, for example (Barsalou, 1999, 585). An ability to construct, maintain, and transform conscious spatial representations is highly correlated with the ability to solve certain mechanical problems, suggesting that conscious mental imagery may be important. This evidence, however, is more consonant with the interpretation that non-visual analogical reasoning is involved, which exploits piecemeal causal knowledge instead of mere holistic inspection of internal images (Hegarty, 2004).

These results do not warrant that even concrete category representations are strictly constituted by low-level sensory and motor properties. For example, the human visual system contains high-level neurons that code qualitative information about the presence of edges or lines without a particular length, position, or orientation. These schematic components may be further integrated to category content which, nevertheless, cannot be represented by any *specific* visual features (Barsalou, 1999). Imagination and perception utilize partly different resources in the way that thinking with images activates perceptual circuits only at the necessary level of specificity. If one wants to infer whether German shepherd dogs are larger than elephants, only higher-level integrated feature detectors may be activated; but if one thinks of whether a German shepherd has pointy or floppy ears, parts of the early visual cortex become active, which are sensitive to fine-grained details lower in the abstraction hierarchy (Pearson et al., 2015). This may explain the phenomenological difference between rich perception and less vivid thought (Brogaard & Gatzia, 2017).

A problem in these considerations is that everyone believes that high-level representations, activated upstream from perceptual areas, are essential for conceptual thought. The only question is do these representations turn into amodal at some point in the processing pathway. I must admit that I am unsure on this; however, *pace* Barsalou, I do not believe that all conceptual content is entirely modal. For example, causal knowledge is not strictly perceptual or

[105] For summary, see e.g. Barsalou et al. (2003); Barsalou (2008); Brogaard & Gatzia (2017).

motor and neither is presumably all the procedural (e.g., inferential) knowledge associated to abstract concepts. Nevertheless, I would rather abstain from this discussion. The standard philosophical argument against naive concept empiricism goes that imagery underdetermines content. To depict fruit, you need to depict a particular fruit. But no image of, e.g., lemon determines if it is about a particular fruit (lemon), a particular type of fruit (citrus), or a symbolic token of fruit(ness) in general—a concept which perhaps cannot be pictorially represented at all. Likewise, if you have an image of Urho Kekkonen, nothing in the image as such conveys the fact that it depicts the 8th and longest acting president of Finland.

As explained above, neo-empiricist theories of mental content are not committed to such naive resemblance theory. However, insisting that in the final analysis cognitive contents must be always reducible to somatosensory neural activation begs for needless philosophical problems. The reader should be aware that the theories discussed in this section are somewhat controversial. It is not at question if sensory and motor processing are involved in conceptual thought, but whether that is the whole story and if the question is even strictly empirical.[106] If we insist that all the higher-order representations are perceptual because they are (in the right circumstances) activated by perceptual input, the claim is weak and likely misleading. Proponents of symbolic mental representations believe that mental symbols are activated this way, and especially causal theorists, such as Fodor, take such causal facts as a cornerstone of their theory. We saw in Chapter 2 that it is theoretically difficult to identify content by such causal covariation. The problems arise especially with abstract content. Because of these problems, Johnson-Laird insisted that mental models are as iconic as possible, but some content is necessarily symbolic. Now, it may be that symbolic content is mentally attached to linguistic tokens that have (multi)modal mental representation, but very few would think that linguistic symbolic content is constituted by the visual or auditory representations of words, for example, rather than inferential and other knowledge associated with them—even if auditory, visual, or (pre)motor processing is factually necessary to use language.

So what good are simulations for? The idea of simulation is basically that of mentally recreating a situation and letting the implicit cognitive processes to do their work. Simulation works essentially as mental self-stimulation that prompts the reward and outcome expectations normally cued by actual perception (Decety & Grèzes, 2006; Gilbert & Wilson, 2007; Barsalou, 2008;

[106] See Machery (2007) for a critical review.

Benoit et al., 2014). Frontal areas of the brain are involved in preparation and control of action. There are extensive neuronal connections feeding back from frontal areas to sensory cortex, and this loop supposedly enables thinking as long chains of simulated actions and perceptions (Hesslow, 2002). I believe that imagining contexts and actions, therefore, works as a memory probe to activate specifically procedural and causal background knowledge. Hence, what these stimulations activate need not be conscious, holistic, or detailed image-like representations even if they employ knowledge extracted in sensorimotor interactions. Instead, simulations are often piecemeal and sketchy, and they can lead to correct reasoning even when the agent lacks correct explicit descriptive knowledge (Hegarty, 2004).

Implicit cognitive capacities that support automated situated reasoning are extremely effective; however, they can not be accessed directly and this sort of self-stimulation is a way to access them indirectly. As a downside, the process depends heavily on specific experiences and how we remember past events. It is essentially like a future-oriented retrospection that can be distorted and, in the words of Daniel Gilbert and Timothy Wilson, "mental simulation is the means by which the brain discovers what it already knows" (Gilbert & Wilson, 2007, 1354). Thus, simulation is embodied and situated reasoning re-enacted. This idea is not pure neo-empiricism but a combination of its core ideas with theories of implicit learning and memory and perceptual recognition theory of cognitive skills. If this is correct, mental images or models involved in simulations do not convey content by resembling their targets. It turns out that mental images and related sensorimotor representations are instead probes to pragmatic inferential content. Hence, the idea is immune to classical arguments against empiricist resemblance theories, and second we do not need to assume a strong form of neo-empiricism that takes into account only sensorimotor content and simulation. Depending on personal experience and specific tasks, some commonsense reasoning may very well also involve the use of explicit rules (Hegarty, 2004).

4.3.2 Procedural knowledge and cognitive control

Perhaps the most long-standing result in the psychology of expertise is that experts make decisions mostly by rapid pattern matching rather than by explicit search through the problem space. Search means systematically figuring out, step by step, how to reach a specific goal from a given starting point by applying a given set of means. In contrast, pattern recognition is a rapid, implicit,

and automatic process that exploits stored knowledge rather than intricate inference. This finding of expertise goes back to Adrian de Groot's (1965) influential research on chess players. His main result was that grandmasters do not think of their moves more than the less competent players. This does not mean that they do not spend time calculating their moves or that thinking is irrelevant in chess but that their advantage is due to intuitive preselection of more relevant options to consider. He also found that more competent players showed better recall of board positions that were displayed to the subjects for a few seconds. While masters could reconstruct the positions of more than 20 pieces out of 25, novices managed to recall only about five pieces correctly.

William Chase and Herbert Simon (1973) explained that this capacity results from expertise dependent ability to chunk individual pieces into larger perceptual units of about five pieces. They found that veridical recall requires the arrangements to be meaningful rather than random positions. They interpreted this result to demonstrate that highly developed skills—at least in chess—depend on perceptual ability to literally see different things on the board. What changes with experience is not the low-level functioning of the visual system but the knowledge base that perception activates, which, in turn, triggers action (Larkin et al., 1980; Campitelli & Gobet, 2010). Since novice and expert performance in the memory task is almost identical if chess pieces are arranged randomly (Gobet & Simon, 1996),[107] the expert ability is not dependent on better (visual) memory but on domain knowledge and the familiarity with the stimuli. The phenomenon is not confined to chess. For example, expert physicians display the same memory phenomena with random versus meaningful patterns of symptoms when compared to medical students (Palmeri, 1997, 346). One can think of these perceptual chunks as similar to categories, that is patterns of features that carry certain (pragmatic) meanings. In chess, their contents are just confined to the specific game that constitutes an isolated system of such meanings. For novices with little know-how of the game, pieces tend to appear as unconnected units while for experts, specific arrangements signify certain meaningful situations in the context of a whole game. Desirable moves become associated with such chunks and eventually the whole board positions rather than calculated for individual pieces, and this

[107] N.b. that the main result of Gobet & Simon (1996) is as per the title: "Recall of rapidly presented random chess positions is a function of skill." That is, experts do show better recall of random board positions than novices. However, the effect size is very small; hence, while the study technically corrects the original finding it still practically corroborates it.

transforms how the board positions and the game are understood (Dreyfus & Dreyfus, 1986, 33–34).

One of the key observations was that experts often cannot explain very detailed reasons for their superior performance, and so the knowledge they acquire is mostly tacit know-how rather than explicit know-that (de Groot, 1965; Dreyfus & Dreyfus, 1986). Although the proportional role of controlled versus automatized processes remains debated (Christensen et al., 2016), these results formed the backbone of modern expertise research: high-level skills are incremental products of vast amount of learning (up to tens of thousands of instances), and they result in a consciously inaccessible capacity to (a) automatically focus on only relevant factors and (b) make competent decisions and engage in reasonable course of action without reflection. This does not mean that experts do not additionally employ different problem-solving strategies compared to novices, but selection of those strategies is also intuitively prefiltered in routine tasks. This makes expert performance different and more effective since experts do not spend much time figuring out how to figure out the problem (Larkin et al., 1980; Chi et al., 1981). Such intuitive competencies are far from mysterious. In fact, they are mostly perceptual capacities, which are rather straightforwardly explained as resulting from a quick access to memory traces of recurring patterns. Therefore, intuitive decision making is in many ways similar to the intuitive interpretation of stimuli (i.e., categorizing). I quote Herbert Simon the second time: "The situation has provided a cue; This cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition." (Simon, 1992, 155)

There are two dominant contemporary paradigms addressing expertise: *naturalistic decision-making* (NDM) and *heuristics and biases* (HB) tradition. NDM has mostly been developed by Gary Klein and his colleagues (Klein et al., 1993; Klein, 1998). It builds on the works of de Groot, Chase, and Simon and emphasizes the reliability and power of human expertise in real-world settings. The basic idea in Klein's model is familiar from our previous discussion. Experts do not compare options but simulate a course of action in their minds and see if it works. If yes, they implement the action. If not, they modify the simulation. If modification seems hard or impossible, they search for another plausible approach. Mental simulation is efficient since it utilizes tacit knowledge and recognition-primed decisions. Given the processing limits of human reflective cognition, following explicit rules and evaluating several options cannot achieve such results in real time with complex and vaguely defined tasks

in the face of multiple constraints, uncertainty, rapidly shifting conditions, unanticipated disturbances, and so on. Moreover, very subtle cues—that often cannot be explicated but need to be experienced—may inform an expert that things are not proceeding normally. In Klein's famous case study, a firefighter lieutenant aborted a mission and pulled out his crew from a seemingly routine kitchen fire because he felt that something else was wrong besides that the fire did not react normally to extinguishment attempts. In hindsight, some untypical features were that the living room was extraordinarily hot and the fire quiet. It turned out that the flames were actually soaring from the basement, and the floor collapsed once the crew exited the building. Klein's book (1998) contains several examples of this sort.

The heuristics and biases tradition is foremost associated with the works of Daniel Kahneman, Amos Tversky, and Thomas Gilovich. Instead of success, it rather focuses on failures of intuitive decision-making. This tradition can be traced to the works of Paul Meehl (1954) and Lewis Goldberg (1979), who studied clinical psychologists' diagnostic judgments (see Kahneman & Klein, 2009). They found that diagnoses were often inconsistent, and simple statistical models proved more accurate. One of the first studies in HB tradition investigated experienced psychologists' and statisticians' intuitions about what would be the appropriate sample sizes in certain psychological experiments Tversky & Kahneman (1971). Participants reached incorrect conclusions and failed to apply statistical rules with which they were certainly familiar with. Research in this tradition has revealed that even formally trained and experienced professionals are susceptible to several such reasoning errors, which are often systematic and predictable between subjects.

Bulk of these reasoning problems appear in statistical and logical inference. This has fueled the hypothesis, generally held in the dual process theories, that formal reasoning is conducted in explicit System 2 and it is hard mainly because of the system's inherent processing limits. This makes people shortcut complex calculations by resorting to intuitive heuristics that may work in many situations but produce predictable biases in others. Both NDM and HB approaches embrace dual-processing framework, and beyond their apparent conflict, they mostly actually agree on the cognitive psychology of expertise. For example, both approaches accept the recognition-based model of skills and the related simulation account of practical reasoning. HB tradition mostly focuses on the simplifying heuristics that are associated with (formal) explicit reasoning and which need not rise from specific experience. Skill learning based on specific experience is the domain of NDM research. Hence, the approaches are funda-

mentally complementary rather than adversarial. Recognition-primed model implies that the environment needs to provide adequately predictive cues to inform action, and the agent needs to have opportunity to learn those cues. An immediate implication of this is that the causal and statistical structure of the environment needs to provide a sufficiently stable array of valid cues to support learning. If that is not the case, intuitive thinking can be defective regardless of the practice invested in skill learning. (Kahneman & Klein, 2009)

For example, expert political forecasts often go wrong because history, as a point of fact, does not repeat itself that much, and therefore the task of learning reliable cues to predict future historical events is simply impossible. Sometimes the task structure is stable and the cues salient but misleading. A famous story of this is told in Daniel Kahneman's autobiography for Nobel Institute.[108] He was lecturing to flight instructors that reinforcing accomplishments is more advantageous for skill learning than punishing poor performance. The audience fiercely protested. They had witnessed that many times yelling and screaming after failure resulted in better performance the next time but praising exceptional feats led only to disappointments. That is an entirely correct observation but because the instructors are witnessing the regression to the mean and not the results of their feedback. The instructors got the statistical structure right but causal structure wrong.

Besides the stability and validity of the environment, expertise needs actual practice to develop—and often lots of it. For example, for chess masters, the gathering of tens of thousands of useful exemplars in their knowledge base is estimated to take about 10,000 hours of dedicated practice stretching over a decade. Similar time frames are estimated for other domains. (Ericsson et al., 1993) People are notoriously poor intuitive statisticians, but our aptness for accurate probabilistic reasoning shows up in many casual everyday contexts (Nisbett et al., 1983; Griffiths & Tenenbaum, 2006). While our explicit statistical reasoning might be defective, we are able to extract a surprisingly accurate statistical structure of our stimulus environment even in laboratory (Reber, 1989). The ability to exploit that information—including base rates—does not seem to depend on declarative knowledge but on implicit procedural knowledge that requires actual interaction to be learned (Koehler 1996; Ahn et al. 2000b, 378), which is also a precondition for a sense of understanding of that information. Arthur Reber (1989) conducted a study on implicit sequence learning. He gave his subjects accurate frequency information about

[108] Available on the official web site of the Nobel price at `https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/kahneman-bio.html` (29.06.2018).

the sequence of appearing lights. The subjects reported that they understood and believed the information they were provided but still felt that this knowledge lacked any meaning they felt they could use. After a real experience with the task, they developed a skill set that directed their predictions about how the sequences proceed. They reported achieving a sense of the nature of the event sequences, which they did not get from the explicit instructions (p. 222). Allen and Brooks (1991) made a similar observation with learning complex non-statistical classification rules.

By and large, these observations side with the highly influential model of expertise developed by Hubert and Stuart Dreyfus (1986). They recognize five stages in a trail from a novice to an expert. In a nutshell, novices adhere rigidly to taught rules without a practical and contextual understanding of proper actions that they consider in isolation. For example, they attend to individual pieces in a chessboard instead of the global situation of the game. After some advancement, they begin to grasp that several situational variables need to be understood in wider contexts, but they fail to rank their relevance for present situations automatically. Beginners move to a third competent stage when they begin to understand the relevance of different aspects in terms of long-term goals. In this stage they have learned some routines that enable them to shift focus from the immediate situation to more strategic aspects of the task. Proficiency is reached when subjects begin to base their task understanding on the holistic appraisal of situations rather than on isolated attributes, and then they also start to notice deviations from normal patterns. At this, stage decision-making is still largely deliberate; however, deliberations are guided by situational factors. At the final stage of true expertise, decision-making is mostly automatized and intuitive. Experts cease to rely on rules and reflection and rather base their decisions on deep tacit understanding of the problem domain. Conscious deliberation occurs only when things are proceeding not as expected.

Note the resemblance of this model to dual-process theories. The top-down learning envisioned above can be interpreted as resulting from the indirect interaction of the two systems through behavior. Reflective System 2 initially drives behavior, and the intuitive System 1 slowly learns the proper responses to task stimuli (Wilson 2002, 121; Toates 2006, 84). If the problem domain is conceptually vague or characterized by complex statistical contingencies, intuitive learning at some point surpasses the accuracy that can be achieved by following explicit rules, resulting in efficient behavior despite the inability to verbalize or even acknowledge what one has learned (Seger, 1994; Sun et al.,

2005). Such implicit learning produces representations that reflect the agent-world interactions, and the statistical and causal structure of the task environment (Reber, 1989). If all goes well, deliberate control becomes advantageous only when a resolution for stimulus can not be retrieved from the long-term memory or existing solutions do not work—i.e., outcomes of actions under stimulus are not as expected. Then attention mechanisms are triggered, and controlled behavior is initiated (Toates, 2006) as per the default-interventionist dual process model of reasoning (Evans & Stanovich, 2013).

But how is procedural knowledge represented in the intuitive mind? Skills and automaticity are often taken to build on situation-action-outcome (SAO) instances (e.g. Logan, 1988; Palmeri, 1997; Sakai, 2008; Koechlin, 2014). When a stimulus is encountered and action taken, these two alongside with the resulting outcome are encoded in long term memory. When the same (or similar) stimuli are encountered later, associated SAO exemplars are automatically retrieved. Encoding and retrieval processes are mandatory and content specific. When more similar exemplars accumulate in memory, the access time speeds up, leading to rapid triggering of action before explicit thinking has time to commence. According to these instance theories, the automatization of skills is, at least for the most part, a memory phenomenon associated with specific contents which does not change the underlying general cognitive capacities. These observations are comparable to what has been reported in exemplar research of categories (Allen & Brooks, 1991).

While the theory has a slight flavor of behaviorism, it is not about simple stimulus–response associations. Internal goals enter into encoding and processing of SAO exemplars, and the idea is that the resultant memory storage is indexed by stimuli and internal goals (Larkin et al., 1980). Since (pragmatic) contexts can be defined by events interpreted in terms of goals, exemplars encode actions interpreted on the contextual basis. This is because overtly the same action can mean different things in different contexts (Vallacher & Wegner, 1987; Christensen et al., 2016). While encoding is mandatory and can happen without awareness, the agent is not a passive receiver but active attentional processes constrain what gets encoded during learning (Logan, 1988; Reber, 1989). Since disturbances are assumed to trigger attention, challenging events receive more attentive processing, and thus we tend to remember and learn more from hard than easy tasks (Christensen et al., 2016, 60). A similar observation is made in category processing where unexpected instances draw attention and have elevated influence in subsequent judgment (Heit, 1998). Lastly, automatic processes are controlled even if they are not

explicitly reflected. This is evident by the fact that disruptions interrupt the flow of action. Christensen et al. (2016) have criticized the Dreyfus model for overemphasizing automation. Skillful action does not rely on the ballistic execution of preprogrammed routines but we actively seek information to maintain situation awareness and guide our behavior online. As noted earlier (see page 184), this is true even when all the relevant information is available: We often shift our focus to retrieve information in piecemeal fashion during performance rather than load everything in working memory in one go.

In brief, fundamental cognitive representations in procedural knowledge are *SAO* exemplars, which serve as inputs to cognitive control. Both representation and processing are composed of endogenous goals and action control with exogenous stimuli and affordances. The learner extracts a stock of rule-like representations that associate specific actions to desired outcomes in a given context.[109] Note that this is an essentially identical theoretical construct with the one postulated by Medin and Schaffer (1978) for exemplar category processing. As noted earlier (e.g., on page 130), these recurring similarities between procedural knowledge research and exemplar theories of categorization are almost certainly not coincidental but a result from the fact that exemplar research is actually probing skill learning. In other words, category learning tasks that demonstrate exemplar-based learning are, in fact, investigating the process of associating proper actions with stimuli, where a successful outcome is defined by making the correct classification response in the context of the experiment.

Much of implicit learning research goes beyond addressing how an action is associated with static stimuli, and focuses on sequence learning, for example. Often such sequences are implemented by a Markov process, which produces branching strings where the next event is determined by the current one by a stochastic function. Learning such temporal structures is essential if organisms are to anticipate how complex events unfold. Related research focuses on sequences that are produced by hierarchical structures generated by recur-

[109] Often rules are thought as language-like symbolic constructs and contrasted with associations (e.g. (Sloman, 1996)). Gigerenzer and Regier (1996) have pointed out that there is no empirically or conceptually principled way to tell simple *if–then* rules apart from one-way associations. I find this seemingly trivial remark as hugely important. It implies that conceiving *SAO* associations as rules does not necessarily commit us to symbolic computationalism. Moreover, it means that if (presumably) associative implicit cognition can learn what (presumably) rule-governed explicit thinking does, *it can still learn the very same things* regardless of the assumed qualitative differences of implicit and explicit representation and processing.

sive rules. These are called artificial grammar tasks because the stimulus sets used are often symbolic formal languages (in the well-defined mathematical sense) and because of the widely held idea in linguistics that natural language competence is a product of such recursive processes. It has been a furiously debated issue for decades whether or not this implies that linguistic processing necessitates a hierarchical execution of formal rules that manipulate symbolic strings in working memory. I would rather not touch that specific issue at all. At this point, it is enough to remember that linguistic structures can be parsed by neural networks that lack such architecture (Socher et al., 2013), and complex hierarchical plans can be learned through reinforcement that maps situations onto policies of actions instead of mapping them onto singular actions only (Russell & Norvig 2010, 856; see also Uddén et al. 2009). Therefore, assuming that representation of actions or policies of actions does not require symbolic rules, it is not necessary to jump to conclusions what implicit hierarchical structure learning might imply about the representation and processing characteristic of the intuitive mind.

As discussed in Section 2.2.2 neural resources behind such hierarchical sequence learning are not associated with language only but with a more general capacity to understand and maintain hierarchically structured sequences and action. To recap from our earlier discussion, it seems that it is a fundamental capacity of the human cognitive system to chunk not only perceptual input but also simple actions to more complex action programs and recursively these programs to more complex units. The result is a hierarchically organized behavior where the highest level is the gist or long-term oriented understanding of the behavior, which is multiply realizable by integrating simpler and more specific actions, which organize the behavior around more immediate goals and affordances. Now, all this is far easier said than done. The perennial problem in cognitive science is to understand how the action selection and control is actually achieved in complex open environments. Curiously, research on elementary learning and decision-making have been advancing quite separately even though related neuroscientific approaches study the functions of the same brain regions (Padoa-Schioppa & Schoenbaum, 2015). Recently, efforts to change this sorry state of affairs have surfaced (Domenech & Koechlin, 2015; Duverne & Koechlin, 2017), and the analysis happens to be relevant to my purposes.

The basic theoretical apparatus behind this endeavor is the combination of task sets and reinforcement learning. The notion of *task set* (or "mental set") dates back to 19th-century German empirical psychology (Monsell, 2003,

135). The basic idea is that the same stimulus can evoke any number of meaningful reactions, depending on the contextual setting where the stimulus is encountered. The mindset or task set is the psychological readiness to select a contextually appropriate course of action. It is trivial to arrange two tasks such that the stimuli are identical in both, but each require different stimulus-response mappings. For example, one can ask subjects to confirm whether a digit is even versus asking them to judge whether it is larger than five. It is equally easy to arrange two tasks such that all stimulus-response pairs are actually identical in both by ensuring that the even digits in the stimulus sets are smaller and odd digits larger that five. Still, even in this identical S-R condition when subjects need to switch from one task to another, a *switch cost* is observed which is a transient decrement of speed and accuracy in the subsequent task even if it is well-learned. This reflects the processing required to adapt to a new context, even though the observable S-R associations remain constant. Much research has been dedicated to investigate how such task sets are established, maintained, and controlled (Meiran et al., 2000; Monsell, 2003; Sakai, 2008).

Philippe Domenech and Etienne Koechlin (2015) have proposed a model, based on computational and empirical neuroscience, of how an adaptive task set construction and control happens in open environments. Simple decisions are learned initially as selective stimulus-action pairs through model-free reinforcement learning (RL). RL is an unsupervised learning technique where rewarding actions get reinforced, producing habitual mappings of $S \rightarrow A$ associations that maximize reward value associated with actions. In RL, the agent does not need to have specific predetermined goals but it can autonomously discover what it finds satisfying through reinforcement. The basic idea is to learn what to do through interaction, which is conceptually quite similar to classical conditioning via trial and error. The premotor cortex encodes actions associated to perceptual cues, and basal ganglia—a subcortical structure present in all vertebrates—maintains ongoing behavioral strategy and adjusts it according to the expected and realized reward values. Such *model-free* RL is roughly the computational equivalent of the influential reward prediction error theory of the dopaminergic system, which models behavioral adjustment through rewards and punishments (Doll et al., 2012).[110]

[110] N.b.: The notion of "reward" is not equivalent to utility or hedonistic pleasure, even though if all goes well rewarded actions are advantageous to the agent. "Reward" in this context is a theoretical notion associated with motivation and learning and refers to aspects of stimuli, actions, or outcomes that reinforce behavior and serve as incentives in decision-

If the environment is stable, cues are encountered sufficiently often, and rewards depend strictly on current stimulus and selected actions, RL tends to converge to behavior strategy that maximizes reward values. However, by design, pure RL overwrites existing associations if the external contingencies or internal rewards change (e.g., due to satiety). This is fine if they change gradually and permanently but maladaptive if the agent needs to cope with abruptly changing and periodically recurring situations. Moreover, simple model-free RL is unsuited to organize sequential behavior where the intermediate steps are not reinforced and distant rewards must be inferred from the current and expected future states (Doll et al., 2012; Koechlin, 2014). According to Domenech and Koechlin (2015), ventromedial prefrontal cortex (PFC) encodes action-outcome associations separately from fixed $S \rightarrow A$ associations for selecting actions according to *outcome* reward values. These $A \rightarrow O$ pairs are learned statistically rather than by reinforcement, and they are predictive instead of selective with respect to actions. In combination, $S \rightarrow A$ and $A \rightarrow O$ pairs make it possible to map stimulus to expected outcomes of actions and dissociate specific stimuli from specific action-outcome pairs. Utilizing (stimulus)-action-outcome exemplars through statistical learning makes it possible to establish *model-based* RL where the SAO knowledge base tracks the functional structure of the task environment. Moreover, dissociating action-outcome associations from stimulus enables to surpass the inherent limitations of model-free RL through task set exploitation.

SAO exemplars enable defining the statistical distributions of outcomes (associated with specific actions in specific situations) in a non-parametric way. In non-parametric models, it is not necessary to prespecify the range of possible outcomes and track posterior distributions every $A \rightarrow O$ contingency in a given situation. Instead, one can cumulate something like tabular or histogram-type representations that contain an accurate description of observed outcomes and their relative proportions. For example, if one gathers $80 \times (S_1 \rightarrow A_1 \rightarrow O_1)$ and $20 \times (S_1 \rightarrow A_1 \rightarrow O_2)$ exemplars, action A_1 seems to lead to outcome O_1 80% of the time in situation S_1. Trying other options (i.e., $A_2, A_3, \ldots$) will generate similar information about their outcomes $O_1, O_2, O_3, \ldots$. At any given time, the agent can discover new actions and outcomes and hence learn completely new contingencies in a piecemeal fashion. Successful (rewarded)

making. For example, in case of addiction, rewards make us do harmful things that are not necessarily enjoyable or gratifying, although dissociation of reward, pleasure, and perceived instrumental utility is not pathological as such. The notions are closely related but distinct both conceptually, psychologically, and neurologically (see Berridge & O'Doherty, 2014).

outcomes are encoded in the action selection ($S \rightarrow A$) model of the task while SAO clusters form the predictive model. Future rewards can be predicted based directly on previous experience or by prospective simulation (Gilbert & Wilson, 2007). Here, I concentrate mostly on the outcome prediction. From a predictive point of view, outcomes are subsequent situations brought about by interventions or manipulations of the current situation. Thus, outcome O becomes situation S_2, and similarly, the agent can produce information about the outcomes of specific actions taken there. In this way, we acquire *action sets* that allow goal-directed outcome prediction and sequential action selection in an environment where contingencies remain relatively stable (Collins & Frank, 2013; Domenech & Koechlin, 2015).

The hard part is to explain how this knowledge is controlled and maintained in open environments. The key is to use *task sets* that are clusters of action sets, which are activated in specific contexts. As explained above, several action sets are associated with a specific task and stimuli. They are adjusted by reinforcement; however, to avoid useful knowledge to be erased, $(S)AO$ traces are permanently stored and only those task-specific action sets that are currently selected to drive behavior are subjected to change through RL. This raises a fundamental problem in reinforcement learning, however: since external conditions can always change, the agent needs to arbitrate when to exploit and adjust previously learned knowledge and when to learn utterly new task sets to maximize long term rewards. This *exploration/exploitation*-dilemma is a computationally intractable problem because, in complex environments, the number of learned behavioral strategies can grow very large, and an optimal solution requires re-evaluating past arbitrations and adjustments and at the same time monitoring the whole repertoire of learned behavioral options online (Koechlin, 2014). According to Domenech and Koechlin (2015), PFC has been evolved to solve this problem.

Basically, an *actor* task set (the active strategy driving behavior) is activated as afforded by the stimuli, that is, the surface content of a situation activates clusters of specific memories that are associated with the current goal. Core executive system in the medial portion of PFC computes the absolute reliability of the actor by using forward inference from stimulus to expected outcomes. If action outcomes match expectations, external contingencies are presumed to remain unchanged, and the actor is maintained. If not, the actor is discarded, and a new one is created from long-term memory.

A distinct inference track in the lateral PFC is assumed to be responsible for new task set creation, and the development of these regions in primates al-

lows for proactive instead of simple reactive inferences, that is, evaluating actor reliability before action (Koechlin, 2014). Thus, action sets are not stimulus-reaction associations. They appear in *between* stimulus and action and serve as inputs to control. Then the agent needs to monitor contextual cues and other external evidence to assess actor reliability before feedback. This happens through forward inference from cues to outcomes perhaps via simulation. It is important to note that task set selection is not based on its expected utility but its reliability (Collins & Koechlin, 2012). Every time a task set is created, it is recoded as a separate trace—at least if found reliable—and the more the similar sets accumulate in memory, the more they contribute to task interpretation and action selection. Successful actions get promoted through RL and more frequently selected actions cluster into long-term memory further promoting their selection. While the underlying selection is fundamentally stochastic, this implements a winner-takes-all mechanism that restricts alternative (hopefully futile) actions from being initiated (Domenech & Koechlin, 2015). This process is reminiscent of the Logan's idea of how accumulating exemplars lead to more rapid but less versatile decision-making and hence automation; only the underlying assumptions of the model are more detailed and complex.

The selection of proper actor task set is a challenge in the first place. Outside the laboratory, we are rarely told explicitly what to do and we need to figure out the pragmatic context by ourselves. Anne Collins and Michael Frank (2013; also see Collins et al. 2014) made an interesting discovery that in novel tasks subjects seem to spontaneously encode parts of the stimuli as contextual cues, and some stimulus features as direct cues for action selection. This happened even in simple tasks without any apparent benefit and without any indication as which input dimension should indicate broader context and which drive immediate action selection. In one experiment, for example, the stimuli consisted of objects with two shapes and two colors, and in the learning task one of four different actions were supposed to be associated with each pair. Thus, one could easily learn simple disjunctive rules: *red squares*$\rightarrow A_1$, *red triangles*$\rightarrow A_2$, *yellow squares*$\rightarrow A_3$, *yellow triangles*$\rightarrow A_4$. Still, about half of the subjects displayed task switch cost on color change and a half on shape change. They seemed to learn rules of the following *hierarchical* kind: (a) if the color is *red*, then *square*$\rightarrow A_1$ and *triangle*$\rightarrow A_2$; (b) if the color is *yellow*, then *square*$\rightarrow A_3$ and *triangle*$\rightarrow A_4$—thus treating color as context for selecting specific $S \rightarrow A$ rules (and *mutatis mutandis* for the group that treated shape as a context).

In short, similar observations revealed that we spontaneously encode stimuli such that some stimulus dimensions S_c are used to retrieve action sets ($S \rightarrow A$ associations, presumably conditioned on goals), and after this, other stimuli S_a activate specific $A \rightarrow O$ associations. In other words, task sets are first triggered by higher-order contextual information from the environment and selected in response to context, and then specific actions are selected in response to specific contextualized stimulus content. Developmental psychologist Renée Baillargeon (2002) has also proposed along similar lines that infants first categorize an event and then use knowledge associated to the event to guide what variables should be attended to. Again, similar proposal was made in Medin and Schaffer's (1978) seminal paper on exemplar based categories.

But, what would be the point of this? Presumably nothing in the above discussed Collins and Frank's austere artificial task. However, in more complex ecologically valid environment sit makes multidimensional decisions easier by breaking it down to two phases: first, interpret the higher-order pragmatic context; and then select relevant actions afforded by specific stimuli (Collins & Koechlin, 2012). The same mechanism may explain the spontaneous creation of *ad hoc* and goal derived categories by constructing object identity and category structure through the preceding selection of the pragmatic context. Such arrangement is also necessary for maintaining hierarchically structured behavior where the long-term goal is multiply realizable, and the specific actions need to be associated with specific affordances, interpreted in terms of the functional context. Lastly, different contexts may utilize similar task sets, and dissociating specific actions (at least partly) from contextual cues allow for analogical transfer. The problem then is to find a proper analogy. The assumption is that contexts are recognized by surface cues (i.e., specific content S) and hence retrieving a valid existing task set (i.e., $A \rightarrow O$ rules to be associated with the novel stimuli to serve as analogy) is not necessarily easy and requires experience to pinpoint the relevant structural cues.

To sum up, reinforcement learning makes us to select actions based on expected rewards under a task set, and task set enables selecting actions based on cues. Notably, this enables the agent to discover its own means and ends. The task sets are selected according to current goals as afforded by stimuli, and maintained or discarded based on their reliability. The realiability and reward is computed by medial portions of PFC. Inference track in the lateral PFC computes the reliability of alternative task sets and creates new ones in the case the active set needs to be switched. Neuronal coupling of these two

systems integrates expected rewards and learned rules in strategy selection (Duverne & Koechlin, 2017).

The practical reality of the agent is conceptually organized according to pragmatic contexts which are defined by goals and concrete affordances. These concext often overlap and they come in different levels of abstraction. Indeed, virtually every situation consist of nested and intersecting contexts, which are delineated by immediate and distant goals. This is reflected in the hierarchical structure in the lateral PFC where the most frontal parts arbitrate between different goals and behavioral strategies the situation affords. In essence, these areas are figuring out what should be done and manage competing goals that may unfold in different time-frames. The neighboring caudal parts process what could be done and manage behavioral strategies as dictated by goals and afforded by the stimuli (Mansouri et al., 2017; Pezzulo et al., 2018), and this inference track extends to premotor areas, which is involved in action preparation and encodes the acute action rules.

Since the idea here is that task sets are essentially the interpretation of context, I presume that analogical transfer would often require an understanding of how the procedural part of situation representations connects remote analogies. Experience is also needed to tell misleadingly similar tasks apart. Recall from the introduction the Brian Ross' (1984) study in which the subjects misapplied a solution procedure that was learned in one task to another task sharing similar surface content. Presumably, the reason was that the subjects interpreted the context correctly: it was about probability calculations. Hence, they were able to retrieve a specific task set associated with such calculations. However, they lacked alternative procedural knowledge attached to those specific contents (e.g., dice rolls), and, for being novice, they did not have relevant experience that could aid them to interpret the task differently (other than in terms of task set attached to probability calculations about dice rolls) and exploit the problem's formal structure in action selection.[111] Hence, they presumably lapsed into exploiting the existing but irrelevant action set in vain rather than engaging in an effortful (and possibly just as futile) explicit search for solution.

Revisiting naturalistic decision-making and the kitchen fire example (page 196), we can hypothesize that the fire's observed lack of response to the extinguishment attempts afforded the first reactive inference that the current behavioral strategy was unreliable. The other cues (e.g., the unusual quietness)

[111] The psychology of surface versus structural content in task interpretation is discussed in detail in the next chapter.

may have worked as less salient prospective cues, signaling that the situation at hand was actually something not expected. Since the fire lieutenant had not experienced such a situation before in any seemingly identical house, no previously stored task set could be activated to interpret the mixture of such cues to inform him that it was a basement fire rather than a kitchen fire. Novel task set creation is supposed to be an effortful off-line process, and hence in such risky and time-critical settings, the best and only option was to pull out and rethink what to do.

Note that the scheme of cognitive control in the above-discussed theory by Koechlin, Domenech, Collins, Frank, et al. is, again, highly similar to the default-interventionist dual-process theory. If the task is considered as familiar, agents execute a routinized course of action. If things fail to go as expected or no interpretation for the situation is found, explicit search for an alternative solution ensues. These writers are unfortunately not very explicit on how the alternative task sets are created, apart from suggesting that they are retrieved from long term-memory by recombining existing actions and action sets. Analogical reasoning clearly fits this bill, especially in case of ill-understood problems. Another evident option is to use domain knowledge when it is available. If the agent has substantial procedural knowledge associated with a context, it is in good position to search for familiar patterns and solutions. This, of course, requires expertise. For educated adults, it is sometimes difficult to remember that seemingly trivial skills such as elementary arithmetic need to be learned at some point in life, and subjects without such skill sets ingrained in long-term memory may be entirely unable to solve or even understand tasks like simple probability calculations. The point of these remarks is that if the putative System 2 is engaged in reflective thinking in case intuitive solution is not available, quite likely this process is mainly devoted to searching analogies and procedural information from the-long term memory rather than executing decontextualized, formal, and general reasoning processes. If so, System 2 contents depend on the idiosyncratic stock of experiences, possessed by the person engaged in reflective thinking.

A critical part of the exploratory behavior in the above-discussed model involves compiling and testing the reliability of novel task sets. However, if absolutely no interpretation for the task can be retrieved, the agent lapses into a model-free exploratory behavior as the last resort. One might assume that this sort of activity is best described as ”random” (e.g. Collins & Koechlin, 2012). However, I presume that this is not the case. Given that human cognition is adapted to extract the causal structure of the environment, one should

expect that the free exploratory behavior displays patterns of optimal data selection in testing causal hypotheses. This seems to be the case in general (Penn & Povinelli, 2007, 104), and it is discussed further in the next chapter. This exploration behavior is generally missing from dual process-theories that emphasize intuitive heuristics to cope with complex and unfamiliar problems. Moreover, one crucial source of behavior strategies seems to be missing in most of these theories: learning from others, which we discussed in Section 2.2.2.

