Otto-Friedrich-Universität Bamberg

Symbol Grounding as the Generation of Mental Representations

Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.), eingereicht an der
Fakultät Wirtschaftsinformatik und Angewandte Informatik der
Otto-Friedrich-Universität Bamberg

von:	Mark Wernsdorfer
Adresse:	Untere Königstrasse 36 96052 Bamberg
Erstgutachter:	Prof. Dr. Ute Schmid
Externer Zweitgutachter:	Prof. Claes Strannegård, PhD. (University of Gothenburg)
Drittes Kommissionsmitglied:	Prof. Michael Mendler, PhD.

Erklärung

Erklärung gemäß §10 der Promotionsordnung der Fakultät Wirtschaftsinformatik und Angewandte Informatik an der Otto-Friedrich-Universität Bamberg:

- Ich erkläre, dass ich die vorgelegte Dissertation selbständig, das heißt auch ohne die Hilfe einer Promotionsberaterin bzw. eines Promotionsberaters angefertigt habe und dabei keine anderen Hilfsmittel als die im Literaturverzeichnis genannten benutzt und alle aus Quellen und Literatur wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

- Ich versichere, dass die Dissertation oder wesentliche Teile derselben nicht bereits einer anderen Prüfungsbehörde zur Erlangung des Doktorgrades vorlagen.

- Ich erkläre, dass diese Arbeit noch nicht in ihrer Gesamtheit publiziert ist. Soweit Teile dieser Arbeit bereits in Konferenzbänden und Journals publiziert sind, ist dies an entsprechender Stelle kenntlich gemacht und die Beiträge sind im Literaturverzeichnis aufgeführt.

Abstract

This thesis deals with the automatic and semantically autonomous construction of the mental representations of an agent—so-called *'symbol grounding'*.

How can a system perform an independent semantic interpretation of its sensorimotor data, that is not just an imitation of the semantics in the head of its designer? The ability to do so is a prerequisite for general learning in unknown environments. Previous approaches try to achieve this in three different ways: by simulating a sufficiently complex biological brain (anatomically motivated), by simulating and combining functional modules of the human psyche (psychologically motivated), and by identifying a fundamental algorithm that enables different types of learning in the same way (holistically motivated).

This work follows the third approach and draws its inspiration from modern *phenomenology,* theories of *embodied cognition, semiotics,* and methods of *machine learning.* Previous approaches to the dynamic generation of representations are presented. After that a new approach in the field of *reinforcement learning* is worked out. Physically present aspects of the environment are captured as sensorimotor activations within a system so that their occurrence can be predicted probabilistically. This is implemented within the theoretical framework of conditional probabilities according to Bayes with an extension for the identification of hierarchical structures in the environment.

It can be shown, on the one hand, that a hierarchical approach exceeds previous methods for sequence prediction. On the other hand, it allows a differentiation of different subsequences and it allows their representation and the modification of these representations at runtime. The possibilities and limitations of the developed algorithm are illustrated and evaluated on the basis of various experiments.

Keywords

Symbol Grounding, Phenomenology, Embodied Cognition, Intentionality, Mental Model, Cognitive Model

Zusammenfassung

Die Arbeit befasst sich mit der automatischen und semantisch autarken Konstruktion mentaler Repräsentationen eines Agenten – dem so genannten *„Symbol Grounding“*.

Wie kann ein System seine sensomotorischen Daten selbständig semantisch interpretieren, ohne die Semantik im Kopf seines Entwicklers zu bemühen? Diese Fähigkeit ist Voraussetzung für generelles Lernen in unbekannten Umgebungen. Bisherige Ansätze versuchen dies auf drei Arten zu erreichen: durch die Simulation eines hinreichend komplexen biologischen Gehirns (anatomisch motiviert), durch die Nachbildung und Kombination funktionaler Module der menschlichen Psyche (psychologisch motiviert) und durch die Identifikation eines grundlegenden Algorithmus der auf die gleiche Weise unterschiedliche Arten des Lernens ermöglicht (holistisch motiviert).

Diese Arbeit folgt dem dritten Ansatz und zieht als Inspiration *Phänomenologie,* Theorien der *Embodied Cognition, Semiotik* sowie Methoden des *Machine Learning* heran. Bisherige Ansätze zur dynamischen Generierung von Repräsentationen werden vorgestellt. Daraufhin wird der eigene Ansatz aus dem Bereich des *Reinforcement Learning* ausgearbeitet. Aspekte der Umwelt werden als sensomotorische Aktivierungen erfasst, so dass ihr Auftreten probabilistisch vorhergesagt werden kann. Informatisch umgesetzt wird dies im theoretischen Rahmen konditionaler Wahrscheinlichkeiten nach Bayes mit einer Erweiterung zur Erfassung hierarchischer Strukturen.

Es kann gezeigt werden, dass ein hierarchischer Ansatz bisherige Methoden der Sequenzprognose übertrifft. Zum Anderen ermöglicht er eine Differenzierung verschiedener Teilsequenzen, für welche Repräsentationen dynamisch zur Laufzeit erstellt und modifiziert werden. Die Möglichkeiten und Grenzen des entwickelten Algorithmus werden anhand verschiedener Experimente dargestellt und evaluiert.

Keywords

Symbol Grounding, Phenomenology, Embodied Cognition, Intentionality, Mental Model, Cognitive Model

Acknowledgements

Contents

List of Figures

List of Tables

List of Algorithms

Over thinking, over analyzing separates the body from the mind.
Withering my intuition leaving all these opportunities behind.

Justin Chancellor

1. Introduction

In conversations about artificial intelligence, the common sentiment is that the concepts of such a system are fundamentally limited by the concepts of its designer. If complex concepts are composed of more basic concepts, so the reasoning, and the most basic concepts of an artificial system are provided by its designer, then the designer's limitations 'carry over' to all the concepts the system can ever have.

The designer's *human* concepts are useful for systems with a human *body,* but probably not optimal to solve *every* problem and certainly problematic for a system on wheels and with infra-red vision. It appears paradoxical to suggest the design of a system that is independent from the concepts of its designer.

A way out of this paradox is represented by *artificial general intelligence* which is supposed to kick off a technological 'singularity', where one artificial system develops a better one *completely autonomously,* without the involvement of human designers and their limited concepts. Eventually, such a development is supposed to result in an explosion in artificial intelligence that rapidly exceeds human abilities.

Symptomatic for the field of artificial general intelligence is, however, that it ignores potentially fundamental conceptual limitations of artificial systems. If the initial artificial intelligence system (i.e. the 'seed AI') is developed from essentially *human* concepts, how can artefacts of this system ever be supposed to exceed this foundation?

1.1. Hypothesis

Any intelligent system must generate its most basic concepts *autonomously.* To establish such a system requires a lot of philosophical groundwork that mostly revolves around the following question: can *artificial* concepts be *actual* concepts?

This