Lastly, it should be noted that talking about input stimuli may mislead thinking in terms of impoverished sense data. However, sensory input is almost always interpreted as a rich causal representation of the surrounding situation. This mainly pertains to familiar everyday situations but less so to artificial stimuli and tasks tested in the laboratory. When dealing with familiar objects and contexts, situations do not need to be familiar in detail to afford intuitive appraisal of the pragmatically relevant causal structure. As discussed in Section 4.2.1, domain knowledge automatically guides our attention toward causally relevant variables and affords the generation of causal hypotheses or exploratory task sets. The predictive models constructed from task sets are fundamentally similar to the causal models discussed earlier. The main difference is that causal model approach emphasize allocentric representation where the effects of interventions are derived from the structure of the model, while predictive models based on task sets primarily contain egocentric information about interventions (i.e., the outcomes of one's actions). Presumably, in the latter structures the knowledge about agent independent event–event causation can be contained as a special case of abstaining from overt action.

5 Abstract thinking with concrete intuitions

Finally, we have covered the basics of concrete category, situation, causal, and procedural knowledge. What we have left on the agenda is the following part of the working hypothesis:

c) Basic reasoning mechanisms are forward causal inference, simulation of situated action, and abstraction by exemplar-based analogical transfer:
   - Surface features are the primary retrieval cues.
   - Valid analogies bind different tasks under shared pragmatic schemata.

Complex skills and commonsense reasoning are both mostly achievements of intuitive cognition, and along the way I have claimed that this also pertains to reasoning with schematic and theoretical concepts—even though intuitive reasoning rests on specific empirical knowledge, and by definition schematic reasoning is non-specific and theoretical concepts are non-empirical. Earlier, I announced that the hard part of theories that emphasize situated concrete reasoning is explaining abstract thinking. However, at this point the hard work has been mostly done already. In the following sections, I first explain how schematic contents can be engendered by situation representations and procedural knowledge in a bottom-up manner and then how discursive reasoning with theoretical and formal concepts can be learned as an inferential capacity than employs similar implicit learning in cultural contexts. While the former part is mostly a recapitulation of our previous discussion, the latter is more hypothetical and focuses on the following planks of the working hypothesis:

A.2. Conceptual understanding builds up as an adaptive cognitive skill. Its cognitive basis is in the intuitive system that gradually learns to exploit context- and goal-relevant regularities in the environment, especially the effects of our own actions in specific situations.

A.3. Often relevant environments are, at least partly, socially constructed. In the case of theoretical concepts, in particular, the relevant regularities are in large part inferential and other discursive commitments. Hence, abstract concept learning is a special case of functional/causal learning in (broadly) social or cultural contexts.

A.4. Initial competencies depend on concrete examples and their surface features or specific content. Extensive learning of procedural knowledge results in a gradual shift of focus from surface cues to structural features of the conceptual domain.

A.5. Through experience, the abstract/practical distinction dissipates both psychologically and phenomenologically.

C.4. Cognitive contents of theoretical and formal concepts are constructed by the agent through social interaction. While grounded in the same capacities, discursive and concrete concept learning often differ qualitatively: Discursive conceptual domains track communal conventions; this makes content intersubjective and normative.

5.1 A case study on conditional reasoning

Thus far we have learned the following: Human intuitive cognition relentlessly tracks the structure of the environment producing causal maps about how entities and their features interact in specific situations. At the same time, we track the effects of our actions, producing predictive models that drive the use of procedural knowledge. Now, let us postpone the obvious question about how these knowledge structures are related. (I will get back to that in the next chapter.) However, both kinds of knowledge are tied to specific contents and situations, and they are activated by perceptual (or simulated) input that contains the stimuli which with they are associated. The causal maps are organized around distinct entities that allow for recombining novel arrangement of familiar things to derive causal predictions in previously unencountered events. Unfortunately, causal maps can quickly become very complex making such information practically useless. Moreover, these causal representations portray mostly the qualitative structure of events, and the effective use of that knowledge (e.g., deriving accurate statistical estimations) generally necessitates practical experience in recurring situations. Fortunately, during learning, humans can extract very fine grained cues to command their actions, and repetition eventually results in a capacity to make complex but accurate judgments in the blink of an eye. Moreover, we spontaneously track contextual cues in the environment that trigger (in combination with exogenous factors such as needs and goals) specific sets of procedural knowledge, which guides what information needs attention. This arrangement makes complex decisions easier since the context, cue, and affordance recognition, which support action, occurs in two phases. All this encoding and retrieving are automatic and mandatory. They often happen without awareness, although selective attention affects both learning and retrieval.

Apart from searching for contextual cues, we are inclined to identify our behavior in the highest level of identity that can be characterized by a coher-

ent, pragmatic context or goal. When specific actions are routinized, control shifts to higher strategic aspects of the task, allowing cognitive resources to be allocated for the pursuit of long-term goals. This necessitates that the more detailed action routines are first mastered intuitively, and if problems arise, focus will again lapse back to lower, more specific aspects of performance. Also, the understanding of generic events, whose defining characteristics are removed from specific objects and acts, requires understanding of lower-level instances for translating schematic descriptions into concrete world knowledge. In any case, the capacity to conceive behavior in the higher level of organization allows integration of simple actions into complex, hierarchically structured tasks. This makes behavior more flexible and promotes new courses of action, because the intention of such higher-level behaviors tend to be multiply realizable. Moreover, all this enables skilled persons to pay attention to more structural elements of familiar situations instead of the detailed implementation of how to cope with them. Unfortunately, the perceptually driven memory mechanisms that activate relevant pragmatic knowledge seem to be triggered by specific contents, even though different contexts may harbor the same functional structure under the seemingly different surface. Hence, the exact routines sometimes need to be learned over and over again in different contexts.

What we have not discussed is the cognitive etiology of rule-based and logical reasoning. Below, we will see how all this plays out in inferential reasoning with rules, and how domain knowledge and schematic concepts build up from similar skill learning and hence inherit its characteristics. For expository and narrative purposes, I tell this story by way of exploring how our previous discussion applies to conceptualize behavior in the Wason selection task—likely the most employed paradigm in the study of human reasoning. The original version (Wason, 1966), often referred to as *the abstract task*, used a deck of cards with letters printed on one side and numbers on the other side of each card. The subjects were told that *if a card has a vowel on one side, then it has an even number on the other side*. Then, for example, the cards A, D, 4, and 7 are displayed and the subjects are told to pick those (an only those) cards which have to be turned over to check whether the claim is true or false.

The logic is straightforward: formally the claim is a universally quantified conditional *for all x: if x has P, then x has Q*; or simply *every P is (or has) Q*. This is arguably the basic logical form of generalizations in natural language. The task can also be framed by a reference to a specific letter and a number, for example: *if there is A on the one side of the card, there is* 7 *on the other*

side. This leaves the logic essentially intact, and does not affect the responses given by the subjects. In what follows, I will use the notation P, $\neg P$, Q, $\neg Q$ to refer to the cards in the selection task, given the rule of the form *if* P *then* Q (e.g., in the original task $P = A$, $\neg P = D$, $Q = 4$, $\neg Q = 7$).

The task is thought (by the experimenters, but perhaps not by the subjects) to be about a test of the validity of a material implication *if* P *then* Q—the most elementary conditional statement. Practically, everyone is perfectly able to reason with such a logical form: if it is the case that A, and also that *if* A *then* B, then it follows that B. Logic does not go much simpler than this, and many conditional rules that we are all familiar with follow this logic. However, surprisingly, most people have problems when trying to comply with the logic in a less direct manner and when confronted with rules they are not familiar with. This is why the task has attracted considerable attention. Despite its simple logical structure, only about 10% of the participants choose the correct cards. Consistently across replications, about half of the subjects choose P & Q, and about one third choose card P only. For testing the rule, card P obviously needs to be turned over. Card $\neg P$ is irrelevant, because the rule only pertains to cards having P, so it is immaterial what there is on the other side of $\neg P$. The problem is with the selection of Q and $\neg Q$. Suppose you turn over the card that has Q and you find P on the other side. The *if* P *then* Q rule is certainly followed then. However, suppose there is $\neg P$ on the other side. Now, we just agreed that cards with $\neg P$ are irrelevant. Therefore, Q card is irrelevant because for the validity of the rule it does not matter what there is on the other side. However, if you turn $\neg Q$ over and find P on the other side, the rule is obviously violated. Hence $\neg Q$ is relevant, and the correct solution is to select cards P and $\neg Q$.

The finding is sometimes taken to demonstrate the inherent fallibility of human reasoning because whatever the subjects are doing it does not look like logical inference. At first sight, the problem might be attributed to the trivial misinterpretation of the rule. We often use conditional *if . . . then* structure to indicate *bi*conditional P *if and only if* Q. However, then all the cards should be turned over, and this rarely happens in the experiments. When asked to justify their choices or (incorrect) choices made by others, subjects mostly produce irrelevant and inconsistent explanations, further indicating that they are not at least *consciously* reasoning according to biconditional or any other Boolean reading of the rule (Wason & Evans, 1975; Evans & Wason, 1976). In principle, they may be *unconsciously* reasoning according to *if* P *then NOT* Q, but that would be a very odd interpretation, indeed. When the subjects are instructed

to turn all the cards and tell which of them would invalidate the rule, usually all give the correct answer (Wason & Shapiro, 1971). Thus, it seems that the subjects understand the relevant logic but are unable to employ it in the selection task.

The reason I discuss this experiment is because for about a half century it has been a sort of a fruit fly of the psychology of reasoning. Virtually every conceivably relevant variable in the experimental setting has been tested by using thematic versus abstract content and familiar versus unfamiliar rules, employing context or a story to frame the task and assessing how cognitive performance measures (e.g., SAT scores) and formal education affects the performance. This has produced a wealth of data about the relations of factors in human reasoning relevant to my argument, the most crucial being expertise, experience with specific content, schematic reasoning in familiar domains, and the role of context. This allows me to point out how these factors hang together in elementary reasoning with a logical rule. Whether classical logic provides the right normative model to evaluate the performance in Wason's task is a contentious issue; however, subsequent research has revealed that specific manipulations of the content and context of the rule allows the majority to conform to the logical model. Whether this means that subjects in these instances are exploiting logic or merely conforming to logic is an interesting issue.

Another reason to discuss this particular experimental paradigm is its important role in theoretical debates of human rationality and reasoning. As such, it seems to support the pessimistic image of human cognitive competence, advertised by the heuristics and biases tradition. It has also been crucial for the development of dual-process theories of reasoning. To my best knowledge the modern version of the theory was first proposed by Wason and Evans (1975) as an explanation of the behavior in the selection task. The interesting theoretical challenge is that even though subjects are not reasoning in accordance with logic, they are not behaving randomly, either. The selection patterns are unevenly distributed and highly predictable between replications. Since subjects are often to unable to provide any sound reasons for their selections, this pattern is thought to emerge from the operations of the intuitive mind.

Some authors have complained that the selection task has received too much attention, and the test itself does not tell us much about human reasoning. For example Dan Sperber et al. (1995; 2002) have argued that pragmatic comprehension mechanisms pre-empt the use of any reasoning that might be needed in the task. The patterns of selection simply follow from the subject's

(in)ability to comprehend what is asked of them, and the task actually taps contextual relevance assessment rather than inferential processes in problem-solving. Evans has made the same point that it is indeed a *selection* task and not a reasoning task. In one of his experiments (Evans, 1996) he used a computer screen to display the cards and instructed the subjects to hover a mouse pointer over each of the cards as they were considering whether it was necessary to turn the card over or not. When a selection was made, the subjects clicked on a card, and the total time spent on that card was recorded. The recorded time for each card strongly correlated with the probability of its selection. Thus, reflection seems not to have much effect on selection decisions, but instead, the subjects spent most of their time thinking about the cards they were going to pick anyway. I think this is what basically makes the task very interesting because relevance is the most important ill-understood issue in human cognition, and such intuitive pre-selection of information is generally thought to be the essence of relevance assessment.

I use the selection task as an example of how the causal and pragmatic reasoning framework explains information selection that guides discursive reasoning. Note that since task set selection is an interpretation of context and production of the predictive causal model is the interpretation of the task content, the difference between inference and interpretation becomes somewhat moot. At least trivial inferences about the relevant variables are already contained in the interpretation, as per the above point made by Sperber and his colleagues. More specifically, I exploit the selection task as a concrete example case to expose how the main theoretical threads of my empirical hypothesis hang together, in particular, content specific causal reasoning, exemplar based inference and memory access, and the reliance of cognitive expertise on specific content and content. The discussion also introduces the idea of how schematic abstractions are engendered bottom-up by analogical transfer—a theme which will be further elaborated in the next section.

5.1.1 Behavior in abstract and unfamiliar selection tasks

In the abstract task, the subjects do not have prior experience with the test rule, and hence they cannot retrieve any meaningful information from long-term memory. Then the task turns into interpreting the *if . . . then* construction and what to do with it in the task context. Unfortunately, conditionals have no standard decontextualized interpretation, or at least it is not the one used in formal logic. Anyone familiar with teaching an introductory level course in logic

knows that the definition of material implication is unnatural, as evidenced by the so-called paradoxes of logical implication. For example, any truth is implied by any proposition, and this is because in logic *if A then B* simply means that $\neg A$ *or B* (*or both*) where A and B do not need to bear any relation. Such truth-functional interpretation of the implication is very remote from its standard semantics in natural language, which is to express some factual (and often causal) dependency between A and B.

Consider the following claim: *In every pub, there is a customer such that if the customer does not drink, then no one drinks in the pub.* Do you think this is true? My guess is that most people find it obviously false. Regardless, it happens to be a logically valid statement. To see this, let's translate it into predicate calculus: In every model M (i.e., *pub*) there is an entity x (a customer) that satisfies the following sentence: $\neg P(x) \rightarrow \forall x \neg P(x)$ (where $P(x)$="x drinks"). Hence, the claim says that in every model $M \vDash \exists x[\neg P(x) \rightarrow \forall x \neg P(x)]$. The formula is equivalent to $\exists x[P(x) \vee \forall x \neg P(x)]$, which is equivalent to $\exists x P(x) \vee \exists x \forall x \neg P(x)$. Since in the right-hand side, x is already universally quantified, the existential quantifier is redundant, and therefore we get: $\exists x P(x) \vee \forall x \neg P(x)$, which is certainly valid because it is equivalent to a trivial tautology: $\exists x P(x) \vee \neg \exists x P(x)$. In plain English: "Either at least someone drinks in the pub or no one drinks in the pub."

So why our intuition resists such an "obvious" logical fact? It is hardly the complexity of the calculations involved. People do understand much more complex claims, and if the complexity was an issue, one probably should expect people to be mostly undecided and sometimes *approve* the sentence rather than reject it outright. The likely explanation is that the "*if ... then*" clause is interpreted as causal, that is, as a claim that in every pub there is someone who would force everyone to abstain from drinking if he has decided to have a sober evening. That is an entirely natural interpretation of the conditional, albeit not logical in the formal sense. As per that interpretation, the claim is definitely false.

Since conditionals are generally used to express such material contingencies, subjects do not have a ready interpretation for the conditional statement if they cannot see what dependency it tries to capture; even while they are perfectly able to process formally identical rules. Using thematic material and context that highlight the relevant aspects of the task remove such problems (Sperber et al., 1995; Sperber & Girotto, 2002). The abstract selection task fails to elicit a meaningful *learned* response, because there is no natural sense in the rule. According to the generic dual-process hypothesis, when this happens, the

control is passed to the explicit system. Unfortunately, the logical solution still resists. Even previous acquaintance with formal logic does not necessarily help much, and neither does scientific education in general; this further underscores that this is a semantic issue. People need to find the proper interpretation of the task (i.e., as a test of the logical implication) to understand what to do with it even if they have the required competence (Cheng et al., 1986; Jackson & Griggs, 1988); and this is not a trivial feat. With proper instructions, it is possible to achieve good performance with subjects not acquainted with formal logic; however, when this facilitation is effective, the experimental instructions basically consists of explaining how to solve the task and therefore the experiment reduces to observing if people are able to follow simple instructions (Platt & Griggs, 1993). In case no instructions or pragmatic understanding is available, subjects presumably resort to fall-back heuristics; for example, they might merely focus on the cards that are mentioned in the rule Evans (1984).

Based on our previous discussion, we should expect that if all else fails, some model-free exploratory behavior ensues. Basically, this means doing at least something and seeing what happens. Generic heuristics may guide this behavior, making it less random, and more specifically we should expect intuitive responses to reflect Bayesian search for regularities between variables occurring in the task. Such an account as a rational analysis of the selection task behavior has been proposed by Mike Oaksford and Nick Chater (1994). Since exploratory search is an inductive and often iterative process, the subjects should not be expected to attempt a one-shot falsification attempt but behave as if it was an interactive process where the cards displayed represent only a sample from a larger domain to be explored. This might be an odd interpretation of the task; however, but this is how many subjects approach the task when tested in a session consisting interaction with the experimenter (Stenning & van Lambalgen, 2004).

Oaskford and Chater's (1994; 1995; 2007) theory is more elaborate in but, to cut the story short, they counted the selection frequency of each individual card in 34 experiments (with a total of 845 subjects) using standard abstract or arbitrary selection rule. The observed ranking was as expected: $P > Q > \neg Q > \neg P$. What one perhaps would not expect is that, given some reasonable assumptions, this ordering represents the most informative ranking of data samples if one wants to test the hypothesis $P \to Q$ against the null hypothesis that P and Q are independent. These arguably reasonable *rarity assumptions* are that P and Q are sufficiently rare with respect to their contrastive classes $\neg P$ and $\neg Q$.

Assume, for example, that you are set out to find out if all crows (P) are black (Q). Assume further that there are far more birds other than crows, and more birds of other colors than black. Informally, the Oaksford and Chater's analysis is that in this case, you will most likely end up with a bunch of white seagulls, green parrots, etc. if you sample the set $\neg Q$ regardless of whether the hypothesis is true or not. Of course, you only need to investigate the set P; however, if your sample size is limited, it is possible that you just accidentally find the few crows which happen to be black. If you take an equally small sample from the set Q, you would expect *not* to find many crows if the foil hypothesis holds, since the statistical distribution of crows should amount to the whole sample space in that set, and there the rarity assumption does its work. Therefore, finding crows from the set of black birds would make the hypothesis more likely by corroborating the otherwise unlikely correlation. This is the Bayesian solution to the famous Hempel's raven paradox (Good, 1960) applied to the selection task. It explains the pattern of selections as reflecting a rational inductive inference in trying to confirm the conditional rather than a confused attempt of deductive inference in trying to falsify it. In that case, it is a matter of your assumptions about the environment whether it is a good idea to focus on Q or $\neg Q$, and this also depends on what you assume about the generalization $P \rightarrow Q$.

More specifically, assume your hypothesis H_1 states that there is a strong but non-deterministic causal mechanism making, say, around 98% of every P to have Q. In the domain you are investigating, you happen to estimate that 5% of the entities have property P and 15% have Q. You consider H_1 equally likely to the foil hypothesis H_0, which states that there is no statistical correlation between P and Q, making property Q distributed evenly in sets P and $\neg P$. Then if you take samples from $\neg Q$, you expect them to be P 2% of the time if H_1 is true and 5% of the time H_0 is true. This may be significant difference in the long run but negligible with small sample sizes. If you take a samples from Q, you find P about 33% of the time if H_1 is true but only 7% of the time if H_0 is true. It is straightforward to show that with these and a wide range of other parameters where the rarity assumption holds, the Bayesian posterior odds for H_1 and H_0 diverge more sharply if samples are consistently taken from Q, rather than $\neg Q$, making it more informative for the induction task.

Therefore, the overall behavior of subjects in the abstract selection task is rational *if* they are doing this kind of inductive exploratory search with the generic assumptions that the dependency of Q on P is strong but non-deterministic (if it exists) and that the rarity assumption holds. This, however,

is not what subjects are supposed to do in the task (under its logical interpretation), and it is dubious if the rarity assumption is reasonable in the original selection task. Nevertheless, the behavior reflects what I previously claimed to be a natural function of the intuitive mind when there is no prior domain knowledge to guide the search for information. Whether the rarity assumption holds or not in most uninformed inductive tasks (making it a rational default assumption) is another matter, which cannot be decided *a priori*. In any case, similar behavior is a robust finding in causal learning research in that subjects are sensitive to the rarity assumption and tend to focus on c^+ and e^+ events and their possible correlation (McKenzie & Mikkelsen, 2007). Note that under this interpretation, there is no single correct answer to the selection task. It is a matter of ecological and resource considerations as to how optimize search, that is what variables to choose and how many require attention.

The optimal data selection model does not readily explain why the logically correct solution P & $\neg Q$ is prevalent in some versions of the selection tasks—especially the ones using thematic rules familiar to the subjects or an explanation which establishes sensible relation between P and Q and renders falsifying instances semantically natural. Abolishing the rarity assumption makes the logically normative selection normative also in the induction task, and Oaksford and Chater (1995) offer this as an explanation for the findings of Sperber et al. (1995). I find the explanation with these particular findings as rather *ad hoc*. The readers are encouraged to consult the cited papers and make their judgment on this, but, in any case, that would be quite odd as a general explanation of such selection task behavior. For how does the rarity assumption rarely hold with familiar contingencies (as we shall see below) if it is supposed to be a sound generic strategy with unfamiliar rules in arbitrary domains?

I think Oaksford and Chater have found what we are disposed to do when confronted with tasks that we cannot resolve by retrieving knowledge from the long-term memory, which is model-free exploration. The different kind of behavior observed with familiar rules and natural contexts can be explained by the exploitation of previously obtained background knowledge. Below, I discuss why the varieties of the selection task cannot be explained by resorting to a core reasoning mechanism but the explanation needs to account for the interplay of several distinct albeit related cognitive processes.[112]

[112] To be sure, Oaksford and Chater (1994) distinguish between rule testing and rule use, and explain reasoning with deontic rules (explained below) by resorting to the latter. On that issue I am not sure if their theory is incompatible with mine. They discuss how expected

Before moving on, I want to clarify a possible misunderstanding. I have claimed that with the rule "if a card has a vowel, then it has an even number" subjects implicitly treat the conditional as causal. I am not, of course, claiming that the participants are tracking whether vowels cause even numbers. That would not make any sense. The cards are artifacts, and with artifacts the attribute correlations are taken to be caused by creator intentions and manufacturing process (unless they result from the intrinsic mechanical structure of the object). Therefore, not all contingency and covariation learning need to be about direct causal mechanisms. The claim is rather that such probabilistic learning is closely connected to causal induction (Nisbett et al., 1983), and that in the abstract selection task the subjects are exploiting the generic mechanism for uncovering causal structure of the environment.

Lastly, the human mind is an engine that tries to make sense of the present and predict future events based on previous experiences, which makes us highly adaptive species not only to ecological but also to cultural environments. The core reasoning capacities are adapted to adapt to local material environments. Now, in addition to the possible relation of variables mentioned in the cards, another set of contingencies that the participants are learning in the selection task are *normative* event–event or situation–action–outcome dependencies. In the debriefing, the participants learn through experimenter feedback what they should have done. It depends on the experimental setting, of course, whether feedback is available, but generally in these kinds of situations subjects are concurrently learning a procedural contingency structure of social practices, that is how they *should* behave. In the abstract selection task, the practice is artificial and isolated, and hence the subjects approach it without any prefigured sense in mind and whatever they learn does not readily transfer to other contexts. This is based on our earlier discussion that subjects are learning the contingency structure of the specific variables in that specific task context and also what counts as successful behavior.[113] In less contrived tasks this procedural know-how should also convey the sense of the task context by connecting it to other meaningful goals and possibly by contextualizing the relevant variables

utility modulates the selections when deontic rules are applied in a context. Below I try to explain how our subjective understanding of these kind of rules is constituted in the first place, which is a prerequisite for their consistent application.

[113] Note that this is quite different from non-social learning where the agent can find its own goals and standards of success based on what it happens to find rewarding. This does not necessarily mean that in social learning there is something essentially different going on, but only that we find positive feedback from our peers and authorities rewarding, as suggested in section 2.2.2.

with more distant goals. Since the stimulus material, contingency structure, and the task demands are social products, tasks of this kind can be seen as an example of elementary contingency learning in cultural contexts (although there is nothing special in the selection task in this respect). This is a highly hypothetical claim; however, the idea is exploited below that cultural learning can be seen as employing the same mechanisms as empirical learning of material contingencies.

5.1.2 The effect of content and domain familiarity

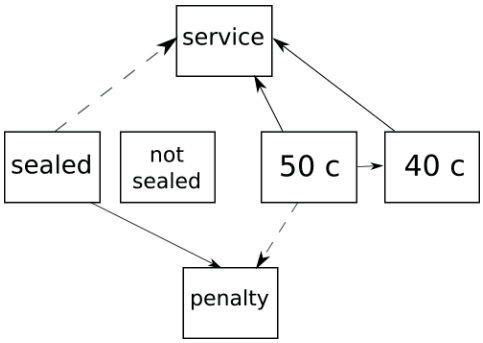
Given that humans are excellent practical reasoners, it is clearly of interest to find out if we reason differently when the selection task is framed with familiar thematic content. A follow-up to Wason's original experiment was conducted with envelopes instead of cards, and the task was to evaluate the rule "if an envelope is sealed, then it has a 50 lire stamp on it." Around 80% of the subjects chose the sealed envelope and the one with less than a 50 lire stamp on it; that is, they made the logically correct selection P & $\neg Q$ (Johnson-Laird et al., 1972). Similar result was obtained with cards containing cities and means of transportation (Wason & Shapiro, 1971).

However, it soon became clear that realistic or concrete material as such was not the facilitating factor. The mentioned results failed to replicate in subsequent studies. It turned out that the postal rule was tested with British subjects that had experience with similar regulation about stamp values and sealed letters; in contrast, North-American subjects without such experience produced the same pattern of responses as in the abstract selection task. These replications were made by Richard Griggs and James Cox (1982). One of their test rule of particular interest was the drinking age rule framed with the following scenario:

> On this task imagine that you are a police officer on duty. It is your job to ensure that people conform to certain rules. The cards in front of you have information about four people sitting at a table. On one side of a card is a person's age and on the other side of the card is what the person is drinking. Here is the rule: IF A PERSON IS DRINKING A BEER, THEN THE PERSON MUST BE OVER 19 YEARS OF AGE. (p. 415)

More than 80% of the participants made the logically correct selection P & $\neg Q$. Virtually no subject chose $\neg P$ or Q cards. The drinking age rule

is generally familiar to the adult population in most countries, and the task has proved to yield a high level of normative performance reliably. Thus, the specific experience with the tested or a similar rule seems to be a decisive factor in competent reasoning.



On the left, there is a sketch of the general sort of knowledge we obtain by such experience. It is an *ad hoc* illustration of a causal model that incorporates the functional knowledge about the postal regulation connecting stamp values and letters being sealed or not. As with the abstract task, we, naturally, do not learn any direct causal information about seals and stamps. Instead, the graph depicts an event–event structure connecting these variables to expectations about services being provided (i.e., letters get sent) and possible penalties imposed (perhaps mere scolding) on attempts to violate the regulation by getting the extra service without paying the extra cost. Solid lines represent generative and dashed preventive relations. Thus, the figure depicts that if you try to post a sealed letter, problems may arise in case you do not pay the required stamp. If you pay the 50c stamp it covers the regular stamp which, is enough to receive the standard service. Hence, by forming this sort of mental model, one can readily grasp that one violates the regulation only by sealing the letter and paying for the lesser stamp.

This kind of situation representation captures one micro-domain of causal knowledge about consequences of actions. More than enabling conditional reasoning indirectly about sealing and stamp values, it also specifies what violation of the rule means in practice: possible penalties and the denial of service. It also contains other pragmatic information; for example that not sealing the letter as such implies no obligations or rights to services. Such representations enable flexible *situated* understanding of contingencies. For example, if the rule is "for access, you must first pay the price" and subjects are told to observe if the authorities enforcing this rule are violating it, they tend to make the selection $\neg P \ \& \ Q$, rarely seen otherwise in the selection task. This is because the authorities are violating the rule if the access is not granted while the price is paid, so the interpretation of the meaning of "violation" depends on individuals' point of view. How specific permissions, obligations, preconditions,

etc. play out depend on the details of the rule and the situations they apply. (Holyoak & Cheng, 1985) Since such knowledge is encoded and processed as situated knowledge rather than decontextualized logical rules, the capacity to give logically correct answers in one task does not readily transfer to other tasks with ostensibly the same logical structure but different content.

However, a curious thing is deontic rules (i.e., rules concerning obligations, permissions, and so on), tend to a yield high level of logically correct responses when unfamiliar or even strange rules are tested, like "if a man eats cassava root, he *must* have a tattoo in his face"—especially if the social character of the rule is made clear (Cosmides, 1989). Thus, we seem to reason differently in deontic matters in comparison to other domains because familiarity is not essential with rules containing permissions and obligations. Perhaps we are hardwired to reason better with social rules? This seems at least *prima facie* plausible for we are inherently social creatures. Famously, Leda Cosmides and John Tooby (1992) have argued for this hypothesis. They claim that we are armed with a cognitive module that tracks cheaters in situations involving social contracts, and hence we readily grasp what violates the rule if it has the form *if you take the benefit, then you pay the cost*, for example. Their theory is more far-reaching and contain a more general hypothesis that the mind relies on a massive amount of self contained modules that enable domain-specific cognitive capacities. This specific explanation, however, is dubious because we find a good level of logically normative reasoning with deontic rules where costs and benefits are not involved. While drinking beer in a pub may plausibly require cheating on the part of a minor, drinking beer is hardly a benefit paid with aging. Subsequent studies manipulated an arbitrary deontic selection task by providing or eliminating a cost/benefit references in an otherwise similar task and found no effect to the normative answer rate (Cheng & Holyoak, 1989). Furthermore, deontic rules lacking any social contract character, such as "if you clean up spilled blood, you must wear rubber gloves", facilitate normative performance (Evans & Over, 1996, 79).

The presence of a narrative that provides a scenario or a rationale for the task also impacts behavior. In one study, the police officer narrative in the drinking age problem was omitted, and the proportion of $P \ \& \ \neg Q$ selections was only 32% (Stanovich & West, 1998) in comparison to 86% in the original study with the narrative. The highly untypical deontic rules that Cosmides and Tooby used were all introduced with a rationale explaining the rule. For example, the cassava-eating rule mentioned above was explained to be about marital status and sexual ethics. Having a tattoo in one's face signifies being

married, and cassava root was told to be a potent aphrodisiac. Given such context, the rule "if a man eats cassava root, he must have a tattoo in his face" corresponds to the familiar norm (even if not nowadays very strongly enforced) condemning extramarital sex. Cheng and Holyoak (1985) found that explaining the rationale of a deontic rule increased the normative selections from about 60% to more than 90%. One of the rules they used was the postal regulation rule "if an envelope is sealed then it *must* have 50 lire stamp on it" (emphasis added). Note that in (Griggs & Cox, 1982), the postal rule was in the indicative form (i.e., without the word "must"), and they found no difference in performance to the abstract task; indeed, only one out of 24 subjects gave the correct answer. However, in Cheng and Holyoak's study, the rule was rendered deontic by adding a mere word and the normative selection rate was found to be 60%. Explaining a rationale had no effect on the subjects who were already familiar with the rule because they already understood the point and performed at the ceiling level. Thus, using specific content with an explanation or rationale seems to make the task psychologically concrete, leading to a performance that is indistinguishable from subjects who have acquaintance with similar rules.

Cheng and Holyoak (1985) used a short narrative with the above-mentioned tasks; however, they also found that deontic rules needed no explanation or specific content to facilitate performance. The schematic rule "if one is to take action A, then one must first satisfy precondition P" yielded the same rate of logically normative responses. Therefore, while deontic rules combined with an explanation leads to response behavior that is indistinguishable from familiar rules, it is enough to indicate that the test rule is about permissions or obligations to achieve a substantially good level of logically normative selections. So qualitatively, this pattern looks much the same as with abstract and familiar non-deontic material: (1) the concreteness of the rule is not a factor but strict familiarity *or* (2) an explanation that makes the task semantically transparent. Logically correct answers are facilitated with whatever content whenever the problem formulation indicates clearly that the task is to focus on $P \,\&\, \neg Q$ (Sperber & Girotto, 2002). The difference is on the base rate of logically correct answers, which is above 50% with unfamiliar or abstract deontic rules and around 10% with unfamiliar or abstract non-deontic rules. Moreover, it takes less effort to instruct the participants to reach the ceiling performance with unfamiliar deontic rules than with the abstract task. Cheng and Holyoak concluded that the exemplar-based explanation of deontic

reasoning cannot hold but the knowledge exploited is encoded as more abstract reasoning schemata.

It has been pointed out (e.g., Stenning & van Lambalgen, 2001, 2004) that the deontic task is logically different from the standard task. In the deontic versions, subjects are not checking whether the rule holds or not. Instead, they are instructed to find possible violations. This removes several possible sources of misunderstanding because, under the deontic interpretation and such instructions, subjects usually understand the relevance of P & $\neg Q$. However, Cheng and Holyoak actually instructed their subjects to monitor if the rule is followed rather than violated. Sieghard Beller (2012) found that when the deontic character of the schematic rule was made clear to the subjects and violation instructions were used, the performance improved from what Cheng and Holyoak found. Hence, deontic reasoning seems to be easy if only one understands to engage in such reasoning, and it is not just the instructions that do the trick but also the subject's previous understanding of deontic concepts they bring to the experimental situation. Beller interpreted these results in the dual-system framework as showing that deontic reasoning is executed by System 2 process because it looks like a rule-like, abstract, and normatively correct capacity.

I have maintained that even if the reasoning is executed as System 2 or any other explicit process, the competence must be rooted in intuitive know-how. If that is the case, a mere reference to System 2 is not an explanation of the competence without explaining the etiology of that know-how. Note also that Evans (1996) provides evidence against Beller's conclusion. Strongly facilitating rules tends to prompt quick and pragmatic rather than slow and analytic processing. (Unfortunately, specifically abstract deontic rules were not tested.) Similarly one can reply to Stenning and van Lambalgen that the facilitation in deontic tasks is not merely the effect of the instructions used. With abstract tasks, the subjects need substantially more coercion to understand and follow the intended logic. Hence, we need to account for where this enhanced ability to understand deontic logic comes from. Below, I explain that the conclusion reached by Cheng and Holyoak is not necessary, and it is probably misleading if it means that deontic reasoning exploits qualitatively different type of knowledge than practical reasoning with familiar content. Specifically, I try to show that (a) there is nothing *psychologically* extraordinary in deontic reasoning but ecologically (or more specifically *socially*) there is compared to many other conceptual domains and (b) the capacity to conform

to and interpret schematic rules does not necessarily indicate explicit rule-based System 2 processing.

5.2 Constructing schematic knowledge bottom-up

The concrete, exemplar-based memory encoding means that whatever we learn through action or observation, the resultant knowledge is associated with specific contents and contexts and does not contain general abstract information. As content manipulations in selection tasks indicate, the capacity to handle one task does not usually transfer to another even if both share the same logical structure. For example, the logical form of the drinking age rule is arguably the same as the postal regulation rule; however, the subjects who were familiar with the former gave different answers in the latter task if they were unfamiliar with the postal regulation. This is what one should expect under the pragmatic situated reasoning theory I that am proposing. If you do not have previous acquaintance with how stamp values relate to sealed letters, then whatever knowledge you may have about stamps does not help you in grasping this specific relation. While the material implication may capture the general form of material reasoning, which is also applicable to this particular task, and we can learn to comply with it in any given domain, human material reasoning is still not guided by form but by content. Thus, the capacity to understand one task does not readily generalize to others. Moreover, while we are able to employ formal rules by the laborious exercise of reflective thought, our fluent intuitions are sensitive to specific contextualized contents.

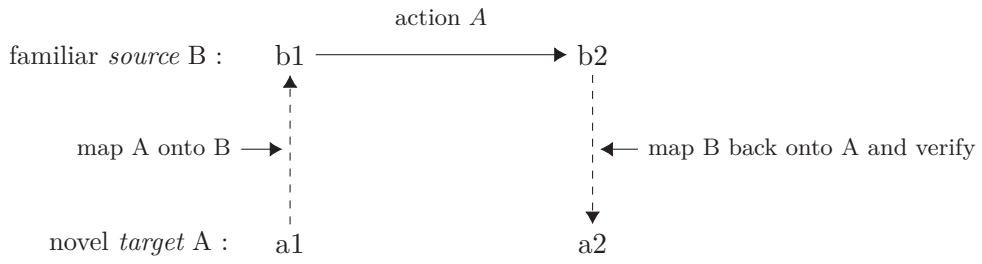
But what is the point of associating procedural knowledge to specific surface attributes rather than to the actually relevant functional form? Logically, nothing for the whole point of logic is to abstract away from content and focus solely on the form. Ecologically, however, it makes sense, given our cognitive limits. Similar surface features may quite reliably indicate specific functional contexts and structures. Indeed, our categorization processes are sensitive to features that indicate functional differences, and they partition the world into pragmatically relevant classes. In specific recurring event types, the members of these classes presumably tend to correlate predictably and provide recurring affordances and cause predictable effects. For agents with a vast behavioral repertoire, knowledge selection is the central problem in complex open environments. Perceptually cued (i.e., surface feature sensitive) pattern recognition allows a memory system with a very efficient content addressable search, as evidenced by our rapid categorization capacity and the long-standing research

in machine learning. Exemplar search is precisely sensitive to feature correlations, and that is what you need to search for when identifying situations. Focusing on correlations reduces the search space by restricting it to attribute conjunctions (or it may weight the search results if exemplars are retrieved *en masse*). Indexing situation exemplars with goals and context cues further narrows search and enables more flexible concrete knowledge representation by allowing different task sets to be associated with stimulus equivalent situations.

All this makes the deployment of procedural knowledge highly effective because it makes knowledge exploitation basically a matter of recognition. This comes with a cost, however, because in novel contexts without familiar contents such recognition is of little use. Learning new things by trial and error can be laborious—especially if it involves learning complex hierarchical plans—and in the worst case even dangerous. Fortunately, clusters of action–outcome pairs encoded in long-term memory can be dissociated from their specific contents and transferred to novel situations. If that works, even approximately, learning does not need to start from scratch and the transferred actions sets can be later optimized for the new contexts through further learning. Hence, while surface content may indicate relevant functional structure, similar functional structures may also unite situations with different surface contents by such transfer. That is the basic idea of analogical transfer and reasoning.

5.2.1 Analogical reasoning

I assume the basics of analogical reasoning are familiar to most but, in a nutshell, it goes like this: You face an ill-understood problem A while you already master another task B; that is, you have a pretty good idea what happens if you manipulate a variable b associated with situation B. Then the thought crosses your mind that how about if you substitute variables in A with variables in B, solve B, and see if that gets you what you want when you map the solution back to A. Here is a picture:



The following classical example in analogical reasoning literature originates from Gick & Holyoak (1983): In radiation therapy, ionizing radiation is used to destroy tumors. The problem is that the ray also destroys healthy tissue along its path. So what to do? The following story may help figure that out: A general is about to capture a fortress. Several roads radiate outward from the fortress; however, they all have been mined so that a small group of soldiers can pass them safely but any large group (necessary for successful attack) triggers the mines and foils a full-scale assault. The general divides his forces into small groups to travel different roads and converge at the fortress at the same time.

Now, substitute a group of soldiers with a single ray and the number of soldiers with the intensity of the ray; here is your answer to the radiation problem. Note that these problems are not entirely analogous: You destroy the tumor with one full intensity ray but you do not conquer the fortress with a full-scale attack from one direction. It is a common feature in analogical reasoning that some relevant aspects of the source and the target problems are similar while they do not share a completely identical structure.

Traditionally, it is held that problems (or situations) can be characterized by (a) *surface (or thematic) content*, which consists of non-relational properties of task elements and (b) a system of relations called a *deep structure*, which determines how the elements are related and what the outcomes are when they are manipulated. Analogical reasoning has taken to consist of the above-depicted sub-processes that are more or less independent: (1) finding a potentially useful source analog from memory, (2) finding a mapping of corresponding elements between the source and the target, (3) deriving inferences based on the mapping, and (4) adapting the solution to satisfy the constraints of the target in case the analogy is imperfect (Holyoak et al., 1994). Accepting a candidate analogy has been considered to depend on the similarity between deep structures of the source and the target, and similarity is traditionally measured as a degree of formal isomorphism between the relational structures (e.g., Gentner, 1983).

In addition to discovering how the deep structure of the source is retrieved and adapted to the target task, the research has been concerned with how knowledge of the structural schemata, which is abstracted away from the surface content of the source, is induced in the process. But thinking of deep versus surface features simply in terms of relational and non-relational properties can be misleading. A better definition is that deep structure consists of properties that are relevant to goal attainment—relational or not—and by implication,

they are not fixed intrinsic features of the source analog but also determined by the current reasoning task (Holyoak & Koh, 1987). Often, though, such properties are relational properties of a causal kind. In what follows, I will use the term *functional* rather than deep structure. This is the causal structure that is tracked by our mental situation models and action–outcome predictive models.

Here I adopt a notion of schemata from Holyoak et al. (2010, 703) that schemata relate to analogical situations the way categories relate to instances: Analogies embody a schematic structure that can be projected across specific situations, and schemata can be abstracted from situations that can be meaningfully grouped by the schema. In the above example, the idea of distributing a powerful force to travel several paths to avoid unwanted destruction and integrating it in a specific point, is the schema projectible across the two tasks. As an example of commonsense schema it is probably not very representative because such a pattern does not frequent in our daily affairs. While the idea is easy to understand, we do not have a name for it and it does not stand out as a lexicalized concept. This is probably why the radiation problem is difficult, even though the solution is quite obvious once pointed out.[114] More familiar schematic concepts are, for example, *obligation* and *getting dressed* and folk physical principles such as *centrifugal force*, which guides our lay understanding of everyday mechanics.

The idea, that analogical transfer is a matter of structural comparison is approximately correct; however, a purely syntactic approach is not. Suppose that a candidate source consists of three causes for effect feature E. Two of them, G_1 and G_2, are generative, and one of them is preventive P. Call this model G_1G_2P. Lee and Holyoak (2008) found that their subjects regarded G_1G_2 to be more similar to the source model than G_1P, demonstrating that similarity judgments are affected by the causal properties of the structure. No such effect was found when comparing G_1G_2 and G_1G_3 against the $G_1G_2G_3$ model. Subjects were explained how features G_1, G_2, and P relate to effect E, and they found this information to more strongly predict the occurrence of E in the G_1G_2 model than in the G_1P model. For example, in one task the subjects were told that "[f]or Animal A, dry flaky skin tends to PRODUCE blocked oil glands; muscular forearms tend to PRODUCE blocked oil glands; a weak immune system tends to PREVENT blocked oil glands" (Lee & Holyoak, 2008,

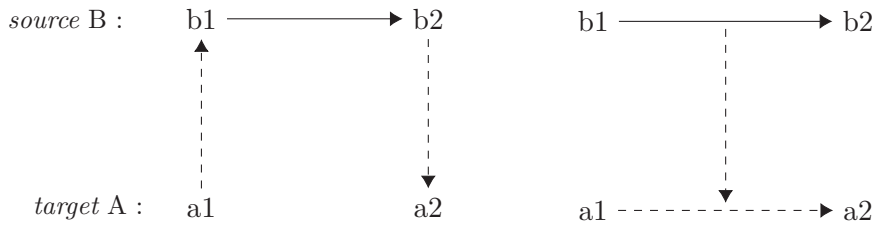
[114] In Gick & Holyoak's (1983) original study, only 10% of participants solved the problem spontaneously. 75% of the subjects managed to solve it once they were given the story about the fortress and hinted that it may help.

1114). Then the participants expected more strongly animal B with dry flaky skin and muscular forearms (G_1G_2) to have blocked oil glands than animal C with dry flaky skin and a weak immune system (G_1P). That is not very surprising given that the source and the targets exhibit the same specific cause properties. However, this effect was also observed in cross-domain inference where the source analogy was a story about a discovery in chemistry (synthesis of a substance) and the target was a hypothesis in astronomy (the forming of a super-star), and thus both the surface properties and the presumed underlying causal mechanisms differed.

Based on these and related findings Lee & Holyoak (2008) and Holyoak et al. (2010) have concluded that causal understanding and analogical reasoning are closely related. Since the 1980s, Holyoak has pursued the idea that component processes in analogical reasoning are constrained by pragmatic factors, namely that analogies are retrieved and evaluated in accordance with agents' current goals (e.g. Holyoak, 1985). The alternative is to emphasize formal structural similarities because not all intelligible analogies are causal or need to appear in problem solving contexts. Gentner (1989) has argued that this is evident in metaphors such as "All rising to a great place is by a winding stair," and purely structural analogies such as "if $abc \rightarrow pqr$ then $abd \rightarrow pqs$." However, intelligibility is one thing and use and retrieval another, and purely structural constraints are rather weak if one wants to exploit existing knowledge for novel problems. Through our earlier discussion of the selection task, we know that structural formal knowledge does not transfer well across tasks if its relevance is not noticed, and this often requires content and context cues. Earlier, we also saw that we are able to employ decent probabilistic heuristics in everyday life, but that ability seems to suddenly disappear in unfamiliar events. People seem to be good at understanding structural analogies; however, they are also weak in spontaneously exploiting structural knowledge. Moreover, meanings of metaphors and idioms are not transparent by their structural correspondences to their intended interpretation, but their meanings are extracted from their use in discursive practices (Keysar & Bly, 1999). In any case, analogical reasoning operates under multiple constraints which are familiar from categorization: at least feature similarity, structure, and purpose (Holyoak & Thagard, 1997).

Despite different emphases and some disagreements, the general significance of pragmatics and causal knowledge has always been widely accepted in the analogical reasoning literature. For a brief discussion see (Holyoak et al., 2010) where the authors also introduce a computational model of analogical

transfer based on Bayesian theory of causal induction. This model utilizes Bayes nets to represent source analogs, build corresponding (partial) models to interpret targets and run the generated target model to see what comes out. Based on our earlier discussion, I accept this as my preferred account of analogical inference. The model makes, *inter alia*, the above-mentioned unique prediction that causal polarity (preventive vs. generative) has an impact on analogical reasoning. Without going into details, the model has a quantitatively imperfect but qualitatively good fit to human data. Note that in this theory, causal knowledge has a dual role: it works as a structural constraint for selecting a source and provides a model for inferences about the operations of the target. Hence, the outcomes of analogical inferences are not directly imported by mapping from the source but simulated at the target. The idea is depicted on the right and the old model of analogical transfer on the left.



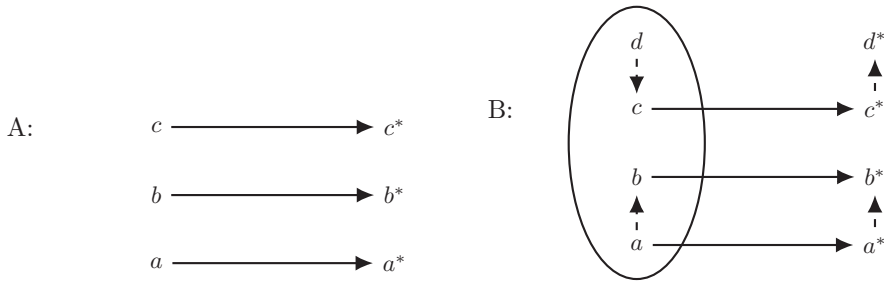
5.2.2 Learning schematic concepts through analogical transfer

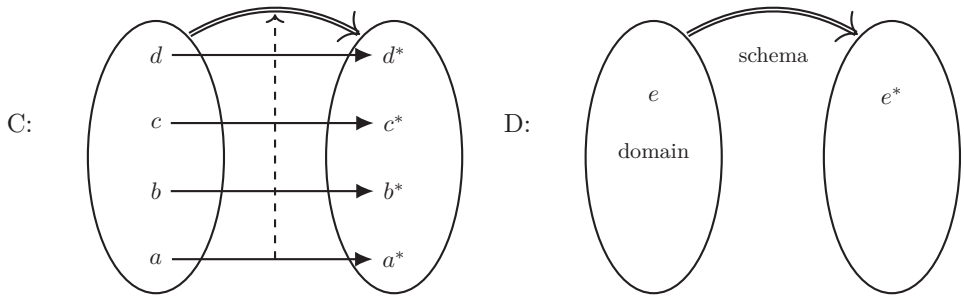
Mixing causal and analogical reasoning may explain how causal hypotheses are generated in the first place by directing focus on the areas of the search space that contain familiar or credible models (Lee & Holyoak, 2008). The surface content is another highly constraining factor, which explains why the cross-domain transfer is difficult. In a familiar domain, models of the target can be partly retrieved by resorting to known regularities between familiar target variables and partly inferred by analogical transfer. Experts know more about a domain than novices and are hence in a better position to retrieve relevant functional information. Moreover, they have a better understanding of the structural principles in their domain of expertise and are more able to project such knowledge across tasks. This suggests that analogical transfer is not only a matter of mapping specific graphs between tasks, but there is also genuine higher level schematic competence that guides domain understanding and is dissociable from specific contents. This is apparently also the case with folk theories and other commonsense conceptual domains like deontic reasoning. By

resorting to the situated representation and processing assumptions discussed earlier, I suggest that this happens in the following way.

Think of an agent who learns to cope with situations a, b, c which are psychologically distinct. That is, the agent grasps the functional structure of these tasks and they are learned separately in different contexts and have sufficiently different surface contents, so no commonalities between them are recognized in the largely stimulus driven processing. This is the situation as depicted in entry (A) below. Unknown to the agent, a substantial overlap happens to be in the functional structure of these situations such that under specific tasks, they are actually analogous. In other words, a general action policy A in situations a, b, and c would lead to predictable outcomes a^*, b^*, and c^*, which would be relevantly similar in terms of goal attainment in each of these situations. Models of these events may be complex, but for the sake of clarity the idea is depicted below as simple situation–outcome pairs $x \rightarrow x^*$ associated with a specific course of action.

(B) In time, the agent learns that something connects these situations so that there is some pragmatically relevant mapping between them, which allows one task—as well as some novel problems (d)—to be resolved in terms of others. At this point, the agent does not need to know what is it exactly that connects these situations. The observation can derive solely from tacit know-how associated with implicit task set selection. Despite decades of research, the process of flexible and creative selection of (especially cross-domain) analogies is not perfectly understood, but we know that (a) explicit attention to the structural commonalities makes analogy learning easier and analogy use more fluent; however, (b) we can learn and often use analogies without explicit awareness of this. In other words, resolution from one problem to another can be transferred implicitly without recollection of the source (Holyoak et al., 1994; Reeves & Weisberg, 1994; Gentner et al., 2003; Kostic et al., 2010).





At point (B), the agent is in the process of learning a domain implicitly by a bottom-up process. The more the structurally identical but superficially divergent exemplars accumulate in the long-term memory, the easier it becomes to find an analogical task set for novel problems. This stems from the availability of more context and stimulus cues and because similar task sets are increasingly replicated in the long-term memory, promoting their selection when exploratory sets are called for. This is a gradual process where problem-solving becomes behaviorally more flexible, but the agent does not need to acknowledge what it has learned or even that it is improving in the mapping aspects of specific situations or "micro-domains" across tasks. The more the clusters grow, the more effective they become in facilitating transfer and the more they function like a domain-specific reasoning schema that can be applied to novel situations.

(C) At some point, some of these structural relations become acknowledged through a recurring practice of exploiting them. A schematic understanding of what unifies these micro-domains is produced—i.e., explicit understanding of their functional structure. This includes event–event causal patterns and *action→outcome* components of task sets, which can then be more flexibly dissociated from specific stimuli and projected across tasks. This stage can be reached without having a specific name for such schemata, but linguistic cues associated with the processing episodes of their constitutive exemplars may serve as contextual markers that enable the selection of relevant task sets in—and hence as a procedural interpretation of—unfamiliar situations. Moreover, as we discussed in connection with psychological essentialism (see also Gentner & Boroditsky, 1999, 245), lexicalization also promotes the search for commonalities (at least with concrete categories), thus possibly facilitating schema induction by directing attention to aid category formation.

This process may give way to stage (D) where schematic knowledge provides know-how that can be exploited without necessary recourse to specific

contents and situations. That is, if novel problem e is identified as belonging to the domain of learned procedural schema, it can be directly exploited to make inferences about e. If these inferences about e are then verified through actual action to be valid, the resultant know-how about e becomes incorporated into the stock of specific SAO knowledge base that constitutes the schema. Such identifiable patterns in our daily practices get labeled, allowing their communal identification without describing their actual structures and associated procedures. (This is similar to how category labels communicate information without describing category contents, which is often implicit and unknown to language users anyway.) This engenders discursive and other symbolic uses of such concepts—or more precisely, uses of *words* that refer to these situations and procedures—which affords communicating and reasoning about them outside the empirical context where they were learned.

That is my explanation of abstract schema induction based on concrete knowledge. Analogical reasoning explains how it is possible to reason in novel situations even while situated reasoning exploits exemplar knowledge associated with familiar events, and analogical transfer explains how the development of domain-specific cognitive expertise emerges from that practice. The shift from novice to expert results from accumulating knowledge and not *primarily* from changes in reasoning procedures or cognitive resources. This incidentally explains the ostensibly unique behavior in deontic selection tasks. Our proficiency in deontic reasoning is not due to specialized processes for social cognition but results from domain general learning mechanisms that use instance-specific encoding and tracks pragmatic causal regularities in our everyday environment. There is thus nothing cognitively special in deontic reasoning.[115] The observed performance stems from the fact that most of the rules we learn throughout our life are by and large deontic. We cope with such requirements on a daily basis—especially in childhood when considerable learning takes place—and we have thus learned especially well what it means to follow or violate rules of this generic type. Thus, obligations, permissions, etc. are heavily present in social learning, rendering learned deontic rules and contexts abundant. This is why any adult who has gone through normal so-

[115] This is not to say that there is nothing special in social cognition in general. I do believe that some capacities are specifically evolved for social cognition, but I also believe that they are mostly irrelevant for explaining how we understand specific conceptual domains, deontic included. In the next section I discuss how social cognitive adaptations may be necessary for explaining general discursive reasoning capacities.

cialization is an expert deontic reasoner, and this is why sparse linguistic cues, at best a mere word, may suffice to prompt a deontic interpretation of a rule.

A narrative or related description of the functional context may further help in finding a relevant analogy. For example, the odd rule "if a man eats cassava root, he *must* have a tattoo in his face" in (Cosmides, 1989) was accompanied by a story which implied that the rule was about sexual ethics. Cassava was told to be a potent aphrodisiac, and the tattoo signified a marriage status.[116] One may object that maybe the context description or a narrative simply makes it easier to imagine the situation and that is what actually helps in resolving the task. But then again, I am here precisely trying to explain what such imagination means and, more to the point, how it helps. Note that if analogical reasoning works by importing functional structure from a source to the target rather than simulating the source—like Holyoak et al. (2010 claim—one will not need to consciously solve the problem by thinking of the source. This is also how task set transfer is supposed to work, that is dissociating the functional knowledge from the source and (provisionally) associating it with the target. Explicit thinking is generally considered to facilitate analogical transfer, and hence it may be that when effortful reasoning helps in these types of tasks, it is mostly due to searching and verifying candidate analogies and task sets rather than due to logical inference (see also Evans, 1984).

This explanation has implications beyond mere deontic reasoning. All familiar domains involved in practical reasoning should be accompanied by a comparable competence of intuitive understanding; for example, folk physics and folk psychology. Amit Almor and Steven Sloman (1996) tested non-deontic rules, which were somewhat transparent analogs of familiar everyday principles. Logically normative answer rate was found to be from around 30% to 60%. Most errors were $\neg P$ card omissions, not Q card inclusions. The best performance was observed in tasks that were designed to map onto folk physical source analogies.

Another implication is that the mere knowledge of formal logic does not help in the abstract task, but subjects that routinely use formal logic should be in a better position to interpret the task as considering the validity of the logical implication. Jackson and Griggs (1988) confirmed this by showing that while the level of education did not predict an advantage, mathematical education did and more so among doctoral than bachelor level students.What these mathematicians were presumably doing was task set retrieval from previous

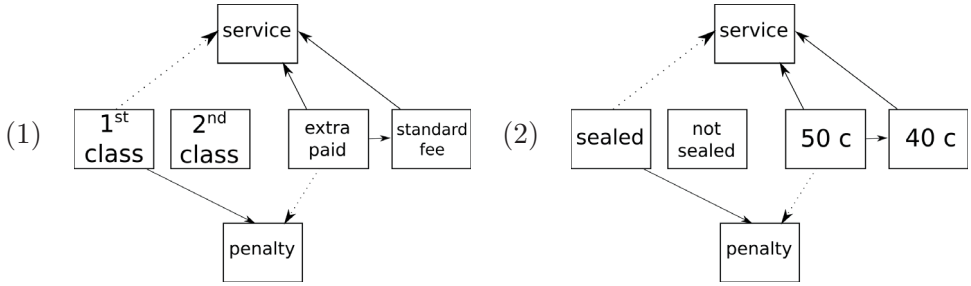
[116] Griggs and Ransdell (1986), for example, have offered similar explanation that selection task performance is best explained by memory cueing plus analogical reasoning.

logic problem contexts. Assuming that the abstract selection task is not obviously a logic problem by its very nature (in a psychological sense), the task set retrieval is psychologically speaking analogical transfer; that is, application of existing knowledge to a novel context. Indeed, I assume the procedural knowledge retrieval in general to exploit exemplar representations of past situations and, as Medin and Schaffer (1978) noted, exemplar classifiers are inherently analogical reasoning systems.

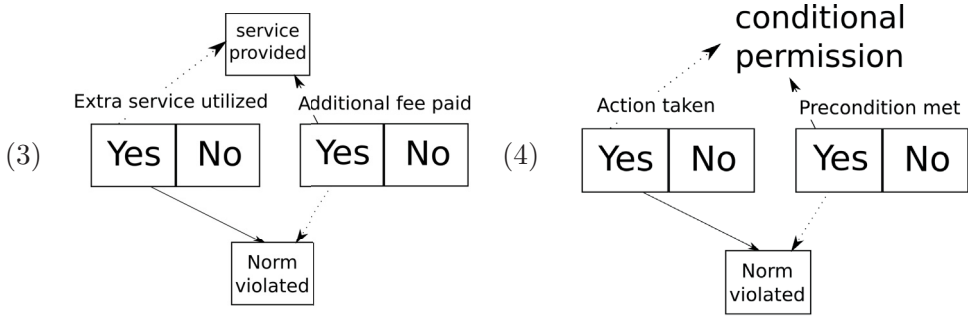
Lastly, the problem of importing knowledge from familiar tasks to unfamiliar ones, especially to abstract selection task, is not only due to different surface contents and learning contexts but also structural features contained in situation representations. The *ad hoc* depiction of the mental representation of the postal regulation is again depicted below. Note that it contains relations to specific services and penalties, but presumably no corresponding events are expected in the abstract selection task. Hence, they should be absent in the task's mental representation. Therefore, whatever logical similarities these rules share (i.e., the abstract rule and postal regulation rule), there is hardly any structural correspondence between their functional mental representation.[117]

While the functional representation of the postal regulation rule does not bear structural correspondence with logical inference rules it corresponds to many real world situations, even if the similarities are not always noticed. I assume that most of us are familiar with the rule that you need to pay for the first-class ticket if you take the first-class seat on a train. Entry (1) depicts the causal model representation of this fact. Many of us are unfamiliar with the postal regulation considering sealed letters; however, it is easy to transfer practically identical functional structure to that rule by pointing out that (2) sealing the letter is an extra service, like the first-class seat, for which you need to pay an extra price. Both rules apply to situations that have mostly the same *causal* structure.

[117] However, inference rules and conditional permissions share certain practical commonalities, in that both are learned in social contexts where the application is socially sanctioned rather than determined by the intrinsic causal properties of the reasoned variables. Of course we do, e.g., *modus tollens* type inferences in material contexts where incorrect reasoning imply incorrect factual predictions, and the feedback comes from non-social environment. However, the point is that these inferences are learned in specific material contexts and consider specific concrete contents. When we learn logical inference rules *as* logical inference rules (i.e., as rule conceptually considering formal logic), the learning contexts are social. This is not a conceptual fact but an empirical fact about the psychological etiology of formal concepts.



Moreover, the explanation connecting (1) and (2) contains the functional commonality between these two regulations, which is a general idea of how additional services work (3). From there, one can take a step further and incorporate other related higher-order pragmatic schemata to the mixture by pointing out that this is just a special case of more general principle about conditional permission where meeting the precondition does not oblige you to anything but releases the ban on taking a specific action (4). This is just another way to present the same point about schematic abstraction as with the figures (A) to (D) earlier.



A few remarks are worth noting here. First, this is not supposed to be any sort of stage theory about abstract concept learning. It resembles closely the *representational redescription* theory proposed by the developmental psychologist Annette Karmiloff-Smith (1992). According to her, the development of domain knowledge goes through specific stages where the initially concrete and procedural knowledge is redescribed to different representational formats in three stages, producing pretty much like the above-described abstraction process. However, the learning process I envision does not produce qualitatively different representations in each step. It is more about developing functional capacities. The shift from (A) to (B) happens gradually as knowledge about different recurring situations accumulates, which leads to the more

or less spontaneous observation of functional correspondences between them. The representations of situations *a*, *b*, *c*, etc. are supposed to be potentially complex models that are also learned in a piecemeal fashion. Hence, it is not just accumulating knowledge of *different* examples that help here but also the increasing knowledge about singular event types that makes it easier to find relevant structural similarities between them. The shift from (B) to (C) happens when the learner spots what these similarities are. This is based on a long-standing result in analogical reasoning research that schema induction involves accumulation and comparison of several examples, and the process can be facilitated by varying the learning contexts and contents, which allow the learner to spot relevant variables and commonalities between instances (Gick & Holyoak, 1983; Holyoak, 1985; Holyoak & Koh, 1987; Novick, 1988; Gentner, 1989; Ross & Kennedy, 1990; Reeves & Weisberg, 1994; Gentner et al., 2003), and supported by similar observations in developmental psychology (Baillargeon, 1998) and exemplar category research (Homa & Vorsburgh, 1976; Medin & Schaffer, 1978).

Moreover, these are mandatory, incremental, and non-terminating learning processes that endure the whole life; although if some procedures adapt to external and internal contingencies optimally, they presumably reach a stable plateau as long as these conditions remain unchanged. Second, the learning trajectory has nothing inherently to do with childhood development but with the development of novel conceptual capacities in general.

But does the shift from (B) to (C) not produce a separate representation of the induced schema? Recall that while task sets are learned in association with specific contents, they contain *action→outcome* elements that can be transferred to other contexts and stimuli. This does not require that procedural knowledge be communicable or consciously accessible. Such transfer can happen completely tacitly, and in order to transfer knowledge at will, the agent needs only to grasp that it has a relevant capacity to cope with specific contingencies that recur in another context and to be able to recognize when such transfer is helpful. Such understanding may be explicit but this does not imply that the subject has access to the procedural knowledge enabling that understanding. The implicit retrieval of such knowledge is the basis of understanding situations, and the understanding of one's competencies stem from actually exercising them rather than from introspective enlightenment. Phenomenologically and psychologically, we attend to aspects of situations rather than our cognitive faculties. Likewise, we know that we know what chairs, birds, or tools are; however, we are unable to access the detailed knowledge

that constitutes our understanding of these things. (That is what similarity-based theories taught us.)

Hence, what is needed for getting from (B) to (C) is a shift of focus from surface cues to specific patterns and practices that recur; it is also necessary to understand what affordances and other functional commonalities they share beyond trivial appearances. The capacity to group distinct situations based on such pragmatically relevant patterns is the key here and not new type of representations. However, I do not have an argument why some kind of representational redescription could not happen in this process. I only try to explain why it is not theoretically necessary, or more specifically, if some redescription is taking place, it only requires bracketing the S part of SAO in task sets or to turning causal graphs nodes from fixed entities to open variables. Producing entirely new types of abstract representations is not needed.

Vinod Goel et al. (2004) conducted a neuropsychological study on Wason selection task. They investigated how the reasoning of patients with frontal lobe damage differed from healthy controls. Previous research had discovered that, consistent with our discussion in this and section 4.3, familiar content activates areas in the lateral PFC and temporal areas dealing with linguistic meaning while abstract and unfamiliar content is processed in distinct areas in the parietal system (Goel, 2007, see also). Accordingly, the performance of healthy controls and patients with frontal lobe lesions were indistinguishable in the abstract selection task. However, when tested with the concrete drinking-age rule and the abstract permission schemata, they failed to benefit from the rule content in *both* conditions without significant difference between the tasks. This is consistent with my claim that pragmatic schemata and concrete knowledge are processed at least in the same cognitive systems even if they do not utilize exactly the same knowledge.

It should be kept in mind that, as with event identities, procedural knowledge contains levels with higher-order identities. Hierarchically structured actions tend to be multiply realizable at higher levels. When executing these higher-order scripts, one needs only to monitor if the lower-level routines are executed successfully rather than mind their detailed implementation. Even if planning and problem-solving is a sort of simulation of concrete events, agents need likewise to attend only to such gist levels of tasks if they know that they can handle the component tasks. Hence, I seem to assume something that can be interpreted as abstract representation of problems or procedures. However, earlier I explained that such a superordinate level understanding depends on grasping lower-level specific events and acts instead of on formal understand-

ing of task structures. Hence, there is no (necessary) switch at any point to explicit symbolic or formal representation. Once the common functional structure that unifies these lower-level situated actions is understood, it becomes easier to incorporate novel behaviors and situations under the existing action schema, thus affording more flexible behavior and necessitating less attention to specific details. So the overall *competence* affords more abstract description; however, this does not necessitate that the processes and representations responsible for that competence are qualitatively different from those underlying concrete situated reasoning. The basic idea is comparable to recursion in computer programming: You give some recurring routines their identity (i.e., define a subroutine or a function), and then you can define procedures at a more abstract level by compounding these routines. The difference is that the psychological subroutines are not rigid sets of well-defined instructions to manipulate symbolic expressions but incrementally learned and open-ended patterns whose identities and contents are constructed by repeated interactions with the environment.

It should be clear that a similar process is happening in schema learning through analogical transfer. This is how skill learning promotes abstract schema learning, given that schemata are functional structures extracted in goal attainment and that higher-level action schemata consists of schematic functional knowledge that incorporates lower-level situational know-how. Indeed, the whole schema induction goes that increasing expertise in a domain produces an intuitive understanding of specific events, which promotes learning their higher-order functional structures. Hence, the more skilled can pay attention to increasingly abstract regularities, and they can more flexibly project highly learned task sets to novel situations. This is a species of domain knowledge where the agent knows lots of contingencies between variables, recognizes numerous recurring patterns, and is able to understand situations on the basis of higher-level regularities. Likewise, in analogical reasoning, the received view is that experts gain schematic domain knowledge, which is based on projecting schematic patterns over sets of varying surface contents. They become more able to comprehend problems and analogies by their structural features, while novices approach problems by concrete exemplars and spot analogies mainly by specific surface contents (Ross & Kennedy, 1990; Reeves & Weisberg, 1994). Moreover experts become better at finding cross-contextual mappings and evaluating their validity (Novick, 1988; Gentner et al., 2003). This is most likely because they can readily retrieve frequently used task sets and because they have more refined and simply more numerous functional schemata available.

So there are marked similarities in how expertise and abstraction are construed in both skill learning and analogical reasoning research, and here is an explanation why.

When it comes to incremental learning of domain knowledge, Karmiloff-Smith's theory can be criticized on the grounds that leaning of conceptual domains does not happen in distinct stages. Instead, it initially develops by learning distinct event types and contingencies. Infants learn to categorize events and predict how specific variables operate in each event type. By implication, they sometimes need to relearn how a variable operates in contexts that they do not identify. Now, there might be an issue with terminology here. One can think of a larger domain such as commonsense mechanics, as constituted by microdomains that are learned separately and integrated later. However, these microdomains, as identified by Baillargeon (2002), for example, are very specific, like occlusion and containment events, balance and works of gravity, etc. So it is perhaps better to speak of generalizations over specific event types or contingencies rather than domains. Observed commonalities generalize to general principles while unexpected outcomes trigger a search for new variables that could explain the discrepancy between expectations and outcomes. These notes on conceptual development, emphasized by Baillargeon match with my account. Proper conceptual domains seem to develop as an incremental mixture of better learned contextual schemata and more specific event exemplars. The learning process tracks functional similarities and differences, producing increasingly refined ontology of the environment. If analogical reasoning is the source of schematic domain knowledge, one should expect that generalizations achieved in one task bolster structure learning in others associated with the same conceptual domain. This is because the intra-domain transfer is facilitated by shared structure and surface cues between better and worse mastered contexts. This may provide a scaffolding that leads to rapid domain-wide learning, yielding cognitively continuous but behaviorally fast stage-like transition in *competence* once some schematic knowledge of the domain is attained.

Theorists who concentrate on top-down learning (e.g., Hubert and Stuart Dreyfus) emphasize how by following rules of thumb people learn procedural knowledge which they are unable to articulate. In bottom-up learning, the main subject in this section, the picture looks somewhat different. People instead learn gradually to articulate the regularities that they discover once they explicitely notice them (see Reber 1989, 229; Sun et al. 2005, 161). However, these accounts are not in conflict and they both claim that the inaccessible part is intricate tacit knowledge, while the verbal reports are sparse and elliptical

description of broad salient aspects of observed recurring patterns. Note that conceptual definitions usually have this character even while analytic definitions are often considered as the prime examples of explicit knowledge. True, in some domains definitions work as accurate constitutive rules. This is the case in mathematics. However, with ordinary concepts, analytical definitions are more akin to heuristic remarks that try to capture important aspects of how we use the concept under analysis. *A bachelor* is *a male that has never been married*; however the pope is, probably, still not a bachelor, and other exceptions are easy to invent. The classical definition of *knowledge* captures the essentials of the notion but arguably it is inaccurate because it misses the Gettier examples. We first tacitly learn to apply these concepts in broader real-life contexts, then figure out the definitions that seem to capture their use, and then use our underlying intuitions to evaluate the proposed definitions. Unfortunately, rarely does the established use reduces to a few neat logical rules which in combination are both sufficiently narrow and broad. In general, even if our ability to describe what we have learned grows out from spotting regularities in our practices, we still often learn more than we can describe.

I suspect that verbalization can come about as early as phase (B) as an articulation of experienced vague commonalities without being able to describe them besides pointing out the similarities between things or events.[118] In general, when the patterns become salient and differentiable through experience, they likely receive their own labels like concrete categories do, especially if they are associated with shared practices and regularities in one's language community (be that the whole culture or an exclusive professional community). Since schematic contents are not borne out of grasping universal analytical content but recurring concrete patterns, they can be quite idiosyncratic. Particularly at phase (B), two subjects may have quite a different grasp of what their linguistic expressions actually pick out. Later when the commonalities become more salient (at C), it may become easier to negotiate what they are. However, since the underlying competencies are mostly tacit, different subjects may have different understandings about the contents of their schemata. Conceivably this may result in different discursive practices and contents associated with the expressions used to communicate and reason about such schemata—especially within different social groups sharing different experiences. Moreover, since the implicit contents are retrieved on a contextual basis and the same linguis-

[118] Indeed, Pine & Messer (2003) found that verbalization does not mark the mature end state of domain learning but happens in very early stages.

tic expressions occur in widely different contexts, there should be contextual variability in the content associated with the expressions.

5.3 Theoretical concepts and discursive reasoning

But what is the status of purely schematic reasoning depicted in entry (D)? Based on our previous discussion, it is somewhat unclear whether there are any such capacities. But surely there seems to be. Formal concepts, by their very nature, are removed from specific contents, and there certainly are practicing logicians. Apart from mathematics, many concepts are theoretical, in the sense that we acquire an understanding of them through education and public reasoning rather than by direct interaction with their putative referents. Scientific concepts are one thing, but discursive schematic reasoning is a prevalent characteristic of common sense also. If you understand what "permission" and "obligation" mean, then you understand that (1) *if you get* PERMISSION *for something, it incurs* OBLIGATIONS *to some other people.* If that is not all transparent, it means that (2) *if you first need to satisfy a precondition C to get a* PERMISSION *to do P, then any authority enforcing such a rule is* OBLIGED *not to interfere with you doing P if you have satisfied C.* For example (3), if it is the case that you need to clean your room to go out and play, then your parents need to let you go out and play, given that you have cleaned your room. (4) Of course it is possible that you do not have a common understanding of what cleaning your room means, but then your parents can instruct you. (5) If the instructions are not understood, they can show you. If that does not help, then probably nothing does, save some actual practice.

So there are descending levels of abstraction from (1) to (5) that terminates in a Wittgensteinian observation whereby eventually descriptions need to be rooted in actual practice in to be understood. Generally, adults understand schematic descriptions at levels (1) and (2), but the expressions can be also explained by giving (mutually understood) examples, such as (3). Now, we have covered all this; however, the point is that these are examples of concepts that afford discursive inferences at the schematic levels (1) or (2) without necessarily reducing the discourse to lower levels, such as (3). For example, if you are OBLIGED to do A, you MUST BE PERMITTED to do so, because no authority can REQUIRE you to do something and simultaneously NOT ALLOW you to do that; or perhaps they can, but everyone understands that this would create a contradictory condition that violates the logic of bans and obligations.

In principle, such logic can be learned through discursive practices that allow one to understand how to use such concepts even without having lower-level empirical content associated with them. For example, deontic logic has been developed as a branch of general modal logic that contains systems with only a formal interpretation. Informal logic of deontic terms, of course, is mostly learned in everyday life; however, the capacity to do such schematic reasoning is probably a result of both (a) empirical knowledge extracted from the specific situations that the expressions refer to and (b) discursive practices of using those expressions in public reasoning and communication. In formal logics, the discursive element is presumably decisive. Similarly, mental contents and skills to use superordinate concepts such as "mammal" contain both empirical and discursive elements, and this is, I presume, the norm with human lexicalized concepts. What sort of knowledge a word use employs is both a matter of idiosyncratic experiences and how the use is related to discursive and material practices in the linguistic community at large.

I maintain that these two types of knowledge are actually products of the same type of pragmatic empirical learning. Discursive knowledge is just extracted in cultural praxis. It is hard to say what sets "natural" or "empirical" apart from "cultural" or "discursive". There is no metaphysical demarcation of these realms, and here I propose that neither are there are any substantial qualitative gap that strictly demarcates empirical and discursive cognitive contents.[119] Nevertheless, discursive and empirical or *de dicto* and *de re* knowledge associated with certain things, situations, or practices can still be exploited in different proportions and ways in various contexts, hence producing competences where (lay) theoretical and empirical content may seem to be dissociated.

Probably there are things such as colors that cannot be fully explained but they need to be experienced to have the same cognitive contents associated with such concepts, which most people have. However, this does not preclude (color) blind people from having a theoretical or discursive understanding of colors. Then there are things such as *transfinite cardinals* and *the cosmological constant* that cannot be directly experienced and we acquire knowledge about them only by theoretical means. However, this does not mean that the related understanding is purely intellectual comprehension of essences. Instead, theoretical contents are culturally produced, even if their referents are not always

[119] See our previous discussion about the impossibility to differentiate inferential and causally determined contents in Section 2.1.3, and similar points about entanglement of culture and nature in Sections 2.2 and 2.3.

cultural products, and grasping the contents is adaptive procedural know-how in discursive domains. This is an interactional (mostly tacit) skill engendered through immersion in the way of life of a linguistic community, be it the society at large or a group of specific experts.

5.3.1 Learning formal domains top-down

The hypothesis I offer is that learning theoretical contents can be explained by resorting to the framework of schema learning, discussed in the last section. Although mathematical and related formal concepts are probably the most abstract ones that one can think of, the idea most straightforward to explain in the context of formal conceptual domains precisely because they are so abstract that the above considerations about intermixing of different sources of knowledge do not complicate the explanation.

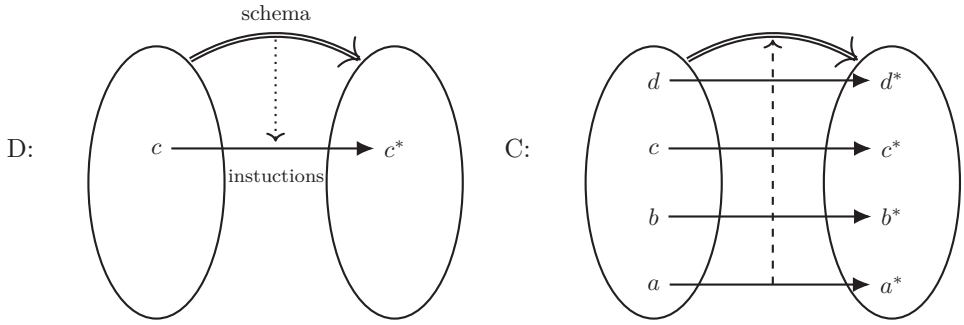
The basic idea is this: When we begin to learn the most simple basics of arithmetic, we count apples and oranges perhaps with fingers and learn how simple addition and terms referring to numerosity work. We are taught how to express quantities and simple arithmetic operations and their outcomes by symbols. We gradually learn to associate specific numbers and operations with certain outcomes (e.g., $2+3=5$). By using pen and a notebook, we learn how to do more complex calculations on paper. While the operations are formal and their correct applications are determined by mathematical facts, what we are primarily learning is procedural know-how to concretely work with external symbols. Presumably, a similar learning trajectory happens when we go to university and learn subjects like theoretical physics or formal logic. Students are given a set of syntactic inference rules (say, the Gentzen's deduction system) and instructions how to apply those rules to manipulate symbolic tokens. At first, students may consider the rules as rather arbitrary manipulations of senseless scribbles. They need not (and perhaps better not) think of the mathematical rationale of these rules, and that is the crucial point: For learning to proceed, students only need to treat the syntactic rules as *socially* permitted concrete operations in that context and be committed to participate in the associated social practice. As for the mathematical aspect of the symbol manipulations, the feedback whether we are doing it right comes from our teachers, tutors, peers, exercise sections of text-books, and in short from social and cultural sources that teach, define, and sanction those practices. In time the students learn the structural properties of the domain and become able to see the reason behind the rules and what is their *mathematical* point.

I assume that the major difference between elementary school and university level mathematics is that as educated adults we are initially equipped with more sophisticated conceptual machinery and skill sets. However, we are similarly introduced to new concepts, labeled by symbols and explained by examples and definitions. More importantly, we are given some principles and techniques as how to use those expressions in reasoning and solving exercises. Such expressions work as concrete variables that afford manipulations. The students can manipulate these tokens in any way they please; yet, not all the possible ways are sanctioned in the institutional setting where the learning takes place. They are embedded in situations where they need to conform to certain practices instituted by rules and enforced by epistemic authorities. The mathematics they already master when they enroll on mathematical curricula puts constraints (i.e., provides background skills) on subsequent learning but most of them are still learning completely novel skills, which are somewhat removed from their existing conceptual structures and knowledge. To learn the rules that constitute those domains, they need substantial practice with concrete problem instances.

Analogies can be used to make sense of some of these new concepts and principles; however, presumably many of them are very abstract and the procedures (like proof methods) strictly foreign. Hence, the procedures and concepts are initially quite rather of sense both psychologically and phenomenologically, and the students are quite unable to understand and apply them properly. The best way to teach them is to concretize them by providing specific examples of general methods and make the students reason through them. When they learn the same principles in various contexts and with various contents, they gradually learn to apply them correctly and more spontaneously to unfamiliar problems. In other words, they begin to see the relevant non-obvious patterns in the problems they are facing which they were unable to spot before and which begin to guide their problem-solving skills.

This is top-down skill learning where we are first introduced some general schematic principles, and through their concrete applications to specific problems, the stock of intuitive know-how about these principles builds up. As a result we learn *in a bottom-up fashion* abstract inference skills that track the intended use of those principles and thus begin to understand the intended meaning of abstract concepts, such as *structural induction*. So what is happening is a controlled build-up of an implicit exemplar-based knowledge structure

(C) under the guidance of explicit instructions concerning general abstract principle depicted in (D):[120]



If this is correct, then an understanding of such abstract principles depends on the implicit know-how of their use, and their understanding is fundamentally a gradually developing procedural know-how. Moreover, then the discussion in the previous section should apply to the learning of such abstract concepts and principles. The formal expressions get their meanings from how they are embedded in inferential practices constrained by schematic rules. This is basically an inferentialist or use theory of formal concepts closely in the spirit of conceptual role semantics but expressly confined to formal domains where such semantics is not very controversial. Moreover, my version of inferentialism does not share the tenet of CRS that the cognitive processes involved are logical computations over symbolic representations—*even if the resulting behavioral competencies are*. In time, at least simple operations are internalized and they can be processed without explicit rule-following and external symbol manipulations. These simple processes then can be simulated mentally. The philosophically relevant point is that simulation of formal calculus is, in fact, executing formal calculus. In this sense, people eventually manage to do logical symbol processing mentally. That capacity, however, is not fundamental characteristic of human cognition but derives from behavioral competences as a mental simulation of those skills.

Once simple operations are mastered and understood, they enable the learning of more complex procedures, which can be used to reason about other abstract principles, and this way domain understanding gradually develops. Not surprisingly, this learning process mirrors the typical organization of math-

[120] The dotted line signifies instructions as to how to apply the generic principle to a particular occasion. Solid lines depict actual application.

ematical textbooks and curricula. First, simple procedures are taught, then more complex theorems are derived, and these new conceptual tools are then used to derive even more general and complex results. This is how complex hierarchically organized skills are learned in general, and how concepts are both means and ends in abstract concept learning. However, mathematical practices often resemble more planning than direct execution of such skills, because not all the lower-level procedures need to be carried out when deriving results in higher levels of abstraction. When deriving complex theorems, the students and practicing mathematicians need only to mind goal-relevant, more or less complex mathematical results, given that they know that these results can be inferred by using more detailed and conceptually lower-level inferences if needed. When students gain more flexible and creative skills with the conceptual machinery, what was once an abstract and semantically opaque set of concepts and methods gradually comes to make sense in their own terms through conceptual development rooted in procedural know-how. The phenomenological and psychological character of abstractness should dissipate during this process similar to any conceptual domains that participate in our daily practices. With well-defined formal and other scientific concepts, we likely gather more explicit knowledge compared to everyday concepts because of the more pronounced role of explicit top-down learning. Thus, the explicit character of the resultant conceptual understanding is not because these concepts are intrinsically processed in System 2 but mostly because of the learning history.

If you happen to entertain yourself by reading popular science books on quantum physics, you inevitably encounter claims that no one really understands it. But surely the practicing physicists who make these claims do understand their trade better than, for example, I do. I gather that this oft-repeated statement means that the reality of fundamental physics is so remote from our commonsense understanding of physical mechanics that there is no natural mapping from that strange reality to our lay conceptions concerning physical objects and principles. Hence there is no "click of comprehension" that is produced by translating the odd nature of fundamental particles to existing intuitive world knowledge. This, however, does not preclude *all* sorts of (theoretical) understanding. The following quotation by Freeman Dyson (1958, 77–78 captures the course of conceptual development "things become to make sense in their own terms":

> I have observed in teaching quantum mechanics, and also in learning it, that students go through an experience similar to the one

> that Pupin describes. The student begins by learning the tricks of the trade. He learns how to make calculations in quantum mechanics and get the right answers, how to calculate the scattering of neutrons by protons and so forth. To learn the mathematics of the subject and to learn how to use it takes about six months. This is the first stage in learning quantum mechanics, and it is comparatively painless. The second stage comes when the student begins to worry because he does not understand what he has been doing. He worries because he has no clear physical picture in his head. He gets confused in trying to arrive at a physical explanation for each of the mathematical tricks he has been taught. He works very hard and gets discouraged because he does not seem to be able to think clearly. This second stage often lasts six months or longer. It is strenuous and unpleasant. Then, unexpectedly, the third stage begins. The student suddenly says to himself, "I understand quantum mechanics," or rather he says, "I understand now that there isn't anything to be understood." The difficulties which seemed so formidable have mysteriously vanished. What has happened is that he has learned to think directly and unconsciously in quantum-mechanical language. He is no longer trying to explain everything in terms of prequantum conceptions.

Note that I am not advocating a strong form of social constructivism concerning natural science—or even folk concepts and theories—as I, hopefully, made sufficiently clear in the latter half of Chapter 2. While our scientific concepts, theories, and beliefs are cultural products, the practices producing these beliefs and concepts are designed to be constrained by human independent empirical reality. While I stress discursive knowledge in using and understanding scientific concepts, my notion of "discursive" is not confined to language but to the whole praxis of giving and asking for reasons within a particular domain and social groups. This means that, for example, evaluating what qualifies as evidence for a particular claim is a discursive act, and so are considerations of how to produce that evidence and the actual practice of gathering it. While gathering evidence is often non-linguistic interaction with material reality (especially in natural sciences), the practice itself (*how* and *why* you do it) is socially constrained to serve private and public reasoning. This is a general point of discursive praxis: Certain facts (evidence) are reasons to say something in accordance with social settings that constitute reasons to make

certain statements in general (like doing science). The general reasons to say something (e.g., stating hypotheses) are, again, reasons to do something (gathering evidence), and specific social norms (e.g., methodological considerations) constrain how this activity should be done in accordance with the mutually understood point of the praxis.

Similarly, I am not selling psychologism about mathematics. My inclinations are somewhat more in line with the formalist program in that mathematical logic is like a game that starts by giving a set of rules about how to manipulate a determined set of tokens, and then we proceed to apply these rules to reach more or less clearly predetermined goals. However, I mean this only as a contrived characterization of mathematical praxis. Our ability to grasp mathematics is necessarily dependent on our cognitive make up, but the understanding eventually comes from our capacity to track certain social discursive practices, which, in turn, are *de facto* constrained by cultural praxis developed during the history of science, thought, and technology. However, I assume that mathematical practices track objectively true facts about quantities, set structures, logical entailment, etc. Such facts are useful in many human affairs, and therefore the practices of exploiting these facts and eventually the science of exploring these facts have been evolved. That is, as far as I can tell, mathematical facts are objective, and when we do mathematics (like arithmetic) we are following intersubjective norms constrained by those facts. Hence, mathematics does not reduce to psychology nor social science. Roughly, its autonomy is guaranteed because no consistent practice exploiting mathematical methods can violate these facts without losing its mathematical character. But this, again, is not necessarily a metaphysical fact about mathematics but a conceptual one, and by "conceptual," I mean that it is a constitutive character of the practice we call "mathematics."

So, again, my claim is that the ability to use and understand formal and other theoretical concepts results from tracking pragmatic regularities in our environment, which are instituted and constantly reproduced by material social practices. Broadly similar ideas, stating that abstract concept learning is cultural learning, have been proposed by Jerome Bruner (1990) and Lev Vygotsky (1986), for example. In the same vein, Murphy and Medin (1985, 309) suggest that when children learn a language they do not learn abstractions but cultural conventions, and this learning is guided and constrained by their world knowledge. Fundamentally, this is a learning of event–event causation in the context of discursive acts. The most natural settings for such learning is mutual reasoning and communication, and formal inference is a special case

of cultural learning in which the manipulation of inert symbolic tokens is emphasized rather than direct social interaction. Still, one can think of exercising such formal capacities as a quite typical case of situated reasoning. The agent faces a specific task that contains a set of variables instantiated by concrete tokens with their contextual affordances. The agent proceeds to manipulate those tokens by applying concrete procedures to transform the situation (or problem state) to another in order to eventually attain its goals.

Many social practices are constituted precisely this way in schools and workplaces. We are not bound by the causal properties of things we manipulate but by the normative constraints of what we are allowed and expected to do; and following or violating those constraints has predictable concrete consequences. The causal graphs in the previous section depicting event and variable contingencies in following regulations are simple examples, although the tasks are often much more complex. As neopragmatists claim, the cognitive contents associated with tokens participating in these practices emerge from the socially constituted procedural understanding of how these communal practices work.

No current research directly addresses if theoretical concept learning proceeds in the way proposed here. To my best knowledge, the studies by Larkin et al. (1980) and especially Chi et al. (1981) on novice and expert physics reasoning come closest to establishing that it is. One possible complication is that the tasks in these studies were introduced by diagrams, and it might be that experts used visually guided preprocessing in selecting the principles and equations to work with. Rich quasi-perceptual processing of problems has been implicated in expert physicists (Clement, 2004), and while these results support my general argument about the applicability of my account of practical situated cognition to scientific reasoning, they do not constitute a strong argument that it applies directly to explicit symbolic thought.

However, there are suggestive commonalities. In schema learning, there is an interaction between explicit and implicit processes. While analogy retrieval and schema induction can happen spontaneously, both can be highly facilitated by an explicit search for commonalities and especially by pointing them out and explaining how specific solutions can work as analogies for other problems. One might suspect that this leads to grasping and storing the common principle explicitly as a rule. But explaining a rule does not functionally lead to instant rule-based behavior: people consistently rely on exemplars even if they have simple and accurate rules available. Early in the learning process we are often quite unable to apply rules instead of specific exemplars in reasoning, and these exemplars also affect how the rule is later used (Ross &

Kennedy, 1990). The content and induction of superordinate categories depend on the amount and variation of the source exemplars (Gick & Holyoak, 1983). The shift of reliance from the surface to structural features is due to an increase in domain-specific knowledge, and this shift parallels the novice to expert development in a given domain (Reeves & Weisberg, 1994). These exemplar effects in later rule application are most evident when the task affords perceptually guided holistic processing (Allen & Brooks, 1991; Brooks et al., 1991); however, at least with novices, mathematical know-how is more strongly cued by task content than structure (Ross, 1984). Already Thorndike (1922) showed that slight changes in the syntax from familiar to unfamiliar in mathematical problems affect reasoning, thus demonstrating that "any disturbance whatsoever in the concrete particulars reasoned about will interfere with reasoning." Barsalou (1999, 606) has proposed the same idea that formal thinking is (quasi-)perceptual simulation of concrete manipulation of external symbolic tokens, and notes that logicians and scientists often construct visual simulations to discover and understand formalisms.

Considering the extended, interactive, and distributed nature of scientific cognition, a relevant strand of research comes from Nancy Nersessian (2008). Her work focuses on creative scientific reasoning and conceptual change and investigates how analogies, (perceptual) simulation, and thought experimens are employed in these processes. Nersessian stresses how scientific reasoning is situated and distributed among scientists and scientific instruments, methods, practices, and representations. Focal representational tools are diagrams and models, which do not convey only symbolic information but analogical and heuristic knowledge about the dynamics and structural properties of the target phenomenon. Diagrams depict the relevant feature of the target, and they can utilize visual devices such as proximity, arrows, brackets, and the like to signify the structural and causal relations of the system. Sketching and tinkering these external representations is an important part of the thinking process as well as gesturing over them. The production of external models is constrained by the subjects' theoretical understanding on the target domain, and the resultant visual features render important information salient for reasoning and afford insights that are hard to process and communicate with propositional statements. Indeed, according to Nersessian, diagrams scaffold simulations. They provide constraints and affordances for inferences and affect what is simulated and how simulations can be transformed.

Hence, external models do not only support cognitive processing but external and internal representations are genuinely coupled through interaction,

where external representations and other devices inform, constrain, and afford reasoning and communication. All this makes scientific thinking domain-specific and situated activity, which is supported by external material devices.[121] Nersessian also emphasizes that analogical reasoning is not a one-shot event that reveals a complete hypothesis of the target phenomena but one of the basic tools for scientists to reason about and improve their models. Lastly, we share the same insight that "propositional articulation at most codifies a reasoning process that already has been carried out by other means. That is, the mental model needs to have been constructed and simulated in order for the propositions that make up the argument to be identified" (Nersessian, 2008, 176).

That statement brings us back to formal symbolic reasoning and my claim that the associated learning is based on the same resources than other top-down discursive reasoning. I must clarify that this does not imply that the development of formal competencies track the development of commonsense concepts. Indeed, my main claim is that formal propositional calculus is not the right instrument to model human thinking, and formal logic is perhaps quite remote from—and maybe even antithetical to—most folk's daily activities. I have also argued in many places that different tasks use different kinds of information and cognitive faculties. If human conceptual understanding develops as a gradual adaptation to our situated pragmatic constraints, we should expect the *domain constraints* to dictate mature expert performance more than the general psychological learning dispositions—within the limits of cognitive constraints, of course. The basic point of my ecological approach to cognition dictates that expert mathematical competence comes eventually to track the structure and constitution of mathematical conceptual domain, and if everyday domains are constituted differently, the end result should look different accordingly.

Conversely, the impact of the general psychological constraints (e.g, the effect specific experiences to learning) should be most obvious in the early stages of learning. Expert cognitive skills such as science, and especially formal disciplines, are interesting for my purposes precisely because the concepts they harbor are often remote from commonsense concepts. Therefore, expert versus novice competence in science should be a good place to study how conceptual understanding develops in adult cognition with a relatively clean slate, even

[121] See especially Chapter 5 in Nersessian (2008). The reader is also advised to consult MacLeod (2016), which contains an excellent brief summary on these issues and especially how domain-specificity of scientific competence affects interdisciplinary research, considering that science is also a social project.

though the subsequent competence probably does not reflect that much the nature of everyday reasoning in other domains.

As for my stand on the cognitive status of elementary logical concepts, such as disjunction, conjunction, and implication, I should add that causal reasoning and representation with Bayesian framework presumes a capacity to process disjunctions, conjunctions, negation and forward inference. Moreover, language and social pragmatic learning presumes a capacity to follow rules and execute recursive procedures. Hence, the key capacities to understand and process simple and combinatorially complex (formal) concepts is presumed in my manner to specify the idea that we are natural born causal and social pragmatists.

In any event, I am not denying the relevance of (propositional) mental representations and formal-logical structures to human thought and especially to scientific reasoning. They are indispensable for scientific praxis. I only maintain that such representations, conceptual structures, and their cognitive contents are not innate but acquired through incremental learning. Symbolic and diagrammatic representations factor into our scientific practices, and their exploitation and internalization happens through interaction, which includes external (and derivately internal) manipulations of these representational devices, sometimes in the guidance of formal rules. Hence, I do not strive for an eliminativist stance to explain logical competencies and representational mental contents away but for a conservative explanation how they are psychologically constituted and how we can have such things in the first place, given that they are not innate.

I take that the understanding of mathematical and logical concepts through formal education develops similar to any other domain-specific cognitive skills and not as a distinct stage in human psychological development. My understanding is that when we employ these skills, we are often doing analogical inference as it is traditionally conceived: We map entities or propositions to symbolic expressions and use our skill to apply formal rules to such expressions to resolve the task we need. Moreover, these competencies can stem from multiple sources and component processes. For example, simple counting may be partly based on an innate ability to perceive and discern small ordinals, a linguistic trait to label them, and a trait to continue sequences (Carey, 2009). Counting with fingers may be cognitively a different ability than doing multiplications on paper, even if both are arithmetic practices. Indeed, arithmetics and other formal competencies are not monolithic skills but they can be learned in different ways in different contexts for specific purposes. These context-

and problem-specific mathematical competencies may not easily transfer to other domains but remain as integral part of the activity they were originally learned, bearing idiosyncratic characteristics of the specific practices and being dependent on the actual material settings where they are exercised.[122]

5.3.2 Informal theoretical and discursive concepts

Despite an obvious overlap, my notion of "theoretical concepts" in this section differs from the *concepts-organized-as-theories* idea in the theory–theory of categories, which holds that conceptual content is constituted by how categories take part in *explaining* events. As discussed earlier, this is a questionable way to formulate the knowledge account because such "explanations" or "theories" do not need to be explicit nor linguistic. A more accurate way to describe human conceptual content is that it is largely determined by how concepts enable us to understand and anticipate events and how they participate in action. Except for the section above, I have mostly avoided discussing language and controlled explicit reasoning; however, by "theoretical concepts" I basically mean linguistic concepts. More accurately, since there are no strictly linguistic and non-linguistic concepts, my treatment of theoretical concepts means theoretical *contents* where "theoretical" refers to discursive know-how, that is procedural knowledge of how we use linguistic expressions in public reasoning and communication. As discussed earlier, expressions such as "mammal" or even "dog" are psychologically associated with discursive knowledge regardless of how concrete their referents may be. This is supposed to be a trivial rather than a profound observation because, at the limit, one discursive use of words is to use them to communicate about, or even just name, specific things in the environment. In Section 2.2, I promoted the commonly held idea that in the development of language this is one of the most primordial discursive acts. I take no stand whether this implies that alarm calls of some animals, for example, should be considered as genuine discursive acts.

The previous section examined a mode of formal learning where symbolic expressions occupy the role of objects to be manipulated by explicit rules. The classical computationalist theory of the mind sees this as the standard model of language processing and thinking: Thinking is logical computation of symbolic

[122] See Lave (1988) for a cognitive anthropological study showing that people acquire arithmetic skills to manage their daily needs in supermarkets, which are distinct from the math skills learned in school and employ their own methods and heuristics; that is, the skills do not correlate strongly with formal mathematical ability and tend to disappear with isomorphic formal tasks. However, they can be (partly) retained in simulated grocery shopping activities.

representations, and language use is overt expression of these computations or their outputs. The principal function of the mental machinery was taken to be problem-solving and deductive inference of new information based on what is already known. In the heyday of the logical model the adaptive capacity of the human mind was attributed to its purported universal and formal nature. The idea was perhaps best captured in the famous *physical symbol system hypothesis* (Newell, 1980). The story was that since the mind is a universal problem-solver, you can throw a cognitive agent in any environment, and it will figure its ways out there. This capacity was supposed to make the difference between genuinely mental agents and simple automata executing trivial reflexes. The model of human cognition that I am proposing also takes the mind to be essentially an adaptive engine; however, the adaptation is a slow learning process, driven by our activity in the environment and constrained by our needs and capabilities as situated embodied agents. I do not aim for a wholesale rejection of the classical rationalist model but rather to put it in its proper place. Yes, we have reflective competencies, and we employ them to figure out what to do when we do not know. We also need explicit controlled cognition to learn various novel skills, such as playing games and doing mathematical logic. However, this is not how our minds guide our behavior most of the time. If we intend to understand what the human mind is and how it works—or how *we* generally work—it is more fruitful to study cognitive processes that are responsible for the etiology of our everyday activity instead of the interesting but rather weak and marginal reflective capacity.

So where does this leave linguistic reasoning and communication? With top-down learned theoretical competencies, such as formal logic, our linguistic expressions may track our actual psychological reasoning processes because we learn these skills by exercising our explicit linguistic and symbolic faculties. However, most of our skills are implicit and learned largely in a bottom-up fashion through interaction. We use our linguistic expressions to communicate the patterns in our environment (i.e., recurring entities, situations, and practices) rather than our (inaccessible) thought processes. Experts are ostensibly better at explaining their decisions; however, we know from expertise research that they are often unable to articulate them in a way that communicates the actual (tacit) reasons and processes underlying their judgment. They rather seem to be better at articulating the relevant aspects of task features in their area of expertise. That is, they gradually learn pragmatically important recurring patterns, and these overt patterns get lexicalized. They also learn how these patterns and the variables in the underlying concrete situations consti-

tuting those patterns hang together. Still, they are often unable to articulate their exact inaccessible thought processes about these matters.

Thus, the communicative function of linguistic expressions mostly describe aspects of recurring situations to peers who ideally share roughly the same experiences and the resultant tacit understanding of what is discussed. Recipients' own comprehension processes fill in the blanks. This tacit understanding that underlies successful communication does not need to be symmetric, however. In educational contexts the tutor tries to explain to the pupils what they should concentrate on and how they should employ the conceptual apparatus they are learning. But these remarks do not only pertain to educational contexts nor expert skills and communities. If common sense is sociocultural expertise, then in the same vein, communication and public reasoning with commonsense concepts rest on the shared understanding of how our everyday environment and social practices work.

In summary, in formal and other (scientific) domains, which are largely learned top-down, our explicit argumentative reasoning protocols track more or less accurately our thinking practices because these thinking practices are products of such protocols, but in complex and vague implicitly learned conceptual domains they often do not. What does this tell us about the nature of commonsense discursive reasoning, then? Hugo Mercier and Dan Sperber (2011; 2017) have proposed that argumentative reasoning has not been evolved to serve the function of improving knowledge or making better decisions. Instead, it has been developed for communicative purposes. The ability to devise arguments has evolved for the purpose to persuade others, and the ability to evaluate arguments has evolved to assess the quality of the information provided by others and to judge their status as reliable information sources. We are always vulnerable to misinformation, and if someone tells us something that we do not already believe, we better ask some justification for the claims—given that we choose not to trust that person blindly; which may be appropriate if the information source is an epistemic authority of some sort. By evaluating the reasons given to believe those facts, we are effectively evaluating the quality of the information. If the given justification is unsound, we have grounds to disregard the information and any considerations based on it. Even if the arguments are sensible, we may disregard the conclusion if we or other people participating in the discursive transaction have more compelling reasons to believe that the claim is not true.

Mercier and Sperber (2011) argue their theory largely by noting that if explains wide swaths of peculiarities observed in human reasoning. One general

phenomenon in this regard is that we reason a lot with our peers and seem to be quite good at it; however, when placed in a laboratory to reason by ourselves, poor judgment is often observed. This probably has to do with the fact that often in everyday contexts we reason about everyday matters we are experts of, but in the laboratory we are often tasked to solve unfamiliar problems. However, this is far from the whole story. For example, if the participants in the abstract selection solved the problem in groups of five to six members, an 70–80% level of correct answers were reached (Moshman & Geil, 1998; Mercier & Sperber, 2011, 63). An obvious explanation is that in large enough groups you happen to find someone who can solve the problem and the correct solution then spreads in the group. This is part of the explanation (Maciejovsky & Budescu, 2007); however, it cannot be the whole story because that is neither sufficient nor necessary for groups to reach the corrects solution.

In collaborative problem solving, groups tend to perform better (but may sometimes perform worse) than their best individuals, and this is be because the collaborative reasoning process is different from individual reasoning processes. When people are requested to explain and defend their judgments in the abstract selection tasks, the odd, confused, and irrelevant character of their justifications are easy to spot and challenge. Subjects are perfectly able to invent explanations for their behavior; however, they do not know the etiology of their intuitive judgment and when they invent reasons *post hoc*, their explanations do not always make much sense. While the logical interpretation of the selection task is not very salient, it is still available to virtually everyone as testified by the debriefing sessions (Wason & Shapiro, 1971). Through peer feedback and collaboration subjects spot problems in their reasoning and partial solutions, which facilitate insight into formal aspects of the task and enables them to converge to the logical solution (Moshman & Geil, 1998). According to Mercier and Sperber (2011), this is a general phenomenon. If one thinks of argumentation and reasoning from information processing perspective, the nature of that process is different between individuals in comparison to what happens inside the participants' heads. Our individual cognition tends to focus on one hypothesis with confirmatory tendencies, while the exchange between individuals employs primarily falsifying strategy toward multiple hypotheses.

Note that the logical interpretation does not necessitate knowledge about formal logic because the logical use is one of the ways that we use conditional statements in everyday language, even if it is not the most usual way. Arguably, the logical reading is the most reasonable interpretation of the selection task, once you think trough what the experimenter might have in mind. But people

do not think it through because when put to reason individually, we just follow our intuitive judgment that feels compelling and then confabulate unchallenged reasons for the judgment. Critical exchange of arguments puts the quality of those confabulations under review, and the joint reasoning tends to converge to the non-obvious but reasonable logical interpretation.[123]

Outside laboratory, the argumentative theory of Mercier and Sperber explains, for example, the notorious *confirmation bias*, that is our robust tendency to pay attention to information that supports our beliefs and hypotheses while disregarding the information that could disconfirm them. Curiously, while people are poor in spotting evidence that goes against their beliefs, they are still good at seeking conflicting evidence for conceptions that they do not personally hold. According to Mercier and Sperber, confirmation bias does not stem from cognitive limitations but from ecological functions for which the argumentative practice has evolved: Our aptitude to device arguments is evolved to persuade others to believe what we believe, and our ability to refute claims is evolved to check the credibility of the information we are fed. Their theory also explains the close cousin of confirmation bias, the *belief bias*, which is our tendency to evaluate the quality of arguments based not on their form but the credibility of conclusions. If you do not believe the conclusion, you are disposed to find weak links in the argument; if you already believe it, you often do not care how it is justified on specific occasions.

Once this practice of giving and asking for reasons is in place, it applies to explaining and arguing for actions and not only claims. I take that this is actually the default way in which argumentative reasoning is used, as we probably care about the quality of information mainly because we use that information for our pragmatic aims. However, this also has a consequence that discursive reasons can become practical reasons for actions. Communal discourse is used to persuade others not just to believe what we believe but also to do what we believe ought to be done. Once these discursive practices are internalized as our own reasons for our own actions, they begin to guide our behavior. People tend to make decisions based on what they find easiest to justify, even if they sacrifice economic utility (Mercier & Sperber, 2017, 255).

[123] Many subjects explain that they are trying to verify the rule (Wason & Evans, 1975). They are actually correct, if the Bayesian explanation of abstract selection task performance is true. The problem, however, is that the participant are unable to give a coherent explanation what their selections has to do with verification. Also, if they are doing Bayesian induction, there is no single correct answer (which is required by the task) and hence the logical interpretation arguably is the rational one.

As we have discussed earlier, discursive reasons are fundamentally rooted in patterns of social behavior. The tacit generalization of these patterns (presumably via analogical transfer) to quite different contexts may further explain some peculiarities conflicting the individualistic economically rational model of the human. For example, the sunk cost fallacy may be explained by the application of the norm "do not waste" in inappropriate places. Sometimes the norms themselves may not be (economically) very rational. For example, you might prefer dish A to B in a restaurant but pick B because your friend first orders A and you feel compelled to follow the maxim "don't pick the same things as others." People use all kinds of lay theories and principles to guide their decisions and sometimes against their apparent interest (Mercier & Sperber, 2011, 70-71).

Also relevant here is the Henrich et al.'s (2005) extensive cross-cultural investigation into how local norms and practices affect individual economical decisions. The study found that subjects' behavior in economical games (*ultimatum*, *public goods*, and *dictator* games) varied in a culture-specific manner and the variation was best explained as reflecting the patterns of interaction encountered in everyday life. "Selfishness axiom" was absent as a norm and the authors remarked that "humans are endowed with cultural learning capacities that allow us to acquire the beliefs and preferences appropriate for the local social environment; that is, human preferences are programmable and are often internalized [...] The preferences become part of the preference function that is maximized in preferences, beliefs, and constraint models. Norms such as "treat strangers equitably" thus become valued goals in themselves, and not simply because they lead to the attainment of other valued goals." (p. 813)

Why certain maxims end up as having normative force in specific societies need not concern us. The important point is that *they have* normative force which makes them pragmatically binding, or at least suggestive, to all parties committed to shared communal practices pertaining to rational action. Mercier and Sperber are unfortunately not very explicit about the cognitive processes involved in devising and evaluating arguments. They attribute these capacities to intuitive cognition, even if the argumentative practices are linguistic and explicit. This is similar to what I have promoted throughout this work; that is, understanding how things, situations, and facts relate to each other in relevant ways depends on a tacit pragmatic understanding of such matters. Producing arguments and explaining reasons is pointing out an approximate path of rationales that the audience needs to be able to follow themselves. This is the norm with all communication. Sentences and gestures generally have no de-

terminate meaning unless the recipient understands the context and can fill in the blanks to retrieve speaker's meaning (see Mercier & Sperber, 2017, 60–63). If these explanations are too vague and complex, they can be broken down and dissected in detail, but eventually mutual understanding needs to be grounded in a tacit know-how shared to a sufficient degree. This means that discursive content partly emerges from non-discursive pragmatic knowledge. Moreover, since discursive reasons are practical reasons to engage in specific policies of actions, discursive norms and acts also make us perform certain non-discursive acts. Hence in social reality—that is, the human reality—there is a two-way interaction, or more like an entanglement, of discursive and non-discursive praxis.

And what does the rehearsal of these points add to our understanding of non-formal theoretical and discursive contents? Remember the remarks a few pages back that expressions referring to abstractions like "permission" and "obligation" derive their contents partly from our pragmatic knowledge of specific situations and partly from their discursive uses. The point is that there often is a two-way interaction between the abstraction levels where our discursive practices with abstract superordinate concepts make us conceive the concrete, more specific situations and actions in new conceptual light.

For example, as children we learn from our parents that we should be fair to other people. What "being fair" really amounts to cannot be captured in simple rules. Through parental feedback in specific events, we learn when we have acted fair and when we have not. Explicit rules may be useful in highlighting which aspects of social interaction we need to pay attention to, but we mostly need to extract the vague sense of the label "fair," which is embedded in intricate pattern of behaviors. Once some understanding of this concept is established, we begin to enforce the norm to our peers. When we socialize with other children, they may have different understanding of what is fair and what is not. Conflicts ensue, but at least we have a shared *word* to negotiate what the culprit is, which helps to mutually adjust our behavior to meet and shape the shared standards. Later in life, we discover that fairness is not something that happens in just a playground but a complex social and political issue, which has a lot to do how we organize our society. Sometimes it is branded as "justice," "equity," "responsibility," and so on. We learn that many people have rather different ideas what it means, and we read books by philosophers and social scientist who have developed intricate theories of what it amounts to. By these conceptual devices we may end up finding new contexts where the concept is relevant and reconsidering that some of our behavior we

have considered as fair may not actually be that, and change our attitudes and behavior accordingly. This conceptual restructuring does not happen solely due to intellectual enlightenment, but also the acceptance and impact of these theories depend on our (internalized) norms, identity, and social practices that delineate who we are and what we take as just and reasonable. So there is a two-way interaction (or perhaps a complex system) between these theoretical revelations and our situated acquaintance with social norms and concepts like *fairness*, and this interaction of levels is made possible by lexicalizing these concepts and thus introducing them to symbolic discursive practices.[124]

So while the understanding of linguistic expressions grows from pragmatic know-how tied to concrete events, the practices of naming things and classifying concrete situations in our practical reality are not the only things we do with words. The discursive practices themselves are practical things we do with words with actual real-life consequences and impact to our intuitive understanding. This connects to our earlier discussion about the neopragmatist theory of linguistic contents in Section 2.2.2 and in particular to the idea that words participate in social interaction much like tools participate in the work done with them.

One can think of theoretical concepts, especially in the early phases of learning, as complex technology we operate attentively. First the focus is on the use of the technologies, and we employ our previous conceptual competencies in reasoning with them and making sense of their nature. In time, the technology fades into the background as we get the feel and sense of how they work and participate in our practices, and the focus shifts to the point of the praxis itself; which is the phenomenological status with our commonsense concepts. Many commonsense concepts reside in both theoretical and concrete realm; *fairness* in one, but notions like *society*, *democracy*, *meaning* have this

[124] Note that while saying is doing, our sayings and corresponding doings do not always happen in the same context. That is, the contexts where we discuss our norms are often different where we actually enact our norms. Since the underlying procedural knowledge and associated incentives in each case are activated on contextual basis, our explicitly endorsed norms do not automatically become enforced in practice, even if they have been internalized. What gets internalized is not norms in the abstract sense but acceptable, sanctioned behaviors; that is, what is a proper thing to say in one context and do in another. Although the sayings and doings can have identifiable common content (because the sayings are often interpreted against the background knowledge about doings) the tendency to say something (and mean it) in one context does not necessarily turn into a tendency to behave accordingly in another relevant context. Therefore, not every case of inconsistency between saying and doing is a case of the weakness of the will. The agent may be completely unaware of this inconsistency.

character also. According to Avner Baz (2016) Wittgenstein's *Philosophical Investigations* can be read as an attempt to show that with words like *know*, *mean*, and *understand*, the referential image of language is inept. These words have conditions of use but the conditions are products of our non-referential discursive practices. While the established use of these concepts may have some general characteristics, they are ill-defined, and explicit rules do not easily capture the adept use. Moreover, proper use tends to differ from context to context. Despite being commonsensical, these concepts are theoretical in the sense of being mostly discursive in content, and I believe that many expert conceptual domains are psychologically constituted similarly. Academic disciplines such as philosophy and theoretical corners of social sciences and humanities are largely constituted by discursive practices revolving around such ill-defined discursive concepts, and indeed much of the professional life in theoretical science is focused on refinement and rethinking our conceptual tools.

If the concepts are ill-defined, they cannot be wholly explicated by rules to guide top-down learning. So how are they learned, then? If my account of conceptual understanding is correct, the explanation is simply that by participating in discursive practices and tracking the use of such concepts in actual situations. Once students enroll in university, they start to catch new terms and reasoning heuristics (like "naturalistic fallacy"), which may be quite abstract and semantically opaque at first. Analogies can concretize the concepts; however, since the analogies tend to be imperfect, the mastery of such concepts needs eventually to be learned by acquaintance through their proper use. We all have some ideas what things like "society" or "knowledge" mean before we participate in philosophy or social science classes but the point of the curricula is to refine and transform these ideas. Textbooks have descriptions and definitions of these concepts; however, they are often inadequately simplistic, and we are in any case quite unable to use these definitions for competent reasoning before practical understanding has sufficiently developed.

Concerning complex theoretical ideas for which definitions and concrete analogies are of little use, I suspect that much learning follows the pattern of ritualistic or model learning, discussed in Section 2.2.2. Under this hypothesis, students first learn to replicate the discursive praxis that they catch from their professors, textbooks, peers, etc. This process is aided by pedagogical tools like analogies, instructions, and heuristics; however, these tools are not mere learning aids but part of the model to be learned, which are then passed to the next generation of students. Students' use of theoretical concepts are first heavily dependent on the actual examples that they have encountered.

They replicate this praxis in essays and discussions and get feedback if they are doing it right. In time, they learn to spot where certain expressions and discursive moves are in or off place. They track these practices as normative; in other words, they are committed to presume that there are proper ways to use the theoretical language, which has normative force over them if they wish to participate in the praxis as competent members. They try to adapt their practices to these norms, which they understand only vaguely at first. Gradually, their understanding gets more nuanced, and they begin to grasp the sense of the intricate patterns of regularities and contextual factors that govern proper use. Eventually, their discursive competence grows more flexible and withdraws from mere imitation. The discursive interaction becomes more contributory and less conformist. Students learn to characterize what they have learned more clearly, formulate arguments for novel ideas, evaluate arguments more reliably, and enforce the norms they have discovered on their peers.

This is the standard story of top-down skill development combined with the hypothesis that learning theoretical discursive skills is social model learning where the constraints of success are intersubjectively determined norms. This is due to the ecological, cognitive, and pragmatic constraints on how this learning takes place. The implication is that certain expressions first find their proper use in specific instances and when they are encountered and applied (correctly or incorrectly) in various contexts, the accuracy and generality of applications gradually improves, and the actual use becomes less dependent on specific models and becomes to track more generally sanctioned, intersubjectively instituted correct use. Still, the contents are products of and determined by the learning history just like any bottom-up learning. Therefore, learners at no point grasp the intellectual essence of these concepts (which the concepts generally do not have), and unfortunately their understanding of them can be contaminated by whatever biases, prejudices, inconsistencies, and other idiosyncrasies present in the learning environment—just like with pathological environments in skill learning in general.

It should be noted that students are not merely replicating the use of specific words. Instead, they are likely imitating the whole praxis itself. This socialization in a cultural niche also incorporates the adoption of values and learning what is important, relevant, (in)appropriate, and so on. As with other skill development, learning new values and ways to think should not necessarily change the previous conceptions; instead, we may learn new attitudes and means of using certain expressions in new contexts, and part of the flexibility

of expert reasoning is the ability to switch from one interpretation (i.e., use) to another at will and understand which interpretation is relevant and proper on which occasions. Hence, refined conceptual competences (e.g., about societal issues, learned in university curricula) do not necessarily replace previous lay theories. It is up to the expert to understand this difference and learn how to exploit and mix their conceptual competencies in different discursive contexts.

This explanation of inferential use of language is not tied to academic and other expert domains but applies also to commonsense discursive reasoning. Very practical commonsense discourses referring to specific concrete matters are one end of the continuum and explicit rule-governed formal practices perhaps the other. I find it difficult to articulate concisely and clearly the nature of the vague, complex, intermixed conceptual competencies in between, which philosophers are mostly interested in and cognitive psychologists practically less so. Avner Baz (2016) has extracted a similar notion of language from the works of Wittgenstein and Merleau-Ponty and empirical psychologists such as Tomasello. I am compelled to quote him at length (69–71):

> Merleau-Ponty, who also comes to compare words to tools, speaks in this connection of "a new meaning of the word 'meaning'," for the reason that here meaning is both perceptually and theoretically inseparable from what "has" it (1996, pp 146, 174). Accordingly, Merleau-Ponty proposes that we learn the meaning of a word in much the same way that we learn the use of a tool—that is, "by seeing it employed in the context of a certain situation" (1996, p. 425).
>
> An important difference between words and the sorts of work-things one may find in a toolbox is that what makes a work-thing—a hammer, say—fit for certain uses but not others is, for the most part, its physical properties, whereas what makes a word suitable for certain uses but not others—call it its "meaning"—is, for the most part, its history, or as Merleau-Ponty puts it, "previous acts of expression" (1996, p. 192). "It is for having been employed in different contexts that the word gradually takes on a meaning that is impossible to fix absolutely," he writes (1996, p. 408; see also p. 194).
>
> Thus, the basic unit of linguistic sense or intelligibility, for Merleau-Ponty as for Wittgenstein, is not the isolated word or the isolated combination of words, but the speech-act as performed by an indi-

> vidual speaker in some particular context—"the total speech act in the total speech situation," as Austin puts it (1999, p. 148).
>
> The open-ended plasticity of language—the fact that words may always, and in a sense must, be used or meant in new ways—means that understanding the speech of another is not a matter of simply putting together the already-fixed meanings of her words. Rather, it is a matter of seeing and aptly responding to the significance, or point, of her utterance—of being able to follow, and follow upon, her act of articulating and taking up a position in an inter-personal world, orienting herself by means of words. This last idea should not be understood psychologistically: seeing the point of an utterance is no more a matter of attributing some inner mental state to the utterer than seeing the point of a move in chess is a matter of such psychologizing (see Turnbull & Carpendale, 1999, p. 343).

Tomasello (2009) pushes these ideas as far as claiming that linguistic syntax is abstracted from the situated communicative patterns encountered in daily life; and Brandom (2000, 128):

> The correct use of these components is then to be understood as determining the correct use also of further combinations of them into novel sentences. The linguistic community determines the correct use of some sentences, and thereby of the words they involve, and so determines the correct use of the rest of the sentences that can be expressed by using those words.

As the reader might guess, I assume that individual reasoning abilities come about as these practices are internalized and simulated; that is, theoretical thinking is simulation of argumentative practices. Just like with formal concepts, the simulation of discursive reasoning *is* discursive reasoning, and importantly, individual reasoning inherits is the normative character from the social practices that it is simulating.[125] In general, explicit processing is ex-

[125] An empirical remark: This implies that abstract thinking recruits linguistic areas of the brain while processing concrete concepts employs sensory areas, as confirmed in a meta analysis (Wang et al., 2010). Note that mathematical reasoning activates areas independent of natural language processing and associated with space and numerosity. There seems to be also a specific sensory area dedicated to visual representation of numbers and equations, which is active during mathematical reasoning but only among expert mathematicians (Amalric & Dehaene, 2016). While mathematical ability can be independent of visual processing,

ploitation of patterns encountered in life, and early in the learning people are not thinking but doing; and this applies also to explicit rule use.

As Mercier and Sperber (2011) remark, to use this capacity in individual reasoning in order to improve knowledge and make better decisions, one needs to acknowledge their limitations and biases and to be able to distance themselves from their opinions to evaluate them critically. In effect, a sound theoretical inference should psychologically proceed like a debate: as a simulated exchange of arguments and counterarguments where the reasoner mentally shifts imagined positions with respect to the argument under development.

My main point is that theoretical reasoning is not abstract formal calculation but we should not be take strictly psychologizing attitude to understanding what people do when they are theorizing. This is because thinking is content-based and theoretical contents are products of cultural practices not of individual mental processes. To understand why individuals reason the way they do, we should take an ecological perspective and look at social norms and practices and their history. Similarly, serious (self-)critical thinking should involve the examination of the cultural and personal etiology of one's concepts, rationales, and norms. While in practice this is often too much to ask, it is crucial to understand that mere rigorous contemplation does not guarantee deep understanding of complex issues if one is not prepared to track rationales of competing viewpoints in good faith. This is because no one's reason is pure and free from the impact of the environment; and because the impact is largely tacit, the more self-evident a personal standpoint feels like, the harder it is to introspectively evaluate whether the reason for this felt givenness is the self-evident nature of the subject matter or one's limited and biased understanding.

When people say that they do not need to examine the sources of their intuitions because they are already thinking logically and objectively, free from the biasing impact of any doctrines, it is almost guaranteed that most of the time they are doing entirely something else. The fact that one cannot find one's biases and socially produced intuitions through serious reflection, no matter how hard one tries, is no guarantee that they are not there. The reluctance to accept that they *might* be there prevents finding them. During the social

rich visuospatial processing is implicated among visually unimpaired (Amalric et al., 2018). Unfortunately, these latter studies do not reveal the developmental trajectory of mathematical skills, but they do point out that the brain recruits a diverse set of resources for different sorts of tasks, and mathematical reasoning might be cognitively quite different from other forms of discursive reasoning. This is not that surprising for, after all, formal symbolic computation is quite different from informal discourse in natural language even while they do share commonalities.

exchange of ideas, biases more easily become manifest and corrected. However, that necessitates some divergence of beliefs between participants to prompt a critical look to spot the fault lines. Too like-minded groups may just end up consolidating shared biases. The process also necessitates taking different rationales, standpoints, and *people* seriously.

Lastly, one might wonder why in the first place we even have practices of discussing things like ”knowledge,” ”meaning,” the nature of ”good” and ”bad,” ”justice,” and ”democracy” if their contents are forged in communal discourse without obvious concrete referents outside mere conversation. This is not a question for psychologists to answer but cultural historians. As with our idiosyncratic implicit concepts, meanings of these shared discursive concepts are also in constant flux while they are transmitted, reinvented, and reinterpreted. The idea that the discussion of these matters in theoretical or philosophical tone is mere only and the discussion of the meanings of things like ”justice” and ”democracy” is mere semantics is seriously misleading. These discourses determine how we reason about such matters and hence shape the norms that affect how we think, talk, and make mutual decisions; critically, they also produce reasons for doing things in certain ways and doing some things at all. Fundamentally, these discourses are founded on a capacity to track and enforce cultural praxis, and together with material conditions of our practical reality, they shape our social order, institutions, and the whole way of life.

6 Conclusions and further reflections

In this work, I have examined a vast range of research in cognitive psychology, including the three traditional category theories, psychological essentialism, causal learning and inference, implicit learning and reasoning, skill learning and procedural knowledge, mental models and simulation, expertise research, conditional inference, analogical inference, and schema induction. I also discussed selected research in the fields of developmental and social psychology, cognitive linguistics, and cognitive neuroscience. It might have appeared that I attempted to forge a grand synthesis of existing research; however, that was not my aim. Rather, as I see it, these various research programs naturally intersect, and my attempt has been to survey what we find from the intersection. "Nothing in biology makes sense except in the light of evolution" was the name of the famous essay by Theodosius Dobzhansky (1973). I am not a biologist but I presume that actually a lot of things in biology make at least some sense without evolutionary considerations. Dobzhansky's point, of course, was that "[w]ithout that light [biology] becomes a pile of sundry facts—some of them interesting or curious but making no meaningful picture as a whole" (p. 129). Similarly, I find it difficult to form a meaningful big picture of cognitive psychology without taking a pragmatic and ecological perspective toward it. After that illumination, it still takes a good deal of work to see how exactly all the pieces fit together.

Certainly I do not assume that I got the picture, or even most fragments of it, completely right. Almost all the empirical results and their theoretical interpretations I have been promoting can be and have been reasonably contested within their respective research programs. Some fields and theories are currently under development (e.g., Bayesian causal induction), some are controversial (e.g., neoempiricist simulation theories), and some are pretty established with their canonical presentation in standard textbooks (e.g., similarity-based category theories). Nonetheless, even the research filed under the established theories are open to new interpretations. This can be especially fruitful when the reinterpretations are not novel competitive theories within the same research tradition but ideas informed by neighboring fields of study because that opens the door for theoretical integration. While integration is generally considered a virtue in its own right, it may also help in explaining inconsistencies in the data. For example, developments in the causal induction theories have enabled the integration of prototype and knowledge theories in the way that explains why we sometimes see phenomena like causal status effect and sometimes

we don't, and why sometimes raw similarity and sometimes category coherence dominates classification. Again, conceiving situation representations as causal models helps integrating analogical and causal reasoning, and so on. What I attempted to show is how component processes *might* hang together in human cognition once seen in the pragmatist light.

As a methodological principle, I have avoided introducing any novel concepts or hypotheses concerning human cognitive processes and representations. The reader should be wary that because of the breadth of the necessary research to compile this picture, I have not been able assess all competing standpoints and fair criticism on the specific empirical theories addressed in this work. While I have made attempts to avoid any cherry picking of evidence, this dissertation is not recommended to be read as a fair and balanced introduction to any of the research programs discussed. I have tried to put together fairly standard cognitive psychology in a way that conserves the spirit of embodied, enactive, and embedded paradigms—even if I have been largely using representationalistic and computationalistic terminology. One (admittedly fairly weak) justification for my specific selection of research is how they conflate naturally as a coherent whole under the pragmatist look on the human mind.

As for the choice of my terminology, although I find moderate scientific realism to exhibit good philosophical taste, I am also somewhat instrumentalist in what comes to scientific concepts. This attitude comes off quite naturally from my conception of concepts in general as being not (intrinsically) referential. For words to be useful, it is not necessary that they pick stable, well-defined classes but, *inter alia*, work as tools for public and private reasoning—and this is especially true of theoretical notions such as "mental representation" and "cognitive processing." If my widespread use of representationalistic terminology excludes me from the club of genuine enactivist, I am completely fine with that. I use that language to discuss standard cognitive psychology, which even radical theorists need to address somehow at some point. Even if the elimination of the remnants of representationalism is the long-term goal, it is probably a good idea to try to bridge these research programs in the meantime and see how it plays out. Enactivist approaches have been criticized for excessive radicalism and getting stuck in accounting for only lower cognition, such as perceptually guided action. I consider my work partly as an attempt to improve this sorry state of affairs by taking an approach sympathetic to enactivist and embodied theories, which I use to interpret the current research on higher cognition and investigate how far you can push the notion of (sim-

ulated) perceptually guided action to understand the higher faculties of the mind.

The critical undertones of this work target principally the classical logical computationalism and associated tenets in the analytical philosophy of mind and language—not computational approaches to the mind as such. The question as to whether the mind or the brain can be modeled as *some sort* of information or signal processing system (and what counts as such a system anyway) is mostly orthogonal to my argument. I have nothing against information processing models of cognition as such but (1) how mathematical logic was generalized as model of language, concepts, and reasoning in analytical philosophy and in particular (2) how these ideas were smuggled into cognitive psychology via the computer metaphor of the mind. Hopefully it is clear that under the view of the human cognition I am been promoting, the right way to understand cognitive processing—higher cognition included—is to look at the pragmatically oriented concrete interactions with the environment and their ecological constraints. Whether or not the dynamical systems approach gives the best tools to model these interactions is a practical matter for which I have no strong opinion. I am in league with the philosophers who maintain that intentional agency necessitates some form of rationality; however, I stress practical rather than formal rationality in this regard.

In practice, my standpoint has been fairly internalist. This is because I have attempting to make a philosophical point about the nature of human mind and intentional content, and hence I have been avoiding much of the discussion on how the specifics of the body and environment contribute to particular factual competencies. In that attempt, however, I have found the idea of cognitive processing as involving the interaction of the brain, body, and the world to be profoundly important. To understand the nature of conceptual cognition we should focus on skill learning but also culture and social interaction to account for normative and discursive reason. Hence, this work does not promote merely ecological but *social ecological* cognitive science.

6.1 Overview and some implications of the empirical argument

The main message of this work is that human conceptual system primarily tracks affordances and other properties that have pragmatic relevance to us as embodied and active organisms and that cognitive processes and representations are highly contextualized, pragmatic, action oriented, and shaped by our human practices, goals, and needs. The notions of *situation* and *action*

rather than *symbol* and *referent* are the keys to understanding human cognition. Concepts are *principally* devices for action, and *derivatively* building blocks of thoughts. In the recapitulation of the argument, I will use square brackets to refer to sections and subsections of this work.

In section 3.2. I began the empirical discussion with a review of the three standard category theories in cognitive psychology. The section is influenced by Edouard Machery's (2009) argument that these theories are not competitive but complementary. I showed that prototype [3.2.1] and knowledge representation [3.2.2] are not, generally speaking, independent but prototypes are actually structured by causal knowledge involving both relations between intrinsic properties of objects [4.1.2.] and how they participate in event causation [3.2.3; 4.1.1 4.2.3; 4.3.1]. At the limit, when no causal information is available, categorization presumably relies on linear combination of features (or on exemplars if the category has no coherent structure); in general, however, it is based on inference based on the causal model of the category [4.1.2]. In Section 4.1.1, I explained how affordances are important to categorization, and category content can even consist of specific movements. "Affordance" is the key notion here because affordances are object properties but they also determined by the situation and the agent and its goals. Hence, in a way, affordances bridge objects with internal and external factors of situations.

In Section 4.1.1, I also explained how category systems and category contents are dependent on the idiosyncratic experiences of the subjects. Category cuts are made on the basis of the property correlations and the causal structure of the environment. However, nothing in the environment as such forces us to carve its joints in any particular way. This is especially true of fineness of the grain of one's basic level categories in a given domain. Experts tend to partition things in their area of expertise in more detail that the rest of us. While ultimately any conceptual system is subjective and a product of idiosyncratic experience, concepts tend to be determined intersubjectively; they also depend on the culture in which one is immersed, be it an expert subculture such as bird watchers, aircraft engineers, or scientists or culture in the broader social and historical sense. This is because culture determines many of our practices and therefore (shared) experiences. For the same reason, people in different cultures and areas of expertise tend to settle to different category systems and with different kinds of information associated with specific categories. Presumably, experts associate more specific affordances and causal information with things that serve as superficial subordinate categories for non-experts. For example, while a regular urban resident sees a bunch of different looking

trees, landscapers and forest industry engineers see elms, oaks, pines, etc. with different practical purposes. For experts, goal-derived ideals may override the similarity or typicality-based categorization of novices (Lynch et al., 2000).

What even counts as a feature is something that needs to be learned, and hence there is no strict separation of perceptual and conceptual [3.2.3; 4.3; 4.1.2]. Here, I wish to highlight that category representations are not products of fixed perceptual modules but in category learning there is also concurrent learning in perceptual level, and features also need to be constructed often in parallel to category construction (see Murphy, 2002, 479).

Two points are worth remembering. First, many people reside in changing intersections of many different cultures and social groups, and hence ”culture” is only a heuristic notion here, characterized by a relatively stable set of norms and practices. Moreover, ”subculture” does not relate hierarchically to ”culture.” For example, one can be a Western European and member of a subculture of professional social scientists; however, not being a Western European obviously does not exclude anyone from the latter social group. Second, our conceptual system is affected by the community in which we are embedded, and we tend to absorb the category system of peers with which we interact. Even if we personally find no use for discerning, say, different genera of trees for our personal goal attainment, we still probably would learn to do that if the people around us find that taxonomy important. We tend to adapt to the basic level category system of our peers, if not for any other reason than to understand what the others around us are talking about and socialize with people we are affiliated with.

Sections 4.1 and 4.3 discussed event representations and how they relate to goals and objects and other specific variables. The latter section explained event comprehension as exploiting quasi-perceptual (causal) models either to execute or to simulate situated reasoning and action [4.3.1]. In between, Section 4.2 reviewed recent advances in Bayesian modeling of human causal knowledge and inference. Some important observations emerged from this discussion. First, causal induction is domain-general but not associative (in the traditional sense). It is guided by generic assumptions about causality (or by implicit learning dispositions that effectively work like assumption). The most important of these is that, in the first place, the world has relatively stable, albeit non-deterministic, causal structure which is responsible for most observed events, even when not all causes are directly observable. Moreover, by default, we seem to extract only coarse causal maps of our environment. Initial learning, in particular, is focused on strong causes that are relatively

sparse. For example, if an effect E results from strong and weak generative causes C_1 and C_2, we may disregard C_2 as one of the unspecified background causes that are responsible for the base rate of E. This obviously leads to less accurate causal maps; however, it may actually improve our ability to predict important events by constraining the complexity of world knowledge search and processing. Optimal accuracy/complexity trade-off is an interactive matter determined by the environment and the capabilities and needs of the agent. Minute outcome optimization by managing weakly causative variables is often irrelevant—unless gains accumulate through frequent repetition, for example; and therefore individual differences are likely. Domain experts presumably extract more detailed causal maps and feature sets to guide their judgment.

We also learned that active production and control over the data are highly useful and sometimes necessary for causal induction. After some knowledge is attained, it comes to guide how causal learning subsequently proceeds, allowing for rapid learning and narrowing down hypothesis spaces. In all, these observations show that humans are both behaviorally and cognitively active producers of causal knowledge and the nature of causal induction cannot be understood strictly by *a priori* analysis but pragmatic and ecological factors need to be accounted for. These sections make the case for the following item of the hypothesis about intuitive conceptual competencies:

a) Inductive (Bayesian) search for causally predictive and pragmatically relevant regularities in concrete situations:
   - Category representations are (prototype) feature clusters structures by causal relations between constituent features.
   - Basic ontology is formed around affordances and features relevant to event–event causation.
   - Basic situation representations are quasi-perceptual model type structures.

Section 3.2.2 discussed also a rather different view on human category representation. We saw that in certain studies categorization previously encountered exemplars instead of causal or summary information about categories. Caution was advised against interpreting these results as telling much about category structure. These exemplar effects are pronounced mostly in ill-structured categories, which are very hard to learn in comparison to most natural categories. They are incoherent and lack any natural sense. They have no point or use outside the very categorization tasks manufactured in the laboratory, and they do not connect to existing knowledge structures. All of these properties are antithetical to human categories in general.

Artificial stimuli are deliberately used to insulate categorization from existing background knowledge. The idea is that the possible distorting or confounding effects of knowledge can be filtered out this way, and therefore in experimental settings, we see pure categorization phenomena instead of effects of content-based inference, for example. However, since categorization actually utilizes category-knowledge and category-based inference, this experimental paradigm mostly dismisses the nature of conceptual representation, which it is set to investigate. Note that comparable issue arises in the abstract Wason selection task. The original abstract task was designed to reveal pure inference without confounding effects of category knowledge; however, because ecologically valid reasoning is content based, the experiment mostly misses the psychology of human deductive inference [5.1].

Nonetheless, there *are* ecologically valid studies that indicate the existence of compulsory and lasting exemplar processing in categorization. Visually guided dermatological diagnostics was discussed as an example. Without a doubt, medical diagnostics is a challenging skill. After Murphy (2002), Smith & Minda (2000), and Johansen & Palmeri (2002), I suggested that exemplar effects do not reflect to category structure but skill learning. The correct way to interpret exemplar categorization research is to think categorization responses in these studies as procedural knowledge and in terms of situated, perceptually guided action. The theme was further elaborated in Section 4.3.2. Indeed, much of the task set selection and switch research are virtually identical to exemplar categorization studies: Generally, in both research paradigms the task of the participants is to make a simple, specific response (e.g., pushing a correct button) to specific stimulus items in which the identifying features correlate arbitrarily. Allen and Brooks (1991) found that exemplar effects were clear with pictorial, easily integrated stimuli but tend to disappear when verbal feature lists are used. This further indicates that exemplar effects may results from perceptually cued access to implicit memory about instances—generally associated to skill learning, as discussed by Murphy (2002, 85–88).

This theme was further elaborated in Section 4.3.2. where the main message was that cognitive skill learning is mostly a memory phenomena with much of the processing happening in perception. In effect, intuitive decision-making and action selection are basically recognition. When a person engages in goal-directed action, the action leaves a memory trace associated with the outcome and the stimulus—or more precisely with the variables of the total stimulus that are perceived to constitute the situation. When the stimulus is reencountered, these SAO exemplars are automatically retrieved, and they

immediately produce subjective expectations of how things in that situation work.

In exemplar categorization research, it has been suggested that every encounter with a stimulus leaves a trace (Murphy, 2002, 58–60). In procedural knowledge research, "encounter" is interpreted in terms of actions taken (with restraining from action as perhaps a special case), so every time a stimulus is enacted, an *SAO*-exemplar gets recoded. These exemplars cluster into action sets that encode non-parametric statistical distributions of action-outcome probabilities. Action sets cluster to task sets that serve as units of procedural knowledge attached to specific tasks. In any given situation, the stimulus environment is first contextualized by selecting an appropriate task set that serves as a pragmatic interpretation of present variables in terms of goals, and then appropriate procedural knowledge (i.e., action set) is triggered by the contextualized stimuli. It is essential to understand how all the components of this process are situation dependent and not only stimulus dependent. This makes the contextualized task sets essential representational tools. The same specific outcome can be a failure in one situation and success in another. The same action can mean different things in different situations, and the same stimulus environment can mean different things and provide different affordances under different goals.

These task sets can initiate action directly and in highly learned tasks they often do. Hence, we can operate most of the time nearly on autopilot, although fundamentally task sets work as inputs to cognitive control. If they lead to unexpected outcomes or are deemed unreliable before action, control shifts to searching and constructing novel task sets from long-term memory. This is often associated with explicit reasoning, and if that fails, behavior lapses to inductive exploration. This dynamics was discussed further in Chapter 5 and particularly in Section 4.3.2, completing the next item of the working hypothesis:

b) Instance-specific encoding of event/action/outcome exemplars:
   - Context- and goal-dependent causal expectations are generated by associative (pattern matching) memory retrieval or analogical mapping from exemplar-based situation representations.
   - Without valid expectations, control shifts to exploration or explicit reasoning.

Now, let us pause for a moment here. It looks like we have found two kinds of fundamental knowledge structures; one seems to be quickly, established object-oriented knowledge with prototype and causal model structure and the

other appears action-oriented knowledge that exploits exemplars associated with actions and outcomes. Their properties are tabulated below:

Type A	Type B
Object oriented	Action oriented
Category prototypes	Event exemplars
Causal models	Procedural knowledge
(Causal structure?)	(Causal strength?)
Sparse structural detail	Rich surface detail
Quickly established	Slowly extracted
Predictive model	Task sets

This looks like a distinction between declarative and procedural knowledge, although that interpretation might be slightly misleading. At least Type A knowledge is not necessarily linguistic nor explicit (because neither prototype nor causal knowledge is), even though this type of knowledge might be, at least partly, more easily communicable than Type B knowledge. This is because compared to procedural knowledge causal structural knowledge is easier to transmit by declarative sentences. (Note that at least this tentative distinction does not seem to reflect the common characterizations of Systems 1 and 2, either.) It is uncertain to me whether causal structure and strength can be dissociated and, in turn, associated according to this sketchy sorting of Type A and B knowledge. But the reason to think so is that while categorization and category-based inference supported by structural causal knowledge is sensitive to qualitative causal strength [4.1.2], we are unable to use frequency information effectively and accurately without extracting it by actual acquaintance. People also form causal models from instructions and then try to map learning data onto this structure (Lagnado et al., 2007, 165). As stated on page 197, subjects often do not feel that the frequency information is meaningful and useful until they experimentally extract it and get a natural sense of the phenomena they are dealing with. Now, although I brought up the distinction, I am more interested in how these two purported knowledge structures are integrated rather than separated.

Active intervention is the most effective way to extract causal knowledge from the source, surpassing the effects of sheer observation or instructions. In fact, intervention may be the only way to find hidden variables and structures [4.2]. So even if A and B knowledge are separate, they are often learned concurrently in the same interactive settings. Since structural information is often not very easy to use without practice (i.e., Type A knowledge tends to be

useless without Type B knowledge), and because initial learning of even very simple general principles or event types (like how gravity and occlusion events work) are learned in a piecemeal fashion during specific concrete events [5.2.2], it seems a reasonable bet that experience-dependent and event exemplar-based Type B knowledge is more fundamental.

As a technical point, not all (or necessarily any) of the causal relations need to be represented in the long term-memory. They can be computed from contingency data when needed. Recall the end of Section 4.2.2 where I brought up the problem of learning weak causes in the long run without tracking all the possible relations between variables and at the same time keeping track of the order of encountered causal events. Here is an idea how this might happen: Assume that events leave memory traces that contain relatively rich but shallow information about events—basically a record of present entities or other variables, their properties or values, actions taken, and associated outcomes. These traces could be indexed and filtered by these properties. Such filtering would be similar to conditioning the data on the values of selected variables. If you have enough of this data, weak signals (i.e., value correlations) should eventually stand out from the noise, and causal networks then can be inferred from masses of such data. Some authors (e.g., Gopnik et al., 2004) concede that these kinds of algorithms may explain human causal induction; however, as Lu et al. (2008b) observe, such data-intensive computations are hardly called for to explain causal learning, which tends to be rapid. Nevertheless, data rich methods appear to be necessary to explain skill learning, which involves statistical clustering of *action*→outcome pairs, conditioned on stimulus or situations [4.3.2]. Now, the technical point I am after is that posterior density distributions of the possible causal relations between any given variables can be approximated from such data by using non-parametric methods—namely computing the density at each point by using nearest neighbor or kernel methods. Both are used in exemplar theories to compute the degree (or probability) of set inclusion of stimuli (Russell & Norvig, 2010, Sections 18.8 & 20.2).

The upside of non-parametric methods is that the learner does not need to pre-specify what input dimensions it is tracking, and weak cues can be extracted from the data at any point of learning. Something like this seems to happen spontaneously in skill learning because learners tend to become incrementally sensitive to subtle cues and configurations without awareness of this happening. However, because of a lack of space and my competence, I dwell no further on computational modeling details. Moreover, I have little knowledge about whether similar formal methods that Collins & Frank (2013),

for example, have used to model procedural knowledge can be used to explain causal learning data. But from the point of view of the process and knowledge integration in elementary learning, it is a suggesting point that similar methods that have been used to model procedural and exemplar (Type B) knowledge can be used to extract—or even directly implement—Type A causal knowledge. This is worthwhile to note especially because much of Type A causal learning depends on interactive settings where Type B knowledge is extracted.

Nonetheless, the above speculation hardly explains how we can readily grasp causal structures from instructions. Moreover, procedural knowledge is supposed to be very situation-specific, and sometimes we need to exploit causal knowledge in novel situations where Type B knowledge is not available. Hence, Type A and B knowledge may be separate. Then again, I have claimed that schematic knowledge and analogical transfer can do the trick here by exploiting existing Type B knowledge in these events. Nonetheless, my claim is *not* that all the core processes of conceptual cognition turn out to be actually identical, and therefore I am not compelled to show that A can be reduced to B knowledge. Quite the contrary: I aim to show that conceptual cognition is a set of functionally integrated capacities. To "have" a concept is not to have information ingrained in one's brain but to understand that information, or more appropriately to understand the environment one is inhabiting, which is fundamentally to act in relevant ways in one's practical reality and this capacity is *supported* by information we tend to associate with concepts.[126]

Type A knowledge supports procedural knowledge precisely in this way. Recall that procedural knowledge is constituted by *situation*→*action*→*outcome* exemplars. If an agent enters an unfamiliar situation for which it has no procedural scheme available, it needs to build and test a novel tasks set from the long-term memory. Type A knowledge supports task set construction by pro-

[126] Having a linguistic concept is slightly more complicated, as discussed in Section 5.3. Understanding linguistic expression is an ability to use it and interpret its use in linguistic practices (to convey information, signal values, opinion or status, device arguments, persuade others etc.). This includes understanding how words are used to refer to certain things, and generally, this necessitates previous non-linguistic understanding of the world and human practices. Concerning mastery of a foreign language, it is perhaps easier to grasp that all this requires practice and skill to learn proper *word–world knowledge* mappings, which can be aided by learning *foreign word—native word knowledge* mappings. The knowledge about (native) language and the world presumably develop concurrently. So *sometimes* we possess knowledge *of* a symbolic entity, which involves understanding how to translate the expression into world knowledge as well as to know how use it in communication. However, this is a limiting case, and understanding natural language necessitates a (non-linguistic) conceptual system that enables us to understand the world around us.

viding affordances and qualitative expectations of how things in a given novel situation interact. This species of knowledge enables us to grasp how things unfold in the absence of our interventions. Hence, one way to conceive object-oriented event–event causal knowledge is to take it as a limiting case of *SAO* knowledge when actions are not taken (hence, properly *SO* knowledge) and then procedural knowledge adds a layer of information about action outcomes. However, this is misleading. In Section 4.1, we saw that object identity is partly dependent on affordances as well as associated movements or motor programs. This means that Type *A* knowledge involves action information and shows again how Type *B* knowledge may directly contribute to category knowledge and category identity. This tight coupling of action and object knowledge is evident if the basic ontology and category cuts are incrementally produced by the agent in interaction with the environment as proposed in section 4.1.1.

Corollary 1: Considering the above discussion, it is safe to assert that **prototype, exemplar, and (causal) knowledge representation are tightly integrated in conceptual processing**. Although this contrasts with Edouard Machery's (2009) well-known claim that these three phenomena form distinct concept representation formats, we both agree that the notion of "concept" cannot be identified with any of these three information structures. Actually, in my account "concept" cannot be strictly identified with any purported representation in the brain for I take that the notion of having a concept should be understood as having a set of functional capacities that support intentional action, including derivately thinking. In any case causal knowledge and prototypes are not separate. Machery (2009, 71–74) and Machery and Seppälä (2009/2010) refer to linguistic evidence to show that similarity and knowledge representations are separate, but I think what they have (correctly) found is that theoretical linguistic knowledge may be separable from other kinds of category knowledge. I shall return to this later.

Note that since event identification and the resultant action selection are supported only by a restricted set of variables, events may be novel in many ways but still remain subjectively familiar if familiar goal-relevant variables are present. More importantly, similarity-based categorization allows category-based inferences with novel entities that share features with familiar things. An essential part of the task set selection process is first to classify an event with regard to a functional context and then focus on action relevant variables,

such as objects or object features [4.3.2]. An agent with sufficiently extensive pragmatic knowledge base should be able to construct causal models of novel situations quite flexibly to support task set construction—even though the automatized access to relevant knowledge is heavily constrained by surface similarity and the ability to spot relevant analogies [5.1,5.2]. Minor variations in routine events should not be even noticed.

If category representations are constituents of situation representations, and contextualized causal information is constitutive of category representations, the distinction becomes somewhat moot. An example of this is our ability to form reasonable goal-dependent and *ad hoc* categories on the fly [3.2.3]. Also, goal-neutral categories often have a graded structure, which is not always stable but varies from occasion to occasion (Barsalou, 1987), meaning that contexts modulate category representation. Even a casual observation verifies that affordances are represented differently in different situations. For example, chairs might appear as *something to stand on to change a light bulb* in one context and *something to hold a door open* in another. A *hammer* may appear as a *tool* or a *weapon* depending on a situation or even a handy *bottle opener*—if you know how to do it without breaking the bottle. Now, if affordances and other functional information are not encoded separately of categories but as constitutive of conceptual content, this cast serious doubts on whether there are invariant category representations in human cognitive system at all, even if *chairs*, for example, will be chairs in every occasion and retain their definite character of *something to sit on* in every context.

Recall that event types are generally conceived by a general gist in mind that delineates what is happening and why. The entities that are listed in taking part of events are often superordinate types, and superordinate categories are mainly characterized by generic functional features [4.1.1]. Every event token necessarily has some specific things filling these functional roles. How the object tokens are conceived depends on the specific situations because affordances, context selection, and object identity go hand in hand. Although this point is hypothetical, it makes sense that the event–event causal knowledge that is activated to support task set construction is similarly filtered according to functional context.

Corollary 2: Object knowledge is (partly) dependent on event knowledge, and object representations are active situated constructs. The point does in no way mean that there is no stable knowledge encoded in the long-term memory. Quite the contrary: The problem with

fluent on-line reasoning is that there is *too much* knowledge in long-term memory. The point instead is that the same exact content does not represent a category in every situation, and the question is then how these *ad hoc* structures are constructed and category information selected on the spot. Rosch (1999) and especially Barsalou (1987; 2003) share the same conviction with me that, psychologically speaking, categories are active constructs rather than static information structures. We have several ways to represent each category, and this flexibility results from selective information retrieval from long-term memory. I maintain that this mechanism is relevant to both abstraction and relevance assessment.

Also, especially Rosch (1999, 72) emphasizes that contexts or situations are the unit that categorization research needs to concentrate. Indeed, as mentioned above, category contents are intertwined with event knowledge, and in my mind, this connection is what actually makes category representations conceptual: Understanding how things interact and participate in different events is what conceptual content is about. Isolated category-specific knowledge does not convey natural intentional content. This is, of course, pretty much what knowledge account of concepts is about [3.2.3]. According to this view, prototypes and other summary category representations are simply clusters of statistical, causal, and other information. When put to use in an actual situation, that information serves functions such as categorization, induction, inference, and word meanings through selective activation of information contained in the cluster. That contextual activation turns the information into actual intentional content via its pragmatic relevance in that specific situation, and gives it meaning.

In section 4.3.1, I suggested that situations are represented as causal models and that thinking is simulation of internalized situated reasoning that draws on tacit conceptual processes to produce situation-relevant and action-related causal expectations [4.3.1]. This is supposed to be precisely the same mechanism that is responsible for cognitive skills as per the recognition-primed theory of decision-making and expertise [4.3.2]. Note that while these theories generally do not make much of an issue about the selection of functional context prior to action selection, it is a vital part of procedural knowledge, rendering it flexible even if it is anchored to specific concrete situations and stimuli. Context and action selection are guided by internal goals and external cues in encountered stimuli; however, this surface content dependency is relaxed when the agent learns higher-order regularities that allow situation and causal under-

standing to be based on event types and superordinate categories. Note that the corollary above pertains especially to superordinate classification, that is selecting the functionally determined category identity based on how the role fillers factor in the external situations and present goals. This ability makes behavior more flexible because it allows situated understanding to be based on schematic knowledge instead of strictly specific variables, and it allows seemingly different things to be incorporated into a shared pragmatic schemata. As discussed in Section 4.3, situations are represented and simulated in variable degrees of surface detail, and schematic abstraction enables hierarchical planning, understanding, and execution of tasks where subgoals are multiply realizable in lower, more detailed levels and therefore need not and cannot be always specifically addressed.

The close connection betwee abstraction and analogical reasoning is discussed in Section 5.2.2. Given the pragmatic view on human cognition, the proposed intimate relation of analogical reasoning and abstraction should not be surprising at all. If one wonders what the functions of schematic abstraction and analogical inference are, the answer is the same for both. Similar to schemata, analogical reasoning makes it possible to use existing functional knowledge to interpret novel problems and incorporate new entities into existing knowledge structures. Both work on a contextual basis. Superordinate classification of entities means treating them as alike in some sense, specified by their functional role in the current situation (as per Corollary 2), and some problems are similarly analogical in some sense, as determined by their functional structure under current goals [5.2.1]. Whether or not there is even a difference may be somewhat a matter of perspective. Think, for example, that you know substances x and y promote tumor growth, and both are antioxidants. You also know that z is an antioxidant and therefore hypothesize that it may also trigger tumor growth. Is this a superordinate category-based inference or an analogical inference? The difference presumably is whether you use superordinate category *antioxidant* to make this deduction or directly the causal knowledge associated with x or y. However, as the discussion in Sections 5.2.2 and 4.1.1 show, there is no fundamental difference because the contents of superordinate categories and schemata are extracted from and mostly reduce to the constitutive exemplars.

This example also shows how causal hypotheses can be generated by analogical reasoning.[127] In any case, discussion of these matters in Section 5.2 completes the following part of the empirical hypothesis:

c) Basic reasoning mechanisms are forward causal inference, simulation of situated action, and abstraction by exemplar-based analogical transfer:
   - Surface features are the primary retrieval cues.
   - Valid analogies bind different tasks under shared pragmatic schemata.

Corollary 3: Regarding information encoding and selection mechanisms, it should be emphasized that **category structure and use in addition to memory search and relevance assessment cannot be understood without a pragmatic and ecological perspective**. The constructive, pragmatic, and idiosyncratic nature of even concrete object concepts was discussed in Chapters 3 through 5, and just above, we saw how the same principles determine superordinate and schematic concepts. In Section 5.1 the same theme was discussed in connection with conditional inference. It is worth recalling that prototype theory presumes that useful feature correlations are found in the environment, which is manifest in how the mechanism breaks down with ill-structured artificial stimuli [3.2.2]. Similarly, surface feature-based exemplar search is presumably useful because surface content reliably guides toward familiar causal structures and affordances [4.3, 5.1, 5.2]. The ecological and cognitive rationale of sparse structural causal induction was discussed above and in Section [4.2.3]. None of the key features of these processes can be understood by contemplating their function in strictly formal register; however, they all make sense, given the cognitive limits of the human mind and the fact that these processes are evolved to make human goal attainment possible in a complex but relatively stable and structured environment where cues for relevant intentional action can be extracted from ambient information. These mechanisms exploit shallow, knowledge-intensive memory search that is poorly suited for formal inference

[127] Penn & Povinelli (2007, 111) briefly mention a similar hypothesis concerning the relations of abstraction and analogical and causal reasoning. They also make an interesting remark that there is no evidence of analogical reasoning in non-human animals apart from one famous unreplicated study with a trained chimpanzee (Gillan et al., 1981). Often the human ability to abstract thinking is attributed to language and controlled explicit reasoning. I agree with this in the case of theoretical concepts [5.3]; however, it is an intriguing observation that presumably uniquely human ability to schematic abstraction may actually stem from a largely unrelated, non-linguistic cognitive capacity to analogical transfer.

and requires considerable experience to be very effective in the wild. Its limitations can be somewhat mitigated by analogical inference and controlled top-down learning [5.3.1], as feeble as these capacities may be. Nonetheless, the real compensation for this trade-off comes from the very efficient real-time guidance of behavior in open environments and abruptly changing contexts after learning.

6.2 The epistemological import of pragmatist cognitive science

Now let us turn to the more philosophically loaded working hypothesis:

A.1. Expert and commonsense reasoning are both grounded in the same cognitive processes.

A.2. Conceptual understanding builds up as an adaptive cognitive skill. Its cognitive basis is in the intuitive system that gradually learns to exploit context- and goal-relevant regularities in the environment, especially the effects of our own actions in specific situations.

A.3. Often relevant environments are, at least partly, socially constructed. In the case of theoretical concepts, in particular, the relevant regularities are in large part inferential and other discursive commitments. Hence, abstract concept learning is a special case of functional/causal learning in (broadly) social or cultural contexts.

A.4. Initial competencies depend on concrete examples and their surface features or specific content. Extensive learning of procedural knowledge results in a gradual shift of focus from surface cues to structural features of the conceptual domain.

A.5. Through experience, the abstract/practical distinction dissipates both psychologically and phenomenologically.

A.6. Still, even later competence is affected by the specific learning history and content. Hence, the result is not an acquisition of general formal reasoning capacity but a gradual transformation in domain understanding.

A.7. Therefore, the process produces domain-specific conceptual competencies by a general adaptive and praxis-oriented learning mechanism. The resultant domains are products of our needs, goals, capabilities, learning environments, culture(s), etc.

At this point, we have covered everything except for A.3. As discussed above, A.1. follows basically from A.2. Cognitively, expertise is mainly an intuitive memory phenomenon supporting procedural know-how but socially, the notion of ”expert” refers to members of exclusive social groups acquainted with certain practices that are uncommon within the larger population. Note that then ”expert” and ”expert skills” are culturally determined social notions. Expertise is something that not everyone has. Some competencies may be

common in rural communities but strike as special skills among the modern urban folk, and vice versa. Often expertise is associated with higher education in specialized fields but, as I explained in the introduction (see footnote on page 22), being an expert does not even necessitate that one has a practically usable skill; only that other people think so and grant that status. Items A.4, A.6, and A.7. were discussed particularly in section 5.2.

Plank A.5. above is a somewhat hypothetical point; however, it is supported by the considerations about superordinate and schematic abstractions and their relation to concrete exemplars and procedural knowledge as well as how people acquire the intuitive sense of instructions, descriptions, and abstract notions through concretization and tacit practical learning [4.1.1, 4.3.2, 5.1, 5.2.2]. The plausibility of Items A.5. and A.1. is perhaps less evident in connection with theoretical concepts and scientific and related expertise. I think this partly stems from our cultural stories that depict deep scientific understanding as requiring extraordinary minds and mental feats. Partly the intuitions that resist Items A.5. and A.1. may come from dual-process and related scientific theories of reasoning, which place linguistic theoretical thinking in its own cognitive realm. I think both of these planks are, by and large, false but it hinges on the plausibility of Item A.3. whereby theoretical concept learning is a special case of functional learning in social or cultural contexts.

I began to unravel this claim in Section 5.1. where I argued that the ability to cope with logical reasoning is a function of expertise with formal logic. While that by itself is banal and obvious, it is important that it also applies to very elementary reasoning tasks. When confronted with a formal problem, we do not necessarily switch to a formal mode of thinking but recruit whatever knowledge we have to contextualize and interpret the task. If this fails, we may resort to exploration rather than simplifying heuristics. People do think in the selection task; however, thinking is mostly associated with rationalizing their decisions rather than engaging in logical inference.[128] The problem the subjects have is not with the ability to understand and reason with logical conditionals (therefore the idea that people use *simplifying* heuristics in this task is likely wrong) but exploiting that knowledge.

These observations suggest that explicit reasoning is not inherently formal or logical. Although controlled explicit thought makes flexible rule-following

[128] Remember that this sort of rationalization is an integral part of discursive reasoning because rationalization fulfills the social coordinative function of argumentation [5.3.2]. Unfortunately, it does not help much when the participants are reasoning alone and have no idea what they should be doing.

possible [5.3.1], I doubt that the notion of "rule" characterizes explicit thought meaningfully. Not all explicit reasoning is rule-following, and we can learn to follow rules implicitly. Rather, rule following characterizes some *tasks* and behavior that can be supported by both implicit and explicit thinking. Generally, tasks are executed by context- and content-sensitive search for implicit procedural knowledge, and already in the introduction, I pointed out that at least novices do not switch to formal or mathematical mode of reasoning when they encounter mathematical problems nor do they extract the formal structure of the tasks but haphazardly resort to whatever procedural knowledge that they have previously learned with related problem contents. This reflects normal skill learning [4.3.2], and it is also consistent with linguistic pragmatism in that linguistically mediated concepts are interpreted by resorting to knowledge associated with actual practices [2.2.2].

I do not rehearse the argument here, but Section 5.3.1 contains the suggestion of how theoretical and particularly formal reasoning is learned as manipulation of external symbols. This manipulation is guided by explicit instructions and learned in a top-down manner. Concurrently, there is bottom-up schema learning going on whereby the agents internalizes the explicit instructions that they follows in specific situations. Through this process, agents also extract more abstract procedural knowledge that gradually tracks the intended meaning of the abstract principles involved (such as proof procedures), and gradually they also learn to make strategic decisions and select relevant tokens and methods for the task at hand, reflecting general implicit skill learning. The constraints of success and correctness of procedures are culturally determined; however, the cultural determination may be further constrained by human independent facts. This hypothesis is a mixture of ideas drawn from the top-down theory of skill learning, dual-process theories, extended cognition, and linguistic pragmatism fused with the theory of schematic abstraction proposed in Section 5.2.

The main difference between common sense and specialized expertise basically reduces to a psychologically irrelevant fact that the former is shared with every member of society while concepts that comprise the latter are mostly shared among a minority who have gone through similar specialized training; although formal training often results in more explicit and regimented conceptual knowledge than with more casual commonsense notions. Throughout, I have been stressing the importance of actual practice and hands-on experience in the development of expertise and understanding. One central point has been that we are quite unable to use domain-general abstract principles spon-

taneously in intuitive processing and that we mostly understand the things with which we interact through intuition. When we grasp new conceptual domains that are remote from our everyday practice, the concepts lack intuitive content and may appear initially as arbitrary, abstract, and formal. This is, I guess, why many scientific concepts are often thought to be abstract; they rarely inhabit a place in the nexus of common everyday activities. However, when thinking, writing, discussing, and otherwise using abstract concepts become a routine activity, we slowly internalize their meaning, which is their appropriate use in public reasoning and communication. During that process, the concepts gradually cease to be alien and abstract in the psychological and phenomenological sense. To put it bluntly, for the practicing scientist, monads and transfinite inductions are embedded in their daily practices essentially in the same way as any routine affairs at large within the common folk.

The proposal is, of course, hypothetical, and my argument mostly aims to show how formal or theoretical concept learning is possible at all in the pragmatic situated reasoning framework. It also demonstrates suggestive parallels between formal concept, schematic concept, and skill learning, which justifies the claim A.5. that the contrived feel of abstract theoretical concepts comes from the lack of intuitive procedural interpretation of the concepts. However, once the competence is achieved, the subjects get the intuitive sense of the associated concepts like in implicit learning tasks where the subjects report to gain the intuitive feel of artificial sequence stimuli after exposure to task structure through interaction.

Relations to dual-process theories: Recall from the introduction the B.1. list of attributes that I was ready to accept as the characteristics of intuitive and reflective cognition.

Intuitive system	**Reflective system**
independent of working memory	*dependent on executive function*
high capacity/low effort, fast	low capacity/high effort, slow
associative	systematic
rigid	flexible
parallel	serial
implicit	explicit
automatic	*controlled*
situated	detached
default process	*inhibitory*

At no point have I really justified the demarcation of implicit and explicit systems or that they can be characterized precisely in this way. I will not do that here, either. In some sections (e.g., 4.3.2), I expressed my support for the default-interventionist dual-process theory of reasoning (see Evans & Stanovich, 2013). The only place I definitely needed the distinction of explicit and implicit processing is with the discussion of controlled top-down learning. What is important is what is *missing* from the list above, which is any references to the idea that the intuitive cognition is an engine of quick and dirty pragmatic heuristics while the explicit system is a cultured, normative, and linguistic symbolic processor. In fact, it is the explicit system that often exploits simplifying heuristics, for example, when we are learning complex new traits by following instructions and rules of thumb.

The italicized entries above are what Evans and Stanovich (2013) take to be the defining features of Systems 1 and 2, and the rest are merely correlated properties. I concur with their analysis that basically explicit system reduces to the executive function consisting of voluntary control and working memory, which enable cognitive decoupling from the current stimulus environment, mental simulation, and hypothetical and counterfactual thinking. I take these three items to be essentially the same thing. In contrast to the traditional formulations of the dual-process theories, I believe that the contents and competencies of explicit thought are mostly intuitive capacities. However, as feeble as the explicit system is, it allows us to learn arbitrary situation-action mappings, which can be internalized as implicit skills, and the resultant capacities can be clustered to support complex, hierarchically organized behavior, planning, and decision-making. Hence, the variability of behavioral and cognitive capacities that the explicit system enables is hugely significant even if it takes time to learn to utilize those feats effectively through automatization. I will leave it to the reader to decide whether I am advocating a version or an antithesis of dual-process theories.

The final section (5.3.2) is perhaps the most chaotically organized part of the text. It is an attempt to discuss commonsense and scientific concepts that are schematic and partly theoretical but defy clear definitions. They are often constituted both taxonomically and theoretically and operate approximately on the level of superordinate categories. Why this sort of concepts defy clear presentation is because many overlapping kinds of knowledge constituting them can be learned by many routes (i.e., by cultural learning, observation, or manual experience), and the relative proportion and the relevance of these factors depend on idiosyncratic and cultural matters. Direct interaction with

category instances provide different, often at least more accurate and rich, conceptual representation in contrast to learning declarative facts. Obviously, this also pertains to event knowledge. For example, you can hear from a friend that pouring water into an oil fire is a bad idea. Once you see that happen, you probably get a better idea of the expected results, and unwisely trying this yourself will most likely provide even more profound knowledge of what happens and how to handle that sort of situations. Generally, the more our interactive pragmatic knowledge of concrete things and events, the deeper our insight of them. Something of this sort is also familiar with formal and theoretical concepts. If you read through a calculus text-book, you either need a spectacular talent or good mathematical background to immediately apply the declarative knowledge contained in the book without gaining procedural knowledge by doing the exercises.

This does not concern only the quantity of information that different sources provide but also a qualitative difference. Many of the concepts that we possess have both discursive content and content derived directly with interacting with category instances. I know that water is H_2O, and this is strictly due to cultural learning. However, almost all I know about the macroscopic properties of water is derived from manual experience. I have direct acquaintance with mammals; however, my notion of "mammal" is, I think, constituted by discursive knowledge, and so on. I have relied on the following tentative typology of concepts and conceptual knowledge: 1° *Object and other concrete concept representations* that allow **taxonomic abstractions**; 2° **schematic abstractions** that are based on concrete *situation representations* and allow pragmatic knowledge to generalize over specific events; and 3° **theoretical concepts** that are produced by learning *discursive practices* in public and private reasoning. This above list is not meant to represent a fundamental or exhaustive typology of concepts nor a complete list of what all things "abstract" might mean but merely as a heuristic description of key components in human conceptual learning and how they relate to different kinds of intentional content.

Corollary 2 states that 1° and 2° are not strictly separated. Clearly, Types 2° and 3° are also mixed since the crux of Type 3° learning is to learn discursive skills, and the theory proposed here claims that all skill learning is based on pragmatic knowledge which, in turn, is engendered by accumulating experience of concrete situations where the concepts are applied. Type 3° concept learning is basically supposed to add a layer of social and linguistic learning, and sometimes it requires controlled and explicit top-down processes such as rule

following. Generally, utterances are acts that we use to make things happen in social settings. This requires considerable knowledge about inferential structure of the context, (shared) word use, what is permissible and relevant both in a social and logical sense, host of epistemological, psychological and perhaps sociological assumptions, and so on. Most of this knowledge is tacit, and these complex discursive interactions are paradigm cases of Type 2° pragmatic knowledge deployment. This was the main subject matter of Section 5.3.2, which consists of a hypothetical application of linguistic pragmatism [2.2.2] on top of the pragmatic reasoning framework articulated earlier. Note that the discussion was not intended as a detailed model of the psychology of linguistic cognition but rather as an empirically informed hypothesis about how human discursive reasoning could be understood in the pragmatic situated reasoning framework as a product of tracking social praxis and in the way that is also consistent with a credible philosophical analysis of human reasoning and intentional content [2.2, 2.3].

In Corollary 1, I mentioned that Machery and Seppälä (2009/2010) have argued that the human conceptual system harbors different category representations for similarity-based an theoretical knowledge. They insist that the necessary condition for these knowledge structures to be aspects of the same concept (and not different concepts), is not only that they are linked but they should also be coordinated in the sense that they do not produce conflicting category judgments. In their experiments the participants agreed, for example, that "in a sense, tomatoes are vegetables" but also that "in a sense, tomatoes are not vegetables." As one might suspect, their subjects explained that botanically tomatoes are fruits (knowledge based classification), but in culinary contexts they are considered as vegetables (similarity based classification). However, I have claimed that conceptual representations, and especially superordinate classes such as "vegetable," are not static information structures but constructed from memory on a contextual basis. Hence, I am inclined to interpret the above result as showing that under different interpretations of the claims about tomatoes, people exploit different information that they have associated with tomatoes in different learning context. "Interpretation" means selecting a pragmatic context that, in turn, activates relevant information to support decision-making concerning specific things in that context.[129]

[129] Note that the subjects initially change the interpretation of *vegetable* and not that of *tomato* on the contextual basis; however, this leads them to conceive tomatoes differently in each contexts; that is, either by activating theoretical botanical or practical culinary information.

As Machery and Seppälä acknowledge, the difference between their heterogeneity hypothesis and integrative accounts like mine, may eventually be terminological. I agree, and we all agree on the empirical fact that they have demonstrated the existence of conceptually relevant and dissociable information structures associated with lexicalized items. However, I think we have a relevant disagreement on what having a concept means, which is not merely a trivial matter of words. For them, having a concept is having an information structure; for me, it principally means having a capacity for intentional action.

These two ways of conceiving concepts are surely compatible, and having a capacity to intentional action presumably often implies having an information structure (at least in some sense). Still, I stress my position because the notion of "concept" plays various epistemic roles in science and philosophy. It is an indispensable *explanans* in social theory and subfields of psychology other than cognitive as well as an important framework concept to coordinate different research programs (see Pöyhönen, 2013). Because "concept" does not refer to a univocal cognitive kind, it may be helpful to identify it with something else to serve these other epistemic purposes; and I think it is both possible and productive to identify it with a complex interlocking set of capacities that support intentional action. The characterization is sufficiently informative and conservative to serve both as a *explanans* in various research fields and as an *explanandum* in cognitive psychology. For many of these epistemic roles I find it important to understand how concepts in psychological sense are unstable active constructs, and how and why they are modulated by contexts. Conceiving different information structures as separate concept representations, instead of component parts of concepts, may lead one to think that we always activate one of these alternative representations on the contextual basis. This may mask the intricate dynamics of situated content construction and in particular the fact that in many contexts the active construction process draws simultaneously and selectively on several of these information pools.

In any case, now that we hopefully agree on A.3., the epistemological version of the working hypothesis below needs no further specification:

C.1. The human conceptual system tracks affordances and other causal properties that have pragmatic relevance to us as embodied and active organisms.

C.2. Intentional content is fundamentally procedural competence to exploit the resulting know-how of what things do and what can be done with them.

C.3. Expertise in abstract expert domains (e.g. science) can be thought of as extended common sense, and common sense can be thought of as a specialized learned skill.

C.4. Cognitive contents of theoretical and formal concepts are constructed by the agent through social interaction. While grounded in the same capacities, discursive and concrete concept learning often differ qualitatively: Discursive conceptual domains track explicit and implicit communal conventions; this makes content intersubjective and normative.

C.5. Two people share the same intuitive understanding as far as they share the same practices, goals, needs, capacities, discursive commitments, environmental demands, affordances, and culture, or, in brief, the same practical reality.

Conceptual understanding as a skill: I presume that the notion of conceptual understanding as a skill is not very controversial in the case of mathematics and related complex domains, which are constituted by well defined inference rules. How we subjectively understand these domains is determined by our ability to exploit their conceptual apparatus and inference methods in goal attainment. Our understanding may be wrong or partial, depending on how accurately our competence tracks the intended use of these concepts. The normative standards are social products, while they may be constrained by non-social factors. Nevertheless, our subjective understanding goes as far as our skills to cope with such conceptual systems. The same story applies to ill-defined and ambiguous discursive notions such as ”society” and ”belief.” A crucial part of fluent know-how is the ability to interpret the context to guide the interpretation of situational factors—be they verbal expressions, reasoning heuristics, or concrete things such as tools—and assess the relevance of affordances they provide for the task at hand. So like with actual tools, in discursive contexts we need to be able to evaluate what conceptual tools are relevant and how they are used in that context to reach a meaningful outcome, such as expressing a contextually reasonable point.

What comes to schematic and superordinate concepts, they are similarly incremental products of our knowledge about how concrete situations and things embedded in them work. As Barsalou (2003, 1177) notes, ”abstraction is the skill to construct temporary online interpretations of a category’s members. Although an infinite number of abstractions are possible, attractors develop for habitual approaches to interpretation.” Again, this is a key part of procedural knowledge associated with context interpretation as explained in Section 4.3.2. It is also closely related to analogical reasoning, which consist largely of skill transfer to novel situations. Indeed, the critical cognitive components in analogical transfer are the same as in recognition-primed decision-making.

I suppose that intuitions may resist the notion of understanding as a skill when it comes to concrete basic level objects, like cats and hammers. Now,

inferences concerning concrete things is an exercise of a skill (and constitutive of category content), medical diagnostics of specific diseases is a skill, chicken sexing is a skill, recognizing tree genera and edible mushrooms are skills (even *expert* skills among moder urban folk), superordinate classification (such as selecting on a contextual basis whether a tomato is or is not a vegetable) is a skill; so it would perhaps be an odd discontinuity if classifying things like cats or hammers and understanding what kind of things they are would be something entirely different. The fact that such feats are automatic and easy and virtually everyone can do them does not mean that they are not skills. Common sense is certainly a skill as well as expert reasoning. Note that mere categorization rarely equates understanding. Phenomenologically speaking, we understand what an object is by being able to take different perspectives and stances towards it; and cognitively this is supported by the automatic activation of situated knowledge concerning what the thing does, how it is used and examined, what it affords, and how it factors in human practices.

This capacity to intuitive perspective taking constitutes something similar to the horizon of the *noematic Sinn* in Husserl's theory of meaning, and definitely, it is the key to practical reasoning and know-how. Getting an intuitive sense of inference supporting knowledge of factors like base rates and rules requires practical acquaintance and the same holds for the construction of the subjective ontology of the world. Grasping the structure of a dynamic phenomena requires implicit procedural know-how and not only declarative knowledge, and so does the grasping of reality in general. In a sense, one can think of categories as variables and the world as a very complex and open dynamical system. Similar to discursive skills, understanding our environment is not a binary matter but determined by our ability to grasp what is happening and how the present variables interact with each other and with ourself.

What I specifically want from the notion of understanding as a skill is to explain what understanding is, how it relates to language and non-linguistic cognition, where meaning, sense, or intentional content comes from and highlight that in order to understand cognitive psychology of concepts, we should conceptualize the notion as both a constitutive part of and something constituted by the core cognitive processes responsible for expertise, cognitive skills, and procedural knowledge. Workings of higher cognition are still somewhat a mystery, and here I try to formulate conceptual tools to reconsider what sort of a thing the mind fundamentally is to guide the framing of hypotheses and questions about the constitution of the higher cognition in a productive way.

If it is still unclear what this could mean in explaining the constitution of complex concepts consider ”climate change.” Its analytical definition is presumably ”a global long-term change in average weather and weather patterns.” If one needs to further unpack this, we all know what weather is based on our casual experience. Through our lived experience we also know that sometimes seasons are untypical and the implication is that the untypical ones will become the norm. In practice this may mean quite different things to city dwellers and farmers and it may carry different connotations to different individual in different places of the world. The context where we need to unpack the meanings of ”weather” and ”average patterns” may also dictate how they are construed in that situation. If the context (or our personal understanding of the climate change) is foremost academic, we may think of weather not as through our personal embodied experience but as a set of quantified parameters, such as temperature, humidity, etc., and construe ”average” in terms of how we have learned to calculate averages in elementary school math class. Some of us with more mathematical tutoring may further notice that it is not just the averages but variances that will be affected.

My point here is not to explain how ”climate change” parses analytically but to describe the process when we actually unpack the notion and this unpacking is a situated process where we use its conceptual components to attain a discursive goal in a (imagined) situation. When we do that, we switch our focus not only from the higher level concepts to more detailed ones but psychologically we also switch contexts. First we may exploit our discursive analytical knowledge on how ”climate change” is generally defined and then switch to consider weather. This contextual control shift may activate detailed knowledge, which may not be hierarchically related to the top-level notion, for example how pleasant or unpleasant exceptional heat waves may be. Moreover, the discussions about climate change carry a subtext that it is an ongoing process of global warming, which is (depending on who you ask) caused by greenhouse gas emissions. We may have theoretical knowledge about the mechanism, which can be concretized by an analogy how actual greenhouses work or what happens in cars that are parked in direct sunlight. If the top-level context (i.e., the task of discussing what ”climate change” means) is primarily politically laden, we may not unravel the notion in terms of the physical process and weather patterns but our discursive knowledge on how the problem is going to change societies and how it connects to our means of production, global economic inequality, etc. Then the knowledge we draw on comes from our acquaintance with public political debates, our personal moral conversations, social science

classes, and so on. Our intuitive situated interpretation how to engage in the task depend on how we and our reference groups thing and talk about the matter and what opinions we have and feel that we should have. Therefore, the question how complex notions like "climate change" are psychologically constituted should be answered by focusing on intricate interactive *processes*, which contain multiple context, control, and task switches that lead to dynamically drawing on different kinds of knowledge pools in different level of detail to attain specific subgoals under the top-level pragmatic context. This process cannot be understood without accounting for what we are actually doing and why and how the specific contextual factors dictate how the discursive situation unfolds. This process is psychologically not determined only by our explicit know-that but critically by procedural know-how guiding how all that knowledge gets used, which is, again, dependent on active construction of situationally relevant *ad hoc* goals and knowledge structures.

Relations to embodied and enactive theories: Here is a rerun of some of the key quotations characterizing embodied and enactive paradigms from Section 2.3.1: "Cognition is the exercise of skillful know-how in situated and embodied action," (Thompson, 2007, 11), and therefore "cognition is not the representation of a pregiven world by a pregiven mind but is rather the enactment of a world and a mind on the basis of a history of the variety of actions that a being in the world performs" (Varela et al., 1991, 9). Because of the nature of cognitive skills and their relation to conceptually structured thought and action, "'higher' cognitive structures also emerge from recurrent patterns of perceptually guided action" (Varela, 1999, 17). Moreover, as Varela continues, I have also maintained that "the world we know is not pregiven; it is, rather, enacted through our history of structural coupling, and the temporal hinges that articulate enaction are rooted in the number of alternative microworlds that are activated in every situation. These alternatives are the source of both common sense and creativity in cognition." These "microworlds" basically are recurring situations, or interpretations of situations, for which we have relevant readiness for action. In philosophical terms, one key observation here is that human cognition does not track the universal conceptual realm but cognition is an interactive constitutive process with the body and the environment involved in creating one.

I could go on with these quotes, but perhaps it would not be very constructive. Although the content of embodied and enactive theories cannot be captured with such catch lines, I take that these programmatic points are what

I just discussed above and have been underlining throughout this dissertation. I have not much discussed the specific embodied nature of cognition nor the details of the structural coupling of the body, brain, and the environment. Mostly, I have focused on the impact of situated action for higher cognition. However, I have all the time worked under the assumption that some account of embodiment is necessarily entailed by my constant talk about the needs and capacities of the agent, even though my research plan was not to articulate the specific import of embodiment to concept constitution. At least, to what specific features in the environment we are sensitive and how they interact in the active construction of situated representations are things that need to be further specified; but I believe that the researchers who are investigating the lower-level processes and detailed coupling with the environment are doing just that. Therefore, I consider my work as a complementary effort to that line of research. What I have provided is an account of cognitive psychology for the more orthodox embodied/enactive researchers with which they can interface their theories, if they choose to interface their research with standard cognitive psychology at all.

Moreover, radical enactivist seem sometimes to get caught surprised how to deal with questions considering how to conceptualize, for example, cognitive competencies of severely paralyzed people who can not feel and move their bodies or how to understand the situated and embodied nature of thought processes in solitary and overtly inactive contemplation. The answers I have heard a few times are that, paralyzed or not, human bodies are still active; our bodies digesting food, pumping blood, there are hormones doing their work, and so on. Further, we are always situated in a physical environment and at least the gravity has a hold on us, so we are never free from the effect of physical forces, and so on. Every plank entirely true, but I still fail to understand how these factors might be constitutive of explanations of our thought processes. In the passages where I have been discussing thinking as internalized situated reasoning, my general point has been that not every mental *episode* need to be explained in this way. The impact of situated, embodied, and enactive theories needs only be to emphasize how mental contents and processes cannot be understood without taking into account the individual stock of experiences engendered in the (bodily) interaction with the environment. It should also stress the general ecological, practical, social, biological, and other material constraints for which the individual mental faculties are adapted in both evolutionary and individual developmental sense. So situated (bodily) interaction is the key to understanding the general characteristics of human cognition (as

per Corollary 3 above) but the body and environment do not need to be factored as a necessary constitutive part into every possible description of mental episodes.

6.3 Closing remarks concerning the nature of human reason

In closing, I want to bring forward some possible implications of the social pragmatic nature of discursive reasoning. Section 5.3.2 combined the argumentative theory of Mercier and Sperber (2011; 2017) with linguistic pragmatism and ritualistic model learning discussed in Section 2.2.2. This amalgamation has some interesting consequences to how to understand the dynamics of public reasoning and decision-making, the social construction of discursive contents, and how these factors shape individual thinking.

Recall the basic idea of Mercier and Sperber that cognitive faculties for argumentative reasoning have not evolved for improving knowledge but for producing reasons for the claims we make in order to persuade others that our input is reasonable and reliable. The capacity to evaluate arguments serves the function to spot faults in the reasons given by others in order to evaluate the credibility of the information they provide. This may look like a very pessimistic image that portray human rationality and reasoning as a competitive undertaking of attacks and defenses and holding your ground as tightly as you can. However, as the authors note, this is not the correct interpretation if one looks at the social dynamics of what is going on in such interactions. Such a competitive setting can still be a cooperative venture to seek the truth sincerely, and if all goes well, it may actually converge to the truth or at least to the acceptance of the best argument. Indeed, as Mercier and Sperber (2011) remark, from the division of epistemic labor perspective, it may be optimal that not all the participants engaged in a complex epistemic project—be it a research program or a formulation of a political program, for example–invest that much in examining the subject matter critically from all the possible angles. Instead, they should perhaps spend their time on devising the best possible justification for their own standpoint and then bring it under public review, which tends to correct individual biases. Thus, especially with complex issues, the best possible outcome may be reached by a congregation of disagreeing people who have a shared interest in the truth.

An important consequence of how I conceive this process is how social and social psychological factors affect the dynamics of public reasoning. I recall some philosopher has proclaimed that a good argument has a nonviolent but

imperative force; that is, when confronted with a persuasive argument, we are compelled to to accept it and to act according to the truth of the conclusion. Something like this is in the heart of the classical rationalist idea of the human. Reason dictates what is rational to accept, and rational, morally competent and utility maximizing subject does what is the best thing to do according to the current beliefs. Of course, weakness of the will and the passions of the soul complicate things but, roughly, this is the classical image of the dynamics of reasoning, decision-making, and action. I think the cited wisdom has truth in it; however, it has less to do with individual rationality or utility maximization than social conformism, which is not a strict constraint of human-decision making but still a compelling force driving social coordination of collective and individual behavior.

Recall that norm enforcement and compliance is modulated by several factors, such as perceived authority and in- or outgroup membership, the need to identify with the model, and peer pressure. Competence to understand and engage in argumentation necessitates a practical understanding of the topic. In practical reasoning, this often means understanding the practices and their material and social constraints; however, in more theoretical settings, this often means understanding how the local language game plays out; that is, what the acceptable and relevant discursive moves are, and how the topic and conceptual domain under discussion is inferentially constituted. Two participants' disagreement may lead to different outcomes in multiple *a*rational ways which are not necessarily dysfunctional (given the social coordinative function of argumentation) but independent of the quality of the arguments.

If mutual understanding is sufficient, and participants have a reasonable trust in each other, and also in the possibility of reaching at least a tentative consensus, things may develop smoothly as dictated by the shared norms of discursive rationality. However, if the participants come from social different groups with different discursive backgrounds, they may have genuine problems in understanding each other because the meaning engendering inferential norms on which the participants tacitly rely may be too divergent. Because the norms are tacit and often complex, the other side may seem inconsistent, incompetent, and incoherent for reasons that may be nearly impossible to discern, making little or no sense at all. Depending on the social factors, failing to make sense may appear as a norm violation leading to penalizing, which may mean hostility, simple indifference, or an attempt to exclude the other from the shared decision-making process and one's own social group. Hence, *trust* is an issue and so is *commitment* to the mutual enterprise in good faith. Of

course, trust may break and commitments may be subjected to review, but presumably respect and close social affiliation help in seeking common grounds for understanding each other and accommodate others' values and ways of reasoning to shared discursive protocols. In antagonistic, settings participants may be reluctant to accept even good arguments if the other is perceived merely as an adversary to be refuted.

In in-group settings, the dynamics may be somewhat different. In case the recipient does not understand the reasoning of the other, arguments and conclusions may still be accepted and perceived as rational and authoritative. This is presumably the case if the other is perceived as an (epistemic) authority (e.g., a teacher or an expert) or a model with whom the recipient wishes to identify. Then disagreement may be taken as a learning opportunity that provides a model for one's own reasoning protocols. If two participants disagree, any third person may be similarly compelled to evaluate the quality of the arguments based on perceived authority. This is, by and large, rational but only if the authority is merited. Several reasons (again, related to social affiliation, status, etc.) may modulate the argument acceptance in these events.

In general, social factors may contaminate our judgment to a larger extent than we acknowledge, rendering our internalized reasoning protocols (i.e., ways of thinking) as rational or biased as the social norms and protocols that support our reasoning are, as well as the whole settings where learn to reason. In certain environments we may develop non-conscious biases to evaluate arguments selectively solely based on what reference group we identify with and what reference group the dissenting arguments come from. The best argument is by no means guaranteed to win if the partisans do not trust, and do not want to trust, each other and are not committed to coordinate their behavior together. Indeed, anecdotal evidence suggests that people may even be susceptible to a reasoning bias which is in a way converse to the belief bias: seeking disconfirmation to arguments whose conclusion they already believe, if it comes from an adversary social group. The implications for political and moral discourse should be obvious. In these contexts, the arational impact may be particularly pronounced for several reasons.

The first and the most obvious reasons is the sociological and social psychological factors that shape political identities of individuals and groups so that the default attitude between groups may become antagonistic. This is especially problematic if it reaches the point where the discursive practices that shape and reproduce the in-group norms are isolated between groups, breaking the critical feedback between dissenting participants. When political discourse

goes theoretical and value based, there may be little external constraints to check the adequacy of the ideas, save the social feedback. This is different from technical and practical problems that are generally constrained by hard external facts. In Section 5.3.2, I proposed that abstract theoretical discourses resemble ritualistic settings which tends to promote the *ways* of doing things together (which is the ways of reasoning things together) tacitly as a normative end in itself, hence suppressing any need and motivation for a critical inter-group feedback. Under extreme political polarization, argumentative practices simply lose their meaning when the social coordinative function breaks down.

Second, it is important to remember that public reasoning shapes conceptual contents, and after internalization they come to predelineate how we conceive the subject matter. Groups with homogeneous values and discursive practices may end up producing highly biased judgments if there is no social pressure to question their rationales and background commitments and make them explicit. This may result in a production of radical and borderline nonsensical ideas within groups where certain initially innocent inclinations are amplified and gradually transformed into locally shared common sense. This is perhaps a problem mostly associated with think tanks and similar institutions; however, the problem may affect any relatively insulated knowledge-producing community.[130] As we have seen, fluent expertise takes years to develop, and during this process, intellectual intuitions may become too deeply entrenched. One of the key features of the intuitive faculty is that when we have developed an appropriate level of understanding of a domain, we cease to think about its fundamental constituents spontaneously. This is why commonsense conceptions are often seen as self-evident truths. Indeed, there is rarely a practical reason to examine them since intuition automatically directs our attention to what is relevant and provides appropriate interpretations for our practical needs. Thus, our intuitions determine our default stance on the matters what our reasoning episodes are about. If research communities that study related problems become insulated from one another, they may develop a fundamentally different and deeply internalized understandings of their research subject and there is the risk that they may cease to understand what others are doing and end up regarding the related research as not complementary but, at worst, incoherent and nonsensical. This would clearly be detrimental to interdisciplinary collaboration. In the worst case, a tradition may dissociate from other research and degenerate into an irrelevant self-contained discourse

[130] This is obviously a problem also for solitary theorizing because the public checks of individual biases are missing, see (Mercier & Sperber, 2017).

where initially arbitrary but compelling ideas end up as axiomatic facts, or at least appear to participants as highly relevant just because they are frequently referenced for the sake of the idiosyncrasies of the tradition.[131]

While the point is more general, the implications to the methodology of philosophy should be, again, quite evident. This theme was discussed already in Section 3.1; however, now we are in a better position to appreciate its implications. Note how the abstraction process proposed in Section 5.2.2 resembles conceptual analysis. Take the concept of "concept," for example. Early analytical philosophers found that mathematical concepts could be broken down into well-defined logical constructs that reduce to self-evident primitives. Some took this as a model of a general analysis of concepts presumably because it was essentially a regimented version of the classical analysis. Later, Wittgenstein remarked that, actually, concepts like "game" are better characterized by a family resemblance structure instead of a set of necessary and sufficient conditions that try to capture essential meanings. Then Kripke and Putnam used natural kind concepts such as "water" to make a case for concept essentialism of sorts. Despite different analyses, all the time most philosophers agreed to a sufficient degree on what they were talking about. Now, my point is that this case-based methodology resembles my claim that we have some shared understanding of the superordinate schematic concepts like "concept," but that understanding is tacit and to explicate it, we need to concretize the schematic concept with specific examples that prompt our intuitions about its application. After that, we generalize what we have learned based on those examples as an explication of the meaning of the superordinate/schematic concept.

In this way, the conceptual analysis does not necessarily add much to our conceptual knowledge but makes aspects of it explicit. This resembles simulation where through self-stimulation "the brain discovers what it already knows" (Gilbert & Wilson, 2007, 1354). I believe the similarity is due to the fact that the thought experiments that the case-based methodology employs *are* mental simulations of imagined and real concept use situations. Nevertheless, the resulting explication may change how the superordinate concepts are subsequently understood by affecting how they are discussed in the following debates, text-books, and so on. (Recall our discussion on how explicating the

[131] See also MacLeod (2016) on how entrenched discipline-specific practices and epistemic values may hinder interdisciplinary collaboration because scientists do not always understand the methods and requirements of their collaborators. This can lead them to consider the input of their associates as arbitrary and incompetent, eroding the necessary trust between the participants.

concept of *fairness* may change how it is subsequently applied, see page 262.) If a one-sided analysis attains an intellectual hegemony, this may be problematic. However, the practice itself is completely innocent and potentially useful. In fact, it is presumably the only available method for investigating our analytic intuitions. The important point is that these intuitions are often simultaneously produced in the process of investigating them.[132]

One should also exercise caution with thought experiments. The underlying message that runs through the literature on human reasoning is that our intuitions are not trustworthy in situations in which we do not have any practical experience. Therefore, discourses based on intuitions about Chinese rooms, zombie worlds, and perhaps political utopias, should be of suspect. This does not mean that these debates are entirely pointless but that care must be taken that the methodology that aims to free our thinking from the confines of common sense does not end up producing common nonsense. The same goes for several maxims entrenched in philosophical training and academic discourse, in general. However, not all thought experiments are problematic. Many arguments in analytical philosophy probe our intuitions about the use of commonsense concepts such as *knowledge* and deploy thought experiments with realistic and well-understood contexts. We should expect to have reliable intuitions in these instances. The famous Gettier problems (Gettier, 1963) may be a an example.[133] However, highly counterfactual thought experiments invent extremely low-validity environments for our intuition to work with, and the resultant judgments may be close to arbitrary. The specific problem with this is that we may initially accept certain intuitions for the sake of argument, but when the discussion proceeds, we may overlearn the logic of the discourse and eventually end up accepting the resulting semantic intuitions as facts of reason like any common truism. This would resemble the impact of so-called "wicked" environments where our intuitive faculties proficiently follow misleading cues (Kahneman & Klein, 2009, 523).

These issues are not limited to groupthink but affect the thought of individuals, as far as individual reason remains a social product. If we are to form a rational analysis of individual reasoning with domains that are even partially socially constructed—which is almost any domain—we have to take

[132] See Baz (2016) for more in-depth discussion about similar points and a more forceful argument aiming to show that the case-based methodology of philosophy seems to stand (and therefore actually fall) with the representational view of language.

[133] Which does not mean that everyone shares the same intuitions about the Gettier cases but that we have a good reason to presume that the intuitions are inherently questionable.

into account the functioning of the complex social networks: how they produce and maintain knowledge and constrain the judgment of their members. Individuals may operate entirely rationally in their respective intellectual environments but still against epistemological norms accepted elsewhere. So it is not just that good discursive and epistemic practices make better knowledge production communities but they help us make better intuitive judgments at the individual level. Recall the words of Alfred North Whitehead (1911, 61):

> It is a profoundly erroneous truism, repeated by all copy-books and by eminent people when they are making speeches, that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them.

While I just provided grounds for being skeptical whether the automatization of thought is always a sign of progress, these lines capture well the insight that the cognitive power of language does not fundamentally originate from its role in individual psychology but from its coordinative function to shape communal practices that can be partly internalized; and these internalized routines support cultural scaffolding of increasingly complex practices and traits. In this section, I have been mapping the possible fault lines of this cultural scaffolding which, when all goes well, extends our capacity for important operations with less need for time-consuming critical thought. I focus on the problems because they may reveal something important about the constitution of cultured thought and not because we should necessarily emphasize the problems over the possibilities of the cultural evolution that it makes possible.

As far as I can tell, not much of what I have said in this section is really news to anyone familiar with the sociology of knowledge. However, I aimed not to deliver novel insights to the reader interested sociology but rather show how common (albeit not universally held) ideas in that field of study conflate rather naturally with the social ecological account of cognitive psychology and how these research programs could, therefore, be mutually informative. I hasted to add that the problems pertaining to group dynamics and social aspects of individual reasoning are not merely due to social incentives. It is well known from the classics of cognitive anthropology that our culture makes us reluctant to publicly reason in specific ways (e.g. Luria, 1976). It is also common wisdom, that we endorse arguments, conclusions, values, and language games based on social affiliation. However, my point is that when we engage in these language

games, they change how we think and understand the world around us, and this impact is mostly tacit. Although social incentives and values often are in play, the point is eventually not about incentives to think or talk in particular ways but about tendencies to intuitively understand the world in particular ways, which are deeply entrenched, automatic, hard to notice, and which unfortunately often suppress understanding of another kinds of perspectives.

I'm afraid that in this concluding chapter I finally sound like a full-blown post-modernist. And yes, I think many of the post-modernists got some aspects of human condition in many ways more right that their adversaries in the universalist enlightenment camp. As for my defense, all I can say is that I think it would still be a sort of naturalistic fallacy to make prescriptive conclusions based on this. I do not know if there are objective values or truth. Note that this is not to suspect the existence of objective reality in which I firmly believe. Without presuming metaphysical material realism, no ecological framework makes any sense. "Truth," remember, is not a property of reality but a property of representational vehicles we use to describe reality. What I have tried to argue is not that facts do not exist or matter but that at least theoretical facts matter only through cultural process. Regardless whether or not there are objective truths, I think it is often a very good idea to at least *pretend* that there are, because such an idea keeps us committed to joint projects and standards even if we can not find reasonable grounds to commit us to each other. Moreover, compared to some radical relativist thinkers, I may be more optimistic about the prospects of finding common grounds between subjects with a divergent stock of experiences—even if all the tacit knowledge we acquire throughout our life cannot be communicated entirely. Much of what unifies us remains invisible, being too commonplace and obvious to take notice. Differences are easier to spot when our tacit expectations about mutual understanding fails, the fault lines break and conflicts become visible. It is just often not a very good idea to enforce the truths and values we find upon others because we can never be really sure if our unshakable deep convictions reflect the objective truths of reason and general human condition or only a confined corner of the complex manifold of potential human experience and understanding.

References

Abraham, Ralph & Shaw, Christopher D. 1992: *Dynamics: The Geometry of Behavior, 2nd ed.* Boston, MA: Addison-Wesley.

Ahn, Woo-kyoung; Gelman, Susan A.; Amsterlaw, Jennifer A.; Hohenstein, Jill & Kalish, Charles W. 2000a: "Causal status effect in children's categorization" *Cognition*, 76(2), B35–B43.

Ahn, Woo-kyoung; Kim, Nancy S.; Lassaline, Mary E. & Dennis, Martin J. 2000b: "Causal Status as a Determinant of Feature Centrality" *Cognitive Psychology*, 41(4), 361–416.

Ahn, Woo-kyoung; Taylor, Eric G.; Kato, Daniel; Marsh, Jessecae K. & Bloom, Paul 2013: "Causal Essentialism in Kinds" *The Quarterly Journal of Experimental Psychology*, 66(6), 1113–1130.

Allen, Scott W. & Brooks, Lee R. 1991: "Specializing the Operation of an Explicit Rule" *Journal of Experimental Psychology: General*, 120(1), 3–19.

Almor, Amit & Sloman, Steven A. 1996: "Is Deontic Reasoning Special?" *Psychological Review*, 103(2), 374–380.

Amalric, Marie & Dehaene, Stanislas 2016: "Origins of the brain networks for advanced mathematics in expert mathematicians" *PNAS*, 113(18), 4909–4917.

Amalric, Marie; Denghien, Isabelle & Dehaene, Stanislas 2018: "On the role of visual experience in mathematical development: Evidence from blind mathematicians" *Developmental Cognitive Neuroscience*, 30, 314–323.

American Psychiatric Association 2013: *Diagnostic and statistical manual of mental disorders (5th ed.)* Washington, DC: APA.

Anderson, John R. 1991a: "Is human cognition adaptive?" *Behavioral and Brain Sciences*, 14(3), 471–517.

Anderson, John R. 1991b: "The Adaptive Nature of Human Cognition" *Psychological Review*, 98(3), 409–429.

Arbib, Michael A. 2005: "From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics" *Behavioral and Brain Sciences*, 28(2), 125–167.

Armstrong, Sharon Lee; Gleitman, Lila R. & Gleitman, Henry 1983: "What some concepts might not be" *Cognition*, 13(3), 263–308.

Ashby, W. Ross 1956: *An Introduction to Cybernetics.* London: Chapman & Hall.

Ashby, W. Ross 1960: *A Design for a Brain (2nd ed.)* New York, NY: John Wiley & Sons.

Austin, John L. 1999: *How to Do Things With Words.* Cambridge, MA: Harvard University Press. Originally published 1962.

Baillargeon, Renée 1998: "Infants' understanding of the physical world" in Sabourin, Michel; Craik, Fergus & Robert, Michèle (eds.) 1998: *Advances in Psychological Science, Vol. 2: Biological and Cognitive Aspects.* Hove: Psychology Press, 503–529.

Baillargeon, Renée 2002: "The Acquisition of Physical Knowledge in Infancy: A Summary in Eight Lessons" in Goswami, Usha (ed.) 2002: *Blackwell Handbook of Childhood Cognitive Development.* Malden, MA: Blackwell Publishing, 47–83.

Barsalou, Lawrence W. 1983: "Ad hoc categories" *Memory & Cognition*, 11(3), 211–227.

Barsalou, Lawrence W. 1985: "Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure in Categories" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4), 629–654.

Barsalou, Lawrence W. 1987: "The instability of graded structure: implications for the nature of concepts" In Neisser, Ulric (ed.) 1987: *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization.* Cambridge: Cambridge University Press, 101–140.

Barsalou, Lawrence W. 1991: "Deriving Categories to Achieve Goals" In Bower, Gordon (ed.) 1991: *The Psychology of Learning and Motivation: Advances in Research and Theory, volume 27.* San Diego, CA: Academic Press, 1–64.

Barsalou, Lawrence W. 1999: "Perceptual symbol systems" *Behavioral and Brain Sciences*, 22(4), 577–660.

Barsalou, Lawrense W. 2003: "Abstraction in perceptual symbol systems" *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1435), 1177–1187.

Barsalou, Lawrence W. 2008: "Grounded Cognition" *Annual Review of Psychology*, 59, 617–645.

Barsalou, Lawrence W.; Simmons, W. Kyle; Barbey, Aron K. & Wilson, Christine D. 2003: "Grounding conceptual knowledge in modality-specific system" *Trends in Cognitive Sciences*, 7(2), 84–91.

Baz, Avner 2016: "On going (and getting) nowhere with our words: New skepticism about the philosophical method of cases" *Philosophical Psychology*, 29(1), 64–83.

Beckers, Tom; De Houwer, Jan & Pineño, Oskar & Miller Ralph R. 2005: "Outcome Additivity and Outcome Maximality Influence Cue Competition in Human Causal Learning" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 238–249.

Beebe, James R. 2012: "Experimental Epistemology" In Cullison, Andrew (ed.) 2012: *The Continuum Companion to Epistemology.* London: Continuum, 248–269.

Beer, Randall D. 1995: "A Dynamical systems perspective on agent-environment interaction" *Artificial Intelligence*, 72, 173–215.

Beer, Randall D. 2000: "Dynamical approaches to cognitive science" *Trends in Cognitive Sciences*, 4(3), 91–99.

Bechtel, William & Abrahamsen, Adele 2002: *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks, 2nd edition.* Malden, MA: Blackwell Publishing.

Beller, Sieghard 2012: "Concrete problems in the abstract deontic selection task—And how to solve them" *The Quarterly Journal of Experimental Psychology*, 65(7), 1414–1429.

Benoit, Roland G.; Szpunar, Karl K. & Schater, Daniel L. 2014: "Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge" *Proceedings of the National Academy of Sciences of the United States of America*, 111(46), 16550–16555.

Berlin, Brent; Breedlove, Dennis E. & Raven, Peter H. 1973: "General Principles of Classification and Nomenclature in Folk Biology" *American Anthropologist, New Series*, 75(1), 214–242.

Berridge, Kent C. & O'Doherty, John P. 2014: "From Experienced Utility to Decision Utility" in Glimcher, Paul W. & Fehr, Ernst (eds.) 2014: *Neuroeconomics: Decision Making and the Brain (2nd ed.)* London: Academic Press, 335–351.

Blair, Mark & Homa, Don 2001: "Expanding the search for a linear separability constraint on category learning" *Memory & Cognition*, 29(8), 1153–1164.

Block, Ned 1986: "Advertisement for a Semantics for Psychology" *Midwest Studies in Philosophy*, 10, 615–678.

Brandom, Robert 1994: *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Cambridge, MA: Harvard University Press.

Brandom, Robert 2000: *Articulating Reasons: An Introduction to Inferentialism.* Cambridge, MA: Harvard University Press.

Brandom, Robert 2008: *Between Saying and Doing: Towards an Analytic Pragmatism.* Oxford: Oxford University Press.

Bransford, John D. & Johnson, Marcia K. 1973: "Considerations of some problems of comprehension" In Chase, William G. (ed.) 1973: *Visual Information Processing.* New York, NY: Academic Press, 383–438.

Brogaard, Berit & Gatzia, Dimitria Electra 2017: "Unconscious Imagination and the Mental Imagery Debate" *Frontiers in psychology*, 8(799), doi:10.3389/fpsyg.2017.00799

Brooks, Lee R. 1978: "Nonanalytic concept formation and memory for instances" In Rosch, Eleanor & Lloyd, Barbara B. (eds.) 1978: *Cognition and Categorization.* Hillsdale, NJ: Lawrence Erlbaum Associates, 169–211.

Brooks, Lee R.; Norman, Geoffrey R. & Allen, Scott W. 1991: "Role of Specific Similarity in a Medical Diagnostic Task" *Journal of Experimental Psychology: General*, 120(3), 278–287.

Brooks, Rodney A. 1999: *Cambrian Intelligence: The Early History of the New AI.* Cambridge, MA: The MIT Press.

Bruner, Jerome S. 1966: *Toward a Theory of Instruction.* Cambridge, MA: Harvard University Press.

Bruner, Jerome S. 1990: *Acts of Meaning.* Cambridge, MA: Harvard University Press.

Campitelli, Guillermo & Gobet, Fernand 2010: "Herbert Simon's Decision-Making Approach: Investigation of Cognitive Process in Experts." *Review of General Psychology*, 14(4), 354–364.

Carey, Susan 2009: *The Origin of Concepts.* Oxford: Oxford University Press.

Carnap, Rudolf 1956: "The Methodological Character of Theoretical Concepts" In Feigl, Herbert & Scriven, Michael (eds.) 1956: *Minnesota Studies in Philosophy of Science, vol. I: Foundations of Science & the Concepts of Psychology and Psychoanalysis.* Minneapolis, MN: University of Minnesota Press, 38–75

Carroll, Lewis 1895: "What the Tortoise Said to Achilles" *Mind*, 4, 278–280.

Chase, William G. & Simon, Herbert A. 1973: "Perception in Chess" *Cognitive Psychology*, 4(1), 55–81.

Cheng, Patricia 1997: "From Covariation to Causation: A Causal Power Theory" *Psychological Review*, 104(2), 367–405.

Cheng, Patricia W. & Holyoak, Keith J. 1985: "Pragmatic Reasoning Schemas" *Cognitive Psychology*, 17(4), 391–416.

Cheng, Patricia W. & Holyoak, Keith J. 1989: "On the natural selection of reasoning theories" *Cognition*, 33, 285–313.

Cheng, Patricia W.; Holyoak, Keith J.; Nisbett, Richard E. & Oliver, Lindsay M. 1986: "Pragmatic versus Syntactic Approaches to Training Deductive Reasoning" *Cognitive Psychology*, 18(3), 293–328.

Chi, Michelene T. H.; Feltovich, Paul J. & Glaser, Robert 1981: "Categorization and Representation of Physics Problems by Experts and Novices" *Cognitive Science*, 5(2), 121–152.

Chomsky, Noam 1957: *Syntactic Structures.* The Hague: Mouton.

Chomsky, Noam 1965: *Aspects of the Theory of Syntax.* Cambridge, MA: The MIT Press.

Chooi, Weng-Tink 2012: "Working Memory and Intelligence: A Brief Review" *Journal of Education and Developmental Psychology*, 2(2), 42–50.

Christensen, Wayne; Sutton, John & Mcilwain, Doris J. F. 2016: "Cognition in Skilled Action: Meshed Control and the Varieties of Skilled Experience." *Mind & Language*, 31(1), 37–66.

Churchland, Paul M. 1989: *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science.* Cambridge, MA: The MIT Press.

Cialdini, Robert B. 2007: "Descriptive social norms as underappreciated sources of social control." *Psychometrika*, 72(2), 263–268.

Clark, Andy 1989: *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* Cambridge, MA: The MIT Press.

Clark, Andy 1997: *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA: The MIT Press.2

Clark, Andy 1999: "An embodied cognitive science?" *Trends in Cognitive Sciences*, 3(9), 345–351.

Clark, Andy 2008: *Supersizing the Mind: Embodiment, action, and cognitive extension.* Oxford: Oxford University Press.

Clark, Andy & Chalmers, David J. 1998: "The Extended Mind." *Analysis*, 58(1), 7–19.

Clark, Andy & Karmiloff-Smith, Annette 1993: "The Cognizer's Innards: A Psychological and Philosophical Perspective on the Development of Thought" *Mind & Language*, 8(4), 487–519.

Clark, Andy & Toribio, Josefa 1994: "Doing Without Representing?" *Synthese*, 101, 401–431.

Clarke-Donae, Justin 2013: "What is Absolute Undecidability?" *Noûs*, 47(3), 467–481.

Clement, John J. 2004: "Imagistic Processes in Analogical Reasoning: Conserving Transformations and Dual Simulations" in Forbus, Kenneth; Gentner, Dedre & Regier, Terry (eds.) 2004: *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates, 233–238.

Cohen, L. Jonathan 1992: *An Essay on Belief and Acceptance.* Oxford: Clarendon Press.

Collins, Anne G. E.; Cavanagh, James F. & Frank Michael J. 2014: "Human EEG Uncovers Latent Generalizable Rule Structure during Learning" *The Journal of Neuroscience*, 34(13), 4677–4685.

Collins, Anne G. E. & Frank, Michael J. 2013: "Cognitive Control Over Learning: Creating, Clustering, and Generalizing Task-Set Structure" *Psychological Review*, 120(1), 190–229.

Collins, Anne G. E. & Koechlin, Etienne 2012: "Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making" *PLoS Biology*, 10(3), doi: 10.1371/journal.pbio.1001293

Cosmides, Leda 1989: "The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task." *Cognition*, 31, 187–276.

Cosmides, Leda & Tooby, John 1992: "Cognitive Adaptations for Social Exchange." In Barkow, Jerome; Cosmides, Leda & Tooby, John (eds.) 1992: *The Adapted Mind: Evolutionary psychology and the generation of culture.* New York, NY: Oxford University Press, 163–228.

Craik, Kenneth 1943: *The Nature of Explanation.* Cambridge: Cambridge University Press.

Dar-Nimrod, Ilan & Heine, Steven J. 2011: "Genetic Essentialism: On the Deceptive Determinism of DNA" *Psychological Bulletin*, 137(5), 800–818.

Davis, Ernest & Marcus, Gary 2015: "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence." *Communications of the ACM*, 58(8), 92–103.

Decety, Jean & Grèzes, Julie 2006: "The power of simulation: Imagining one's own and other's behavior" *Brain Research*, 1079(1), 4–14.

Dennett, Daniel C. 1971: "Intentional Systems." *The Journal of Philosophy*, 68(4), 87–106.

Dennett, Daniel C. 1984: "Cognitive Wheels: The Frame Problem of AI." In Hookway, Christopher (ed.) 1984: *Minds, Machines And Evolution.* Cambridge: Cambridge University Press, 129–152.

Dennett, Daniel C. 1987: *The Intentional Stance.* Cambridge, MA: The MIT Press.

Dennett, Daniel C. 1991a: *Consciousness Explained.* New York, NY: Back Bay Books.

Dennett, Daniel C. 1991b: "Real Patterns." *The Journal of Philosophy*, 88(1), 27–51.

Di Paolo, Ezequiel A.; Rohde, Marieke & De Jaegher, Hanne 2011: "Horizons for the Enactive Mind: Values, Social Interaction, and Play." In Stewart, John; Gapenne, Olivier & Di Paolo, Ezequiel A. 2011: *Enaction : Towards a New Paradigm for Cognitive Science.* Cambridge, MA: The MIT Press.

Dobzhansky, Theodosius 1973: "Nothing in Biology Makes Sense except in the Light of Evolution" *The American Biology Teacher*, 35(3), 125–129.

Doll, Bradley B.; Simon, Dylan A. & Daw, Nathaniel D. 2012: "The ubiquity of model-based reinforcement learning" *Current Opinion in Neurobiology*, 22(6), 1075–1081.

Domenech, Philippe & Koechlin, Etienne 2015: "Executive control and decision-making in the prefrontal cortex" *Current Opinion in Behavioral Sciences*, 1, 101–106.

Dretske, Fred I. 1981: *Knowledge and the Flow of Information.* Cambridge, MA: The MIT Press.

Dretske, Fred I. 1988: *Explaining Behavior: Reasons in the a World of Causes.* Cambridge, MA: The MIT Press.

Dreyfus, Hubert L. 1965: *Alchemy and Artificial Intelligence.* Santa Monica, CA: The RAND Corporation.

Dreyfus, Hubert L. 1992: *What Computers* Still *Can't Do: A Critique of Artificial Reason.* Cambridge, MA: The MIT Press.

Dreyfus, Hubert L. & Dreyfus, Stuart E (with Athanasiou, Tom) 1986: *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* New York, NY: The Free Press.

Duverne, Sandrine & Koechlin, Etienne 2017: "Rewards and Cognitive Control in the Human Prefrontal Cortex" *Cerebral Cortex*, 27(10), 5024–5039.

Dyson, Freeman 1958: "Innovation in Physics" *Scientific American*, 199(3), 74–83.

Enfield, Nick J. 2015: *Natural causes of language: Frames, biases, and cultural transmission.* Berlin: Language Science Press.

Ericsson, K. Anders; Krampe, Ralf Th.; Tesch-Romer, Clemens 1993: "The Role of Deliberate Practice in the Acquisition of Expert Performance" *Psychological Review*, 100(3), 363–406.

Evans, Jonathan St. B. T. 1984: "Heuristic and analytic processes in reasoning." *British Journal of Psychology*, 75, 451–468.

Evans, Jonathan St. B. T. 1991: "Adaptive cognition: The question is how" *Behavioral and Brain Sciences*, 14(3), 493–494.

Evans, Jonathan St. B. T. 1996: "Deciding before you think: Relevance and reasoning in the selection task." *British Journal of Psychology*, 87, 223–240.

Evans, Jonathan St. B. T. 2003: "In two minds: dual-process accounts of reasoning" *Trends in Cognitive Sciences*, 7(10), 454–469.

Evans, Jonathan St. B. T. 2008: "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition" *Annual Review of Psychology*, 59, 255–78.

Evans, Jonathan St. B. T. 2009: "How many dual-process theories we need? One, two, or many?" (Evans & Frankish, 2009, 35–54)

Evans, Jonathan St. B. T.; Baston, Julie L. & Pollard, Paul 1983: "On the conflict between logic and belief in syllogistic reasoning." *Memory and Cognition*, 11(3), 295–306.

Evans, Jonathan St. B. T. & Frankish, Keith (eds.) 2009: *In Two Minds: Dual Processes and Beyond.* Oxford: Oxford University Press.

Evans, Jonathan St. B. T.; Newstead, Stephen E. & Byrne, Ruth M. J. 1993: *Human Reasoning: The Psychology of Deduction.* Hove: Psychology Press.

Evans, Jonathan St. B. T. & Over, David 1996: *Rationality and Reasoning.* Hove: Psychology Press.

Evans, Jonathan St. B. T. & Stanovich, Keith E. 2013: "Dual-Process Theories of Higher Cognition: Advancing the Debate" *Perspectives on Psychological Science*, 8(3), 223–241.

Evans, Jonathan St. B. T. & Wason, Peter C. 1976: "Rationalization in a Reasoning Task" *British Journal of Psychology*, 67(4), 479–486.

Fadiga, Luciano & Craighero, Laila 2006: "Hand actions and speech representation in Broca's area" *Cortex*, 42(2), 486–490.

Fadiga, Luciano; Craighero, Laila; Fabbri-Destro, Maddalena; Finos, Livio; Cotillon-Williams, Nathalie; Smith, Andrew T. & Castiello, Umberto 2006: "Language in shadow" *Social Neuroscience*, 1(2), 77–89.

Faisal, Aldo; Stout, Dietrich; Apel, Jan & Bradley, Bruce 2010: "The Manipulative Complexity of Lower Paleolithic Stone Toolmaking" *PLoS ONE*, 5(11), 1–11.

Fast, Cynthia D. & Blaisdell, Aaron P. 2011: "Rats are sensitive to ambiguity" *Psychonomic Bulletin and Review*, 18(6), 1230–1237.

Feldman, Jerome A. & Ballard, Dana H. 1982: "Connectionist Models and Their Properties" *Cognitive Science*, 6, 205–254.

Field, Hartry 1977: "Logic, Meaning, and Conceptual Role" *The Journal of Philosophy*, 74(7), 379–409.

Fitch, W. Tecumseh 2010: *The Evolution of Language.* Cambridge: Cambridge University Press.

Floreano, Dario; Kato, Toshifumi; Marocco, Davide & Sauser, Eric 2004: "Coevolution of active vision and feature selection" *Biological Cybernetics*, 90, 218–228.

Floreano, Dario & Mattiussi, Claudio 2008: *Bio-Inspired Artificial Intelligence: Theories, Methods, and Technologies.* Cambridge, MA: The MIT Press.

Fodor, Jerry A. 1975: *The Language of Though.* Cambridge, MA: Harvard University Press.

Fodor, Jerry A. 1980: "Methodological solipsism considered as a research strategy in cognitive psychology" *Behavioral and Brain Sciences*, 3, 1, 63–73.

Fodor, Jerry A. 1983: *The Modularity of Mind.* Cambridge, MA: The MIT Press.

Fodor, Jerry A. 1987: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, MA: The MIT Press.

Fodor, Jerry A. 1990: *A Theory of Content and Other Essays.* Cambridge, MA: The MIT Press.

Fodor, Jerry A. & Lepore, Ernest 1991: "Why Meaning (Probably) Isn't Conceptual Role" *Mind & Language*, 6(4), 328–343.

Fodor, Jerry A. & Lepore, Ernest 1996: "The red herring and the pet fish: why concepts still can't be prototypes" *Cognition*, 58(2), 253–270.

Fodor, Jerry A. & Pylyshyn, Zenon W. 1988: "Connectionism and cognitive architecture: A critical analysis" *Cognition*, 28(1–2), 3–71.

Frankish, Keith 2004: *Mind and Supermind.* Cambridge: Cambridge University Press.

Frankish, Keith & Evans, Jonathan, St. B. T. 2009: "The duality of mind: A historical perspective" (Evans, 2009, 1–29)

Føllesdal, Dagfinn 1966: "Husser's Notion of Noema" *The Journal of Philosophy*, 66(20), 680–687.

Gallagher, Shaun & Miyahara, Katsunori 2012: "Neo-pragmatism and enactive intentionality." In Schulkin, Jay (ed.) 2012: *Action, Perception and the Brain: Adaptation and Cephalic Expression.* Basingstoke: Palgrave Macmillan, 117–146.

van Gelder, Timothy 1995: "What Might Cognition Be, If Not Computation." *Journal of Philosophy*, 92(7), 345–381.

van Gelder, Timothy & Port, Robert F. 1995: "It's About Time: An Overview of the Dynamical Approach to Cognition." (In Port & van Gelder, 1995, 1–43)

Gelman, Susan A. 2003: *The Essential Child: Origins of Essentialism in Everyday Thought.* Oxford: Oxford University Press.

Gelman, Susan A. 2004: "Psychological essentialism in children." *Trends in Cognitive Sciences*, 8(9), 404–409.

Gelman, Susan A.; Coley, John D. & Gottfried, Gail M. 1994: "Essentialist beliefs in children: The acquisition of concepts and theories" In Gelman, Susan A. & Hirschfeld, Lawrence A. 1994: *Mapping the Mind: Domain Specificity in Cognition and Culture*; Cambridge, Cambridge University Press, 341–365.

Gelman, Susan A. & Markman, Ellen M. 1986: "Young Children's Inductions from Natural Kinds: The Role of Categories and Appearances." *Child Development*, 58(6), 1532–1541.

Gelman, Susan A. & Medin, Douglas L. 1993: "What's So Essential About Essentialism? A Different Perspective on the Interaction of Perception, Language, and Conceptual Knowledge" *Cognitive Development*, 8, 157–167.

Gelman, Susan A. & Wellman, Henry W. 1991: "Insides and essences: Early understanding of the non-obvious" *Cognition*, 38(3), 213–244.

Gendler, Tamar Szabó 2010: *Intuition, Imagination, and Philosophical Methodology.* Oxford: Oxford University Press.

Gentner, Dedre 1983: "Structure-mapping: A theoretical framework for analogy" *Cognitive Science*, 7(2), 155–170.

Gentner, Dedre 1989: "The mechanisms of analogical learning" in Vosniadou, Stella & Ortony, Andrew (eds.) 1989: *Similarity and analogical reasoning.* Cambridge: Cambridge University Press, 199–241.

Gentner, Dedre & Boroditsky, Lera 1999: "Individuation, relativity, and early word learning" in Bowerman, Melissa & Levinson, Stephen C. (eds.) 1999: *Language acquisition and conceptual development.* Cambridge: Cambridge University Press.

Gentner, Dedre; Lowenstein, Jeffrey & Thompson Leigh 2003: "Learning and Transfer: A General Role for Analogical Encoding" *Journal of Educational Psychology*, 95(2), 393–408.

Gettier, Edmund L. 1963: "Is Justified True Belief Knowledge?" *Analysis*, 23(6), 121–123.

Gibson, James J. 1979: *The Ecological Approach to Visual Perception.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gick, Mary L. & Holyoak, Keith J. 1983: "Schema Induction and Analogical Transfer" *Cognitive Psychology*, 15, 1–38.

Gigerenzer, Gerd & Regier, Terry 1996: "How Do We Tell an Association From a Rule? Comment on Sloman (1996)" *Psychological Bulletin*, 119(1), 23–26.

Gillan, Douglas J.; Premack, David & Woodruff, Guy 1981: "Reasoning in the Chimpanzee: I. Analogical Reasoning" *Journal of Experimental Psychology: Animal Behavior Processes*, 7(1), 1–17.

Gilovich, Thomas; Griffin, Dale W. & Kahneman, Daniel (eds.) 2002: *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge: Cambridge University Press.

Gobet, Fernand & Simon, Herbert A. 1996: "Recall of rapidly presented random chess positions is a function of skill" *Psychonomic Bulletin & Review*, 3(2), 159–163.

Goel, Vinod; Shuren, Jeffrey; Sheesley, Laura & Grafman, Jordan 2004: "Asymmetrical involvement of frontal lobes in social reasoning" *Brain*, 127(4), 783–790.

Goel, Vinod 2005: "Cognitive Neuroscience of Deductive Reasoning" In Holyoak, Keith J. & Morrison, Robert G. (eds.) 2005: *The Cambridge Handbook of Thinking and Reasoning.* Cambridge: Cambridge University Press, 475–492

Goel, Vinod 2007: "Anatomy of deductive reasoning" *TRENDS in Cognitive Sciences*, 11(10), 435–441.

Goldberg, Lewis R. 1970: "Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences" *Psychological Bulletin*, 73(6), 422–432.

Goldstone, Robert 1994: "Influences of Categorization on Perceptual Discrimination" *Journal of Experimental Psychology: General*, 123(2), 178–200.

Good, I. J. 1960: "The Paradox of Confirmation" *The British Journal for the Philosophy of Science*, 11(42), 145–149.

Gopnik, Alison; Glymour, Clark; Sobel, David M.; Schulz, Laure E. & Kushnir, Tamar 2004: "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets" *Psychological Review*, 111(1), 3–32.

Greene, Joshua & Haidt, Jonathan 2002: "How (and where) does moral judgment work?" *Trends in Cognitive Sciences*, 6(12), 517–523.

Grice, H. Paul 1957: "Meaning" *The Philosophical Review*, 66(3), 377–388.

Gilbert, Daniel T. & Wilson, Timothy D. 2007: "Prospection: Experiencing the Future" *Science*, 317(5843), 1351–1354.

Griggs, Richard A. & Cox, James R. 1982: "The elusive thematic-materials effect in Wason's selection task" *British Journal of Psychology*, 73, 407–420.

Griggs, Richard A. & Ransdell, Sarah E. 1986: "Scientists and the Selection Task" *Social Studies of Science*, 16(2), 391–330.

Griffiths, Thomas L. & Tenenbaum, Joshua B. 2005: "Structure and strength in causal induction" *Cognitive Psychology*, 51(4), 334–384.

Griffiths, Thomas L. & Tenenbaum, Joshua B. 2006: "Optimal Predictions in Everyday Cognition" *Psychological Science*, 17(1), 767–773.

Griffiths, Thomas L. & Tenenbaum, Joshua B. 2009: "Theory-Based Causal Induction" *Psychological Review*, 116(4), 661–716.

de Groot, Adriaan D. 1965: *Thought and Choice in Chess.* The Hague: Mouton. Originally published as a doctoral thesis in Dutch *Het Denken van den Schaker*, 1946.

Hagmayer, York & Waldmann, Michael R. 2000: "Simulating Causal Models: The Way to Structural Sensitivity" in Gleitman, Lila R. & Joshi, Aravind K. (eds.) 2000: *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates, 214–219.

Haidt, Jonathan 2001: "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review*, 108(4), 814–834.

Haidt, Jonathan & Bjorklund, Fredrik 2007: "Social intuitionists answer six questions about morality" In Sinnott-Armstrong, Walter (ed.) 2007: *Moral psychology, Vol. 2: The cognitive science of morality.* Cambridge, MA: The MIT Press.

Hampton, James A. 1982: "A demonstration of intransitivity in natural categories" *Cognition*, 12(2), 151–164.

Hampton, James A. 1993: "Prototype Models of Concept Representation" In Van Mechelen, I.; Hampton, J.A.; Michalski, R.S.; & Theuns, P. 1993: *Categories and concepts: Theoretical views and inductive data analysis.* London: Academic Press, 67–95.

Hampton, James A. 1995: "Testing the Prototype Theory of Concepts" *Journal of Memory and Language*, 34, 686–708.

Hampton, James A. 1998: "Similarity-based categorization and fuzziness of natural categories" *Cognition*, 65(2–3), 137–165.

Hampton, James A. 2006: "Concepts as Prototypes", in Ross, Brian H. (ed.) 2006: *The Psychology of Learning and Motivation: Advances in Research and Theory, vol. 46.* London: Academic Press, 79–113.

Harman, Gilbert 1974: "Meaning and Semantics", in Munitz, Milton & Unger, Peter (eds.) 1974: *Semantics and Philosophy.* New York, NY: New York University Press, 1–16.

Harman, Gilbert 1975: "Language, Thought, and Communication" in Gunderson, Keith (ed.) 1975: *Minnesota Studies in the Philosophy of Science, vol. VII: Language, Mind, and Knowledge.* Minneapolis, MN: University of Minnesota Press, 270–298.

Harman, Gilbert 1987: "(Non-Solipsistic) Conceptual Role Semantics" in Lepore, Ernest (ed.) 1987: *New Directions in Semantics.* London: Academic Press, 55–81.

Harnad, Stevan 1990: "The Symbol Grounding Problem" *Physica D*, 42(1–3), 335–346.

Harnad, Stevan 2005: "To Cognize is to Categorize: Cognition is Categorization" In Cohen, Henri & Lefebre, Claire 2005: *Handbook of Categorization in Cognitive Science.* Amsterdam: Elsevier, 19–43.

Hatfield, Gary 2007: "The *Passions of the soul* and Descartes's machine psychology" *Studies in History and Philosophy of Science*, 38(1), 1–35.

Haugeland, John 1990: "The Intentionality All-Stars." *Philosophical Perspectives*, 4, 383–427.

Hauser, Marc; Cushman, Fiery; Young, Liane; Jin, R. Kang-Xing & Mikhail, John 2007: "A Dissociation Between Moral Judgments and Justification." *Mind & Language*, 22(1), 1–12.

Hegarty, Mary 2004: "Mechanical reasoning by mental simulation" *Trends in Cognitive Sciences*, 8(6), 280–285.

Heit, Evan 1998: "Influences of Prior Knowledge on Selective Weighting of Category Members" *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24(3), 712–731.

Heit, Evan & Rubinstein, Joshua 1994: "Similarity and Property Effects in Inductive Reasoning" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(2), 411–422.

Henrich, Joseph; Boyd, Robert; Bowles, Samuel; Camerer, Colin; Fehr, Ernst; Gintis, Herbert; McElreath, Richard; Alvard, Michael; Barr, Abigail; Ensminger, Jean; Henrich, Natalie Smith; Hill, Kim; Gil-White, Francisco; Gurven, Michael; Marlowe, Frank W.; Patton, John Q. & Tracer, David 2005: "'Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies" *Behavioral and Brain Sciences*, 28(6), 795–855.

Hergenhahn, Baldwin R. & Henley, Tracy B. 2014: *An Introduction to the History of Psychology, 7th ed.* Belmont, CA: Cangage Learning.

Hesslow, Germund 2002: "Conscious thought as simulation of behaviour and perception" *Trends in Cognitive Sciences*, 6(6), 242–247.

Hoch, Stephen J. & Tschirgi, Judith E. 1985: "Logical knowledge and cue redundancy in deductive reasoning" *Memory & Cognition*, 13(5), 453–462.

Holyoak, Keith 1985: "The Pragmatics of Analogical Transfer" in Bower, Gordon H. 1985: *The Psychology of Learning and Motivation, vol. 19.* New York, NY: Academic Press.

Holyoak, Keith J. & Cheng, Patricia W. 1995: "Pragmatic Reasoning with a Point of View" *Thinking and Reasoning*, 1(4), 289–313.

Holyoak, Keith J. & Cheng, Patricia W. 2011: "Causal Learning and Inference as a Rational Process: The New Synthesis." *Annual Review of Psychology*, 62, 135–163.

Holyoak, Keith J. & Koh, Kyunghee 1987: "Surface and structural similarity in analogical transfer" *Memory & Cognition*, 15(4), 332–340.

Holyoak, Keith J.; Lee, Hee Seung & Lu, Hongjing 2010: "Analogical and Category-Based Inference: A Theoretical Integration With Bayesian Causal Models" *Journal of Experimental Psychology*, 139(4), 702–727.

Holyoak, Keith J.; Novick, Laura R. & Melz, Eric R. 1994: "Component Processes in Analogical Transfer: Mapping, Pattern Completion, and Adaptation." In Holyoak, Keith J. & Barnden, John A. (eds.) 1994: *Advances in Connectionist and Neural Computation Theory, vol. 2: Analogical Connections.* Norwood, NJ: Ablex, 113–180.

Holyoak, Keith J. & Thagard, Paul 1997: "The Analogical Mind" *American Psychologist*, 52(1), 35–44.

Homa, Donald & Vosburgh, Richard 1976: "Category Breadth and the Abstraction of Prototypical Information" *Journal of Experimental Psychology: Human Learning and Memory*, 2(3), 322–330.

Hornik, Kurt 1991: "Approximation Capabilities of Multilayer Feedforward Networks." *Neural Networks*, 4, 251–257.

Jackendoff, Ray 1999: "Possible stages in the evolution of language capacity." *Trends in Cognitive Sciences*, 3(7), 272–279.

Jackson, Sherri L. & Griggs, Richard A. 1988: "Education and the selection task." *Bulletin of the Psychonomic Society*, 26(4), 327–330.

James, William 1890: *The Principles of Psychology.* New York, NY: Henry Holt & co.

Johansen, Mark K. & Palmeri, Thomas J. 2002: "Are there representational shifts during category learning?" *Cognitive Psychology*, 45(4), 482–553.

Johnson-Laird, Philip N. 2004: "The History of Mental Models" in Manktelow, Ken & Chung, Man Cheung (eds.) 2004: *Psychology of Reasoning: Theoretical and Historical Perspectives.* Hove: Psychology Press, 179–212.

Johnson-Laird, Philip N. 2008: *How We Reason?* Oxford: Oxford University Press.

Johnson-Laird, Philip; Legrenzi, Paolo & Legrenzi, Maria Sonino 1972: "Reasoning and a Sense of Reality." *British Journal of Psychology*, 63(3), 395–400.

Jones, John T.; Pelham, Brett W. and Carvallo, Mauricio & Mirenberg, Matthew C. 2004: "How Do I Love Thee? Let Me Count the J:s: Implicit Egotism and Interpersonal Attraction." *Journal of Personality and Social Psychology*, 87(5), 665–683.

Jones, Matt & Love, Bradley C. 2011: "Bayesian Fundamentalism or Enlightenment: On the explanatory status and theoretical contributions of Bayesian models of cognition" *Behavioral and Brain Sciences*, 34(4), 169–231.

Jones, Susan S. & Smith, Linda B. 1993: "The Place of Perception in Children's Concepts" *Cognitive Development*, 8, 113–139.

Juslin, Peter & Persson, Magnus 2002: "PROBabilities from EXemplars (PROBEX): a "lazy" algorithm for probabilistic inference from generic knowledge." *Cognitive Science*, 26, 563–607.

Kahneman, Daniel 2011: *Thinking, Fast and Slow.* New York, NY: Farrah, Straus and Giroux.

Kahneman, Daniel & Klein, Gary 2009: "Conditions for Intuitive Expertice: A Failure to Disagree." *American Psychologist*, 64(6), 515–526.

Karmiloff-Smith, Annette 1992: *Beyond Modularity: A Developmental Perspective on Cognitive Science.* Cambridge, MA: The MIT Press.

Keil, Frank C. 1992: "The Origins of an Autonomous Biology" In Gunnar, Megan R. & Maratsos, Michael (eds.) 1992: *Modularity and Constraints in Language and Cognition, Minnesota Symposia on Child Psychology, volume 25.* New York, NY: Psychology Press, 103–138.

Keil, Frank C. 1993: "The birth and nurturance of concepts by domains: The origins of concepts of living things" in Hirschfeld, Lawrence A. & Gelman, Susan A. 1994: *Mapping the Minds: Domain Specificity in Cognition and Culture.* Cambridge: Cambridge University Press.

Kenward, Ben 2012: "Over-imitating preschoolers believe unnecessary actions are normative and enforce their performance by a third party" *Journal of Experimental Child Psychology*, 112(2), 195–207.

Kenward, Ben; Karlsson, Markus & Persson, Joanna 2011: "Over-imitation is better explained by norm learning than by distorted causal learning" *Proceedings of the Royal Society B: Biological Sciences*, 278(1709), 1239–1246.

Keren, Gideon 2013: "A Tale of Two Systems: A Scientific Advance or a Theoretical Stone Soup? Commentary on Evans & Stanovich (2013)" *Perspectives on Psychological Science*, 8(3), 257–262.

Keupp, Stefanie; Behne, Tanya; Zachow, Joanna; Kasbohm, Alina & Rakoczy, Hannes 2015: "Over-imitation is not automatic: Context sensitivity in children's overimitation and action interpretation of causally irrelevant actions." *Journal of Experimental Child Psychology*, 130, 163–175.

Keysar, Boaz & Bly, Bridget Martin 1999: "Swimming against the current: Do idioms reflect conceptual structure?" *Journal of Pragmatics*, 31(12), 1559–1578.

Klein, Gary 1998: *Sources of Power: How People Make Decisions.* Cambridge, MA: The MIT Press.

Klein, Gary 2008: "Naturalistic Decision Making" *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3), 456–460.

Klein, Gary; Orasanu, Judith; Calderwood, Roberta & Zsambok, Caroline (eds.) 1993: *Decision Making in Action: Models and Methods.* Norwood, NJ: Ablex Publishing Corporation.

Knill, David C. & Pouget, Alexandre 2004: "Bayesian brain: the role of uncertainty in neural coding and computation" *TRENDS in Neurosciences*, 27(1), 712–719.

Knobe, Joshua & Nichols, Shaun 2008: "An Experimental Philosophy Manifesto." In Knobe, Joshua & Nichols, Shaun (eds.) 2008: *Experimental Philosophy.* Oxford: Oxford University Press, 3–14.

Koechlin, Etienne 2014: "An evolutionary computational theory of prefrontal executive function in decision-making" *Philosophical Transactions of The Royal Society B: Biological Sciences*, 369(1655):20130474, doi:10.1098/rstb.2013.0474

Koehler, Jonathan J. 1996: "The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges" *Behavioral and Brain Sciences*, 19(1), 1–53.

Koellner, Peter 2006: "On the Question of Absolute Undecidability" *Philosophia Mathematica*, 14(2), 153–188.

Kosslyn, Stephen Michael 1973: "Scanning visual images: Some structural implications" *Perception & Psychophysics*, 14(1), 90–94.

Kostic, Bogdan; Cleary, Anne M.; Severin, Kaye & Miller, Samuel W. 2010: "Detecting analogical resemblance without retrieving the source analogy" *Psychonomic Bulletin & Review*, 17(3), 405–411.

Kripke, Saul A. 1971: "Identity and Necessity." In Munitz, Milton K. (ed.) 1971: *Identity and Individuation.* New York, NY: New York Unirvesity Press, 135–164.

Kripke, Saul A. 1980: *Naming and Necessity (2nd ed.* Oxford: Blackwell. Originally published in Davidson, Donald & Harman, Gilbert (eds.) 1972: *Semantics of Natural Language.* Dordrecht: D. Reidel, 253–355.

Kroon, Frederick W. 1987: "Causal descriptivism." *Australasian Journal of Philosophy*, 65(1), 1–17.

Kushnir, Tamar; Gopnik, Alison; Lucas, Chris & Schulz, Laura 2010: "Inferring Hidden Causal Structure" *Cognitive Science*, 34(1), 148–160.

Lagnado, David A.; Waldmann, Michael R.; Hagmayer, York & Sloman, Steven A. 2007: "Beyond Covariation: Cues to Causal Structure" in Gopnik, Alison & Schulz, Laura (eds.) 2007: *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Lake, Brenden M.; Salakhutdinov, Ruslan & Tenenbaum, Joshua B. 2015: "Human-level concept learning through probabilistic program induction" *Science*, 350(6266), 1332–1338.

Lakoff, George & Johnson, Mark 1999: *Philosohpy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* New York, NY: Basic Books.

Larkin, Jill H.; McDermott, John; Simon, Dorothea P. & Simon, Herbert A. 1980: "Expert and Novice Performance in Solving Physics Problems" *Science*, 208(4450), 1335–1342.

Lave, Jean 1988: *Cognition in Practice: Mind, mathematics and culture in everyday life.* Cambridge: Cambridge University Press.

Lee, Hee Seung & Holyoak, Keith 2008: "The Role of Causal Models in Analogical Inference" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1111–1122.

Legare, Christine; Wen, Nicole J.; Herrmann, Patricia A. & Whitehouse, Harvey 2015: "Imitative flexibility and the development of cultural learning" *Cognition*, 142, 351–361.

Lin, Emilie L. & Murphy, Gregory L. 1997: "Effects of Background Knowledge on Object Categorization and Part Detection" *Journal of Experimental Psychology: Human Perception and Performance*, 23(4), 1153–1169.

Logan, Gordon D. 1988: "Toward an Instance Theory of Automatization" *Psychological Review*, 95(4), 492–527.

Lu, Hongjing; Rojas, Randall R.; Beckers, Tom & Yuille, Alan 2008a: "Sequential Causal Learning in Humans and Rats" In Love, Bradley C.; McRae, Ken & Sloutsky, Vladimir M. (eds.) 2008: *Proceedings of the 30th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society, 185–190.

Lu, Hongjing; Yuille, Alan L.; Liljeholm, Lili; Cheng, Patricia W. & Holyoak, Keith 2008b: "Bayesian Generic Priors for Causal Learning" *Psychological Review*, 115(4), 955–984.

Luhmann, Christian C. & Ahn, Woo-Kyoung 2007: "BUCKLE: A Model of Unobserved Cause Learning" *Psychological Review*, 114(3), 657–677.

Luhmann, Christian C.; Ahn, Woo-Kyoung & Palmeri, Thomas J. 2006: "Theory-based categorization under speeded conditions" *Memory and Cognition*, 34(5), 1102–1111.

Luria, Aleksandr 1976: *Cognitive Development: Its Cultural and Social Foundations.* Cambridge, MA: Harvard University Press. First published in Russian (1974) and translated into English by Martin Lopez-Morillas and Lynn Solotaroff.

Lyn, Heidi; Russell, Jamie L.; Leavens, David A.; Bard, Kim A.; Boysen, Sarah T.; Schaeffer, Jennifer A. & Hopkins, William D. 2014: "Apes communicate about absent and displaced objects: methodology matters" *Animal Cognition*, 17, 85–94.

Lynch, Elizabeth B.; Coley, John D. & Medin, Douglas L. 2000: "Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices" *Memory & Cognition*, 28(1), 41–50.

Machery, Edouard 2007: "Content empiricism: A methodological critique" *Cognition*, 104, 19–46.

Machery, Edouard 2009: *Doing without Concepts.* Oxford: Oxford University Press.

Machery, Edouard & Seppälä, Selja 2009/2010: "Against Hybrid Theories of Concepts" *Anthropology & Philosophy*, 10(1-2), 99–127.

MacLeod, Miles 2016: "What makes interdisciplinarity difficult? Some consequences of domain specificity in interdisciplinary practice" *Synthese*, 195(2), 697–720.

Maciejovsky, Boris & Budescu, David D. 2007: "Collective Induction Without Cooperation? Learning and Knowledge Transfer in Cooperative Groups and Competitive Auctions" *Journal of Personal and Social Psychology*, 92(5), 854–870.

Mandler, Jean M. 2004: *The Foundations of Mind: Origins of Conceptual Thought.* Oxford: Oxford University Press.

Mansouri, Farshad Alizadeh; Koechlin, Etienne; Rosa, Marcello G. P. & Buckley, Mark J. 2017: "Managing competing goals — a key role for the frontopolar cortex" *Nature Reviews Neuroscience*, 18(11), 645–657.

Marcus, Gary F. 1998: "Rethinking Eliminative Connectionism" *Cognitive Psychology*, 37(3),243–282.

Margolis, Eric 1994: "A reassessment of the shift from the classical theory of concepts to prototype theory" *Cognition*, 51(1), 73–89.

Markman, Arthur B. & Ross, Brian H. 2003: "Category Use and Category Learning" *Psychological Bulletin*, 129(4), 592–613.

Markman, Ellen M. & Hutchinson, Jean E. 1984: "Children's Sensitivity to Constrains on Word Meaning: Taxonomic versus Thematic Relations" *Cognitive Psychology*, 16(1), 1–27.

Matthews, Robert J. 1997: "Can Connectionism Explain Systematicity?" *Mind & Language*, 12(2), 154–177.

McCarthy, John & Hayes, Patrick J. 1969: "Some Philosophical Problems from the Standpoint of Artificial Intelligence" in Meltzer, Bernard & Michie, Donald (eds.) 1969: *Machine Intelligence 4.* Edinburgh: Edinburgh University Press, 463–502.

McClelland, James L.; Rumelhart, David E. & the PDP Research Group 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models.* Cambridge, MA: The MIT Press.

McCulloch, Warren S. & Pitts, Walter 1943: "A Logical Calculus of Ideas Immanent in Nervous Activity" *Bulletin of Mathematical Biophysics*, 5, 115–133.

McIntyre, Ronald 1986: "Husserl and the Representational Theory of Mind" *Topoi*, 5(2), 101–113.

McIntyre, Ronald & Smith, David Woodruff 1989: "Theory of Intentionality" In Mohanty, Jitendra Nath & McKenna, William R. (eds.) 1989: *Husser's Phenomenology: A Textbook*, Washington, D.C.: Center for Advanced Research in Phenomenology & University Press of America, 147–179.

McKenzie, Craig R. M. & Mikkelsen, Laurie A. 2007: "A Bayesian view of covariation assessment" *Cognitive Psychology*, 54(1), 33–61.

Medin, Douglas L. 1989: "Concepts and Conceptual Structure" *American Psychologist*, 44(2), 1469–1481.

Medin, Douglas L.; Coley, John D.; Storms, Gert & Hayes, Brett K. 2003: "A relevance theory of induction" *Psychonomic Bulletin & Review*, 10(3), 517–532.

Medin, Douglas L.; Gerald, Dewey I. & Murphy, Timothy D. 1983: "Relationships Between Item and Category Learning: Evidence That Abstraction Is Not Automatic" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4), 607–625.

Medin, Douglas L. & Schaffer, Marguerite M. 1978: "Context Theory of Classification Learning" *Psychological Review*, 85(3), 207–238.

Medin, Douglas L. & Schwanenflugel, Paula J. 1981: "Linear Separability in Classification Learning" *Journal of Experimental Psychology*, 7(5), 355–368.

Medin, Douglas L. & Shoben, Edward J. 1988: "Context and Structure in Conceptual Combination" *Cognitive Psychology*, 20(2), 158–190.

Meehl, Paul E. 1954: *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Minneapolis, MN: University of Minnesota Press.

Meiran, Nachshon; Chorev, Ziv & Sapir, Ayelet 2000: "Component Processes in Task Switching" *Cognitive Psycholohy*, 41(3), 211–253.

Mercier, Hugo & Sperber, Dan 2011: "Why do humans reason? Arguments for an argumentative theory" *Behavioral and Brain Sciences*, 34(2), 57–74.

Mercier, Hugo & Sperber, Dan 2017: *The Enigma of Reason.* Cambridge, MA: Harvard University Press.

Merleau-Ponty, Maurice 1996: *Phenomenology of Perception.* London: Routledge. Translated by Colin Smith from French original *Phénoménologie de la perception* (1945).

Mervis, Carolyn B.; Catlin, Jack & Rosch, Eleanor 1976: "Relationships among goodness-of-example, category norms, and word frequency" *Bulletin of the Psychonomic Society*, 7(3), 283–284.

Millikan, Ruth Garrett 1989a: "In Defence of Proper Functions" *Philosophy of Science*, 56(2), 288–302.

Millikan, Ruth Garrett 1989b: "Biosemantics" *The Journal of Philosophy*, 86(6), 281–297.

Minsky, Marvin 1974: *A Framework for Representing Knowledge.* MIT-AI Laboratory Memo 306, June, 1974.

Minsky, Marvin & Papert, Seymour A. 1988: *Perceptrons: An introduction to computational geometry (Expanded edition).* Cambridge, MA: The MIT Press.

Miyahara, Katsunori 2011: "Neo-pragmatic intentionality and enactive perception: a copromise between extended and enactive mind" *Phenomenology and the Cognitive Sciences*, 10, 499–519.

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

Morgan, T. J. H.; Uomini, N. T.; Rendell, L. E.; Chouinard-Thuly, L.; Street, S. E.; Lewis, H. M.; Cross, C. P.; Evans, C.; Kearney, R.; de la Torre, I.; Whiten, A. & Laland, K. N. 2015: "Experimental evidence for the co-evolution of hominin tool-making teaching and language" *Nature Communications*, 6:6029, doi: 10.1038/ncomms7029

Moshman, David & Geil, Molly 1998: "Collaborative Reasoning: Evidence for Collective Rationality" *Thinking and Reasoning*, 4(3), 231–248.

Murphy, Gregory L. 2002: *The Big Book of Concepts.* Cambridge, MA: The MIT Press.

Murphy, Gregory L. 2016: "Is there an exemplar theory of concepts?" *Psychonomic Bulletin & Review*, 23(4), 1035–1042.

Murphy, Gregory L. & Allopenna, Paul D. 1994: "The Locus of Knowledge Effects in Concept Learning" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), 904–919.

Murphy, Gregory L. & Medin, Douglas L. 1985: "The Role of Theories in Conceptual Coherence" *Psychological Review*, 92(3), 289–316.

Nersessian, Nancy J. 2008: *Creating Scientific Concepts.* Cambridge, MA: The MIT Press.

Nelson, Katherine 1974: "Concept, World, and Sencence: Interrelations in Acquisition and Development" *Psyhcological Review*, 81(4), 276–285.

Newell, Allen 1980: "Physical Symbol Systems" *Cognitive Science*, 4, 135–183.

Newell, Allen & Simon, Herbert 1976: "Computer Science as Empirical Enquiry: Symbols and Search" *Communications of the ACM*, 19(3), 113–126.

Nisbett, Richard E.; Krantz, David H.; Jepson, Christopher & Kunda, Ziva 1983: "The Use of Statistical Heuristics in Everyday Inductive Reasoning" *Psychological Review*, 90(4), 339–363.

Nisbett, Richard E.; Peng, Kaiping; Choi, Incheol & Norenzayan, Ara 2001: "Culture and Systems of Thought: Holistic Versus Analytic Cognition" *Psychological Review*, 108(2), 291–310.

Nisbett, Richard E. & Wilson, Timothy D. 1977: "Telling More Than We Can Know: Verbal Reports on Mental Processes" *Psychological Review*, 84(3), 231–259.

Noë, Alva 2004: *Action in Perception.* Cambridge, MA: The MIT Press.

Norton, Alec 1995: "Dynamics: An Introduction" (Port & van Gelder, 1995, 46–68)

Nosofsky, Robert 1986: "Attention, Similarity, and the Identification-Categorization Relationship" *Journal of Experimental Psychology: General*, 115(1), 39–57.

Novick, Laura R. 1988: "Analogical Transfer, Problem Similarity, and Expertise" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3), 510–520.

Novick, Laura & Cheng, Patricia 2004: "Assessing Interactive Causal Influence" *Psychological Review*, 111(2), 455–485.

Oaksford, Mike & Chater, Nick 1994: "A Rational Analysis of the Selection Task as Optimal Data Selection" *Psychological Review*, 101(4), 608–631.

Oaksford, Mike & Chater, Nick 1995: "Information gain explains relevance which explains the selection task" *Cognition*, 57, 97–108.

Oaksford, Mike & Chater, Nick 2007: *Bayesian Rationality: The probabilistic approach to human reasoning.* Oxford: Oxford University Press.

Osherson, Daniel N. & Smith, Edward E. 1981: "On the adequacy of prototype theory as a theory of concepts" *Cognition*, 9, 35–58.

Osherson, Daniel N.; Smith, Edward E.; Wilkie, Ormond & López, Alejandro 1990: "Category-Based Induction" *Psychological Review*, 97(2), 185–200.

Over, Harriet & Carpenter, Malinda 2012: "Putting the Social Into Social Learning: Explaining Both Selectivity and Fidelity in Children's Copying Behavior" *Journal of Comparative Psychology*, 126(2), 182–192.

Padoa-Schioppa, Camillo & Schoenbaum, Geoffrey 2015: "Dialogue on economic choice, learning theory, and neuronal representations" *Current Opinion in Behavioral Sciences*, 5, 16–23.

Palmeri, Thomas J. 1997: "Exemplar Similarity and the Development of Automaticity" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(2), 324–354.

Passingham, Richard E. & Wise, Steven P. 2012: *The Neurobiology of the Prefrontal Cotex.* Oxford: Oxford University Press.

Pearl, Judea 1985: "Bayesian networks: A model of self-activated memory for evidential reasoning" Technical report CSD-850017, UCLA Computer Science Department: Los Angeles, CA.

Pearl, Judea 2000: *Causality: Models, Reasoning, and Inference.* Cambridge: Cambridge University Press.

Pearson, Joel; Naselaris, Thomas; Holmes, Emily A. & Kosslyn, Stephen M. 2015: "Mental Imagery: Functional Mechanisms and Clinical Applications" *Trends in Cognitive Sciences*, 19(10), 590–602.

Pendlebury, Michael 1998: "Intentionality and normativity" *South African Journal of Philosohpy*, 17(2), 142–151.

Penn, Derek C. & Povinelli, Daniel J. 2007: "Causal Cognition in Human and Non-human Animals: A Comparative, Critical Review" *Annual Review of Psychology*, 58(1), 97–118.

Peer, Eyal & Gamliel, Eyal 2013: "Heuristics and Biases in Judicial Decisions" *Court Review*, 49, 114–118.

Perlman, Mark 1997: "The Trouble with Two-Factor Conceptual Role Theories." *Minds and Machines*, 7, 495–513.

Petiot, Jean; Varela, Francisco J.; Pachoud, Bernard & Roy, Jean-Michele (eds.) 1999: *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science.* Stanford, CA: Stanford University Press.

Pettit, Philip 1996: *The Common Mind: An Essay on Psychology, Society, and Politics.* Oxford: Oxford University Press.

Pezzulo, Giovanni; Rigoli, Francesco & Friston, Karl J. 2018: "Hierarchical Active Inference: A Theory of Motivated Control" *Trends in Cognitive Science*, 22(4), 294–306.

Piaget, Jean 1952: *The Origins of Intelligence in Children.* New York, NY: International Universities Press. Translated by Margaret Cook from French original *La naissance de l'intelligence chez l'enfant* (1936).

Piccinini, Gualtiero 2004: "Functionalism, computationalim, and mental states." *Studies in History and Philosophy of Science*, 35, 811–833.

Pine, Karen & Messer, David 2003: "The development of representations as children learn about balancing" *British Journal of Developmental Psychology*, 21(2), 285–301.

Platt, Richard D. & Griggs, Richard A. 1993: "Facilitation in the Abstract Selection Task: The Effects of Attentional and Instructional Factors" *The Quarterly Journal of Experimental Psychology*, 46A(4), 591–613.

Port, Robert F. & van Gelder, Timothy (eds.) 1995: *Mind as Motion: Explorations in the Dynamics of Cognition.* Cambridge, MA: The MIT Press.

Posner, Michael I. & Keele, Steven W. 1968: "On the Genesis of Abstract Ideas" *Journal of Experimental Psychology*, 77(3), 353–363.

Powell, Derek; Merrick, M. Alice; Lu, Hongjing & Holyoak, Keith J. 2013: "Generic Priors Yield Competition Between Independently-Occurring Causes" in Knauff, Markus; Pauen, Michael; Sebanz, Natalie & Wachsmuth, Ipke (eds.) 2013: *Cooperative Minds: Social Interaction and Group Dynamics. Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, Austin, TX: Cognitive Science Society, 1157–1162.

Prinz, Jesse J. 2002: *Furnishing the Mind: Concepts and Their Perceptual Basis.* Cambridge, MA: The MIT Press.

Psillos, Stathis 2012: "Causal descriptivism and the reference of theoretical terms" In Raftopoulos, Athanassios & Machamer, Peter (eds.) 2012: *Perception, Realism, and the Problem of Reference.* Cambridge: Cambridge University Press, 212–238.

Putnam, Hilary 1975: "The Meaning of 'Meaning' " In Gunderson, Keith (ed.) 1975: *Minnesota Studies in the Philosophy of Science, vol. VII: Language, Mind, and Knowledge.* Minneapolis, MN: University of Minnesota Press, 131–193.

Pylyshyn, Zenon 1981: "The Imagery Debate: Analogue Media Versus Tacit Knowledge" *Psychological Review*, 88(1), 16–45.

Pöyhönen, Samuli 2013: "Natural Kinds and Concept Eliminativism" In Karakostas, Vassilios & Dieks, Dennis (eds.) 2013: *EPSA11 Perspectives and foundational Problems in Philosophy of Science.* Cham: Springer, 167–179.

Quine, Willard V. O. 1960: *Word and Object.* Cambridge, MA: The MIT Press.

Rapaport, William J. 1995: "Understanding Understanding: Syntactic Semantics and Computational Cognition" *Philosophical Perspectives*, 9: *Connectionism and Philosophical Psychology*, 49–88.

Reber, Arthur S. 1989: "Implicit Learning and Tacit Knowledge" *Journal of Experimental Psychology: General*, 118(3), 219–235.

Reber, Arthur S. 1993: *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious.* Oxford: Oxford University Press.

Reeves, Lauretta M. & Weisberg, Robert W. 1994: "The Role of Content and Abstract Information in Analogical Transfer" *Psychological Bulletin*, 115(3), 381–400.

Rehder, Bob 2003: "A Causal-Model Theory of Conceptual Representation and Categorization" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1141–1159.

Rehder, Bob 2007: "Essentialism as a generative theory of classification" In Gopnik, Alison & Schulz, Laura (eds.) 2007: *Causal learning: Psychology, philosophy, and computation*, Oxford: Oxford University Press, 190–207.

Rehder, Bob 2009: "Causal-Based Property Generalization" *Cognitive Science*, 33(3), 301–344.

Rehder, Bob & Kim, ShinWoo 2010: "Causal Status and Coherence in Causal-Based Categorization" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(5), 1171–1206.

Rey, Georges 1983: "Concepts and stereotypes" *Cognition*, 15(1–3), 237–262.

Rifkin, Anthony 1985: "Evidence for a basic level in event taxonomies" *Memory & Cognition*, 13(6), 538–556.

Rips, Lance J. 1989: "Similarity, typicality, and categorization" in Vosniadou, Stella & Ortony, Andrew (eds.) 1989: *Similarity and analogical reasoning.* Cambridge: Cambridge University Press, 21–59.

Rips, Lance J.; Shoben, Edward J. & Smith, Edward E. 1973: "Semantic Distance and the Verification of Semantic Relations" *Journal of Verbal Learning and Verbal Behavior*, 12, 1–20.

Rips, Lance J.; Smith, Edward E. & Medin, Douglas L. 2012: "Concepts and Categories: Memory, Meaning, and Metaphysics" In Holyoak, Keith J. & Morrison, Robert G. (eds.) 2012: *The Oxford Handbook of Thinking and Reasoning*, Oxford: Oxford University Press, 177–209.

Rorty, Richard 1991: *Objectivity, Relativism, and Truth: Philosophical Papers, Volume 1.* Cambridge: Cambridge University Press.

Rosch, Eleanor 1973: "Natural Categories" *Cognitive Psychology*, 4(3), 328–350.

Rosch, Eleanor 1975: "Cognitive Representations of Semantic Categories" *Journal of Experimental Psychology*, 104(3), 192–233.

Rosch, Eleanor 1978: "Principles of Categorization" in Rosch, Eleanor & Lloyd, Barbara B. (eds.) 1978: *Cognition and Categorization.* Hillsdale, NJ: Lawrence Erlbaum Associates, 27–48.

Rosch, Eleanor 1999: "Reclaiming concepts" *Journal of Consciousness Studies*, 6(11-12), 61–77.

Rosch, Eleanor E. & Mervis, Carolyn B. 1975: "Family Resemblances: Studies in the Internal Structure of Categories" *Cognitive Psychology*, 7(4), 573–605.

Rosch, Eleanor; Mervis, Carolyn B.; Gray, Wayne D. Johnson, David M. & Boyes-Braem, Penny 1976: "Basic Objects in Natural Categories" *Cognitive Psychology*, 8(3), 382–439.

Rosenblatt, Frank 1958: "The Perceptron: A probabilistic model for information storage and organization in the brain" *Psychological Review*, 65(6), 386–408.

Rosenblatt, Frank 1962: *Principles of Neurodynamics: Perceptrons and the theory of brain mechanisms.* Washington, D.C.: Spartan Books.

Ross, Brian H. 1984: "Remindings and Their Effects in Learning a Cognitive Skill" *Cognitive Psychology*, 16(3), 371–416.

Ross, Brian H. & Kennedy, Patrick T. 1990: "Generalizing From the Use of Earlier Examples in Problem Solving" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1), 42–55.

Ross, Brian H.; Perkins, Susan J. & Tenpenny, Patricia L. 1990: "Reminding-Based Category Learning" *Cognitive Psychology*, 22(4), 460–492.

Ross, Brian H.; Wang, Ranxiano Frances; Kramer, Arthur F.; Simons, Daniel J. & Crowell, James A. 2007: "Action information from classification learning" *Psychonomic Bulletin & Review*, 14(3), 500–504.

Rumelhart, David E.; McClelland, James L. & the PDP Research Group 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations.* Cambridge, MA: The MIT Press.

Rumelhart, David E.; Smolensky, Paul; McClelland, James L. & Hinton, Geoffrey E. 1986: "Schemata and Sequential Thought Processes in PDP Models." In McClelland, James L.; Rumelhart, David E. & The PDP Research Group (eds.) 1986: *Parallel Distributed Processing. Explorations in the Microstructure of Cognition, volume 2: Psychological and Biological Models.* Cambridge, MA: The MIT Press, 7–57.

Rupert, Robert D. 1999: "The Best Test Theory of Extension: First Principle(s)" *Mind & Language*, 14(3), 321–355.

Russell, Stuart & Norvig, Peter 2010: *Artificial Intelligence: A Modern Approach (3rd ed.)* Upper Saddle River, NJ: Pearson.

Ryle, Gilbert 1949: *The Concept of Mind.* London: Hutchinsons University Library.

Sachs, Carl B. 2014: *Intentionality and the Myths of the Given: Between Pragmatism and Phenomenology.* London: Pickering & Chatto.

Sakai, Katsuyuki 2008: "Task Set and Prefrontal Cortex" *Annual Review of Neuroscience*, 31, 219–245.

Schneider, Walter & Shiffrin, Richard M. 1977: "Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention" *Psychological Review*, 84(1), 1–66.

Schopenhauer, Arthur 1966: *The World as Will and Representation, Volume II.* Translated by E. F. J. Payne from German original *Die Welt als Wille und Vorstellung* (1818).

Searle, John R. 1980: "Minds, brains, and programs" *The Behavioral and Brain Sciences*, 3(3), 417–457.

Seger, Carol A. 1994: "Implicit Learning" *Psychological Bulletin*, 115(2), 163–196.

Sellars, Wilfrid 1956: "Empiricsm and the Philosophy of Mind"In Feigl, Herbert & Scriven, Michael (eds.) 1956: *Minnesota Studies in the Philosohpy of Science, volume I: The Foundations of Science and the Concepts of Psycholohy and Psychoanalysis.* Minneapolis, MI: University of Minnesota Press, 253–329.

Shafir, Eldar B.; Smith, Edward E. & Osherson, Daniel N. 1990: "Typicality and reasoning fallacies" *Memory & Cognition*, 18(3), 229–239.

Shepard, Roger N. & Metzler, Jacqueline 1971: "Mental Rotation of Three–Dimensional Objects" *Science*, 171(3972), 701–703.

Shiffrin, Richard M. & Schneider, Walter 1977: "Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending, and a General Theory" *Psychological Review*, 84(2), 127–190.

Siegelmann, Have T. & Sontag, Eduardo D. 1991: "Turing Computability of Neural Nets" *Applied Mathematics Letters*, 4(6), 77–80.

Siegelmann, Hava T. & Sontag, Eduardo D. 1995: "On the Computational Power of Neural Nets" *Journal of Computer and System Sciences*, 50(1), 132–150.

Simon, Herbert A. 1955: "A Behavioral Model of Rational Choice" *Quarterly Journal of Economics*, 69(1), 99–118.

Simon, Herbert A. 1992: "What is an "Explanation" of Behavior?" *Psychological Science*, 3(3), 150–161.

Skinner, Burrhus F. 1965: *Science and Human Behavior (New impression edition).* New York, NY: Free Press. First edition published in 1953.

Sloman, Steven A. 1996: "The Empirical Case for Two Systems of Reasoning" *Psychological Bulletin*, 119(1), 3–22.

Sloman, Steven A. 1998: "Categorical Inference Is Not a Tree: The Myth of Inheritance Hierarchies" *Cognitive Psychology*, 35(1), 1–33.

Sloman, Steven A. 2005: *Causal Models: How People Think About the World and Its Alternatives.* Oxford: Oxford University Press.

Sloman, Steven A. & Malt, Barbara C. 2003: "Artifacts are not ascribed essences, nor are they treated as belonging to kinds" *Language and Cognitive Processes*, 18(5/6), 563–582.

Smith, Eliot R. & Collins, Elizabeth C. 2009: "Dual-process models: A social psychological perspective" (Evans, 2009, 197–216) In Evans, Jonathan St. B. T. & Frankish, Keith 2009: *In Two Minds: Dual Processes and Beyond*

Smith, Elior R. & DeCoster, Jamie 2000: "Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems" *Personality and Social Psychology Review*, 4(2), 108–131.

Smith, David Woodruff & McIntyre, Ronald 1982: *Husserl and Intentionality: A Study of Mind, Meaning, and Language.* Dordrecht: D. Reidel.

Smith, J. David & Minda, John Paul 1998: "Prototypes in the Mist: The Early Epochs of Category Learning" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6), 1411–1436.

Smith, J. David & Minda, John Paul 2000: "Thirty Categorization Results is Search of a Model" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1), 3–27.

Smith, J. David; Murray, Morgan J. Jr. & Minda, John Paul: "Straight Talk About Linear Separability" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(3), 659–680.

Smith, Edward E. & Osherson, Daniel N. 1984: "Conceptual Combination with Prototype Concepts" *Cognitive Science*, 8, 337–361.

Smith, Edward E. & Sloman, Steven A. 1994: "Similarity- versus rule-based categorization" *Memory & Cognition*, 22(4), 377–386.

Sobel, David M. & Kirkham, Natasha Z. 2007: "Bayes nets and babies: infant's developing statistical reasoning abilities and their representation of causal knowledge" *Developmental Science*, 10(3), 298–306.

Socher, Richard; Bauer, John; Manning, Christopher D. & Ng, Andrew Y. 2013: "Parsing with Compositional Vector Grammars" *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 455–465.

Sperber, Dan; Cara, Francesco & Girotto, Vittorio 1995: "Relevance theory explains the selection task" *Cognition*, 57, 31–95.

Sperber, Dan & Girotto, Vittorio 2002: "Use or misuse of the selection task? Rejoinder to Fiddick, Cosmides, and Tooby." *Cognition*, 85(3), 277–290.

Sripada, Chandra Sekhar & Stich Stephen 2007: "A framework for the psychology of norms" In Carruthers, Peter; Laurence, Stephen & Stich, Stephen 2007: *Innateness and the Structure of the Mind, Vol. II.* Oxford: Oxford University Press, 280–301.

Stalnaker, Robert 1976: "Propositions." In MacKay, Alfred & Merrill, Daniel (eds.) 1992: *Issues in the Philosophy of Language.* New Haven: Yale University Press, 76–91.

Stanovich, Keith E. 1999: *Who Is Rational? Studies of Individual Differences in Reasoning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Stanovich, Keith E. & West, Richard F. 1998: "Cognitive Ability and Variation in Selection Task Performance" *Thinking and Reasoning*, 4(3), 193–230.

Stanovich, Keith E. & West, Richard F. 2000: "Individual differences in reasoning: Implications for the rationality debate?" *Behavioral and Brain Sciences*, 23(5), 645–726.

Stenning, Keith & van Lambalgen, Michiel 2001: "Semantics as a Foundation for Psychology: A Case Study of Wason's Selection Task" *Journal of Logic, Language and Information*, 10(3), 273–317.

Stenning, Keith & van Lambalgen, Michiel 2004: "A little logic goes a long way: basing experiment on semantic theory in the cognitive science of conditional reasoning" *Cognitive Science*, 28(4), 481–529.

Steyvers, Mark; Tenenbaum, Joshua B.; Wagenmakers, Eric-Jan & Blum, Ben 2003: "Inferring causal networks from observations and interventions" *Cognitive Science*, 27(3), 453–489.

Stich, Stephen 1983: *From Folk Psychology to Cognitive Science: The Case Against Belief.* Cambridge, MA: The MIT Press.

Stich, Stephen 1996: *Deconstructing the Mind.* Oxford: Oxford University Press.

Stout, Dietrich & Chaminade, Thierry 2012: "Stone tools, language and the brain in human evolution" *Philosophical Transactions of The Royal Society*, 367, 75–87.

Sun, Ron 2002: *Duality of the Mind: A Bottom Up Approach Toward Cognition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sun, Ron; Slusarz, Paul & Terry, Chris 2005: "The Interaction of the Explicit and the Implicit in Skill Learning: A Dual-Process Approach" *Psychological Review*, 112(1), 159–192.

Tanaka, James W. & Taylor, Marjorie 1991: "Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder?" *Cognitive Psychology*, 23(3), 457–482.

Tarski, Alfred & Vaught, Robert L. 1957: "Arithmetical Extensions of Relational Systems" *Compositio Mathematica*, 13(2), 81–102.

Tenenbaum, Joshua B. & Griffiths, Thomas L. 2001: "Structure learning in human causal induction" in Leen, Todd K.; Dietterich, Thomas G. & Tresp, Volker (eds.) 2001: *Advances in Neural Information Processing Systems 13.* Cambridge, MA: The MIT Press, 59–65.

Tenenbaum Joshua B.; Griffiths, Thomas L. & Niyogi, Sourabh 2007: "Intuitive Theories as Grammars for Causal Inference" In Gopnik, Alison & Schulz, Laura (eds.) 2007: *Causal Learning: Psychology, Philosophy, and Computation.* Oxford: Oxford University Press, 301–322.

Thorndike, Edward L. 1922: "The Effect of Changed Data Upon Reasoning" *Journal of Experimental Psychology*, 5(1), 33–38.

Toates, Frederick 2006: "A model of the hierarchy of behaviour, cognition, and consciousness" *Consciousness and Cognition*, 15(1), 75–118.

Tolman, Edward C. 1932: *Purposive Behavior in Animals and Men.* London: The Century Co.

Thompson, Evan 2007: *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge, MA: Harvard University Press.

Tomasello, Michael 2009: "The usage-based theory of language acquisition" in Bavin, Edith (ed.) 2009: *The Cambridge Handbook of Child Language.* Cambridge: Cambridge Univeristy Press, 69-88.

Tomasello, Michael; Carpenter, Malinda; Call, Joseph; Behne, Tanya & Moll, Henrike 2005: "Understanding and sharing intentions: The origins of cultural cognition" *Behavioral and Brain Sciences*, 28(5), 675–735.

Turnbull, William & Carpendale, Jeremy I. M. 1999 "A Social Pragmatic Model of Talk: Implications for Research on the Development of Children's Social Understanding" *Human Development*, 42(6), 328–355.

Tversky, Amos 1977: "Features of Similarity" *Psychological Revies*, 84(4), 327–352.

Tversky, Amos & Kahneman, Daniel 1971: "Belief in the law of small numbers" *Psychological Bulletin*, 76(2), 105–110.

Tversky, Amos & Kahneman, Daniel 1983: "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment" *Psychological Review*, 90(4), 293–315.

Tweney, Ryan D. & Yachanin, Stephen A. 1985: "Can Scientists Rationally Assess Conditional Inferences?" *Social Studies of Science*, 15(1), 155–173.

Uddén, Julia; Araújo, Susana; Forkstam, Christian; Ingvar, Martin; Hagoort, Peter & Peterson, Karl Magnus 2009: "A Matter of Time: Implicit Acquisition of Recursive Sequence Structure" in Taatgen, Niels & van Rijn, Hedderik (eds.) 2009: *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society, 2444–2449.

von Uexküll, Jakob 1926: *Theoretical Biology.* New York, NY: Harcourt, Brace & Company. Translated by D. L. Mackinnon from German original "Theoretische Biologie" (1920).

Uomini, Natalie Thaïs & Meyer, Georg Friedrich 2013: "Shared Brain Lateralization Patterns in Language and Acheulean Stone Tool Production: A Functional Transcranial Doppler Ultrasound Study." *PLoS ONE*, 8(8), e72693, 1–9.

Vallacher, Robin R. & Wegner, Daniel M. 1987: "What Do People Think They're Doing? Action Identification and Human Behavior" *Psychological Review*, 94(1), 3–15.

Varela, Francisco J. 1992: *Ethical Know-How: Action, Wisdom, and Cognition.* Stanford, CA: Stanford University Press. Translated by the Board of Trustees of the Leland Stanford Junior University from Italian original *Un Know-How per l'Etica* (1992).

Varela, Francisco J.; Thompson, Evan & Rosch, Eleanor 1991: *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge, MA: The MIT Press.

Vygotsky, Lev 1986: *Thought and Language (Revised edition).* Cambridge, MA: The MIT Press. Translated and edited by Alex Kozulin from Russian original *Myshlenie i rech* (1934).

Waldmann, Michael R. (ed.) 2017: *The Oxford Handbook of Causal Learning.* Oxford: Oxford University Press.

Wang, Jing; Conder, Julie A.; Blitzer, David N. & Shinkareva, Svetlana V. 2011: "Neural Representation of Abstract and Concrete Concepts: A Meta-Analysis of Neuroimaging Studies" *Human Brain Mapping*, 31(10), 1459–1468.

Wason, Peter C. 1966: "Reasoning" in Foss, Brian M. (ed.) 1966: *New Horizons in Psychology 1.* Harmondsworth: Penguin, pp. 135–155.

Wason, Peter C. & Evans, Jonathan St. B. T. 1975: "Dual processes in reasoning?" *Cognition*, 3(2), 141–154.

Wason, Peter C. & Shapiro, Diana 1971: "Natural and Contrived Experience in a Reasoning Problem." *Quarterly Journal of Experimental Psychology*, 23, 63–71.

Wattenmaker, William D.; Dewey, Gerald I.; Murphy, Timothy D. & Medin, Douglas, L. 1986: "Linear Separability and Concept Learning: Context, Relational Properties, and Concept Naturalness" *Cognitive Psychology*, 18(2), 158–194.

Whitehead, Alfred N. 1911: *An Introduction to Mathematics.* London: Williams & Norgate.

Wilson, Timothy D. 2002: *Strangers to Ourselves: Discovering the Adaptive Unconscious.* Cambridge, MA: The Belknap Press.

Wittgenstein, Ludwig 1953: *Philosophical Investigations.* Oxford: Blackwell. Translated from German manuscripts by G. E. M. Anscombe and edited with Rush Rhees.

Wright, J. Edward 2000: *The Early History of Heaven.* Oxford: Oxford University Press.

Zahavi, Dan 2004: "Phenomenology and the project of naturalization." *Phenomenology and the Cognitive Sciences*, 3, 331–347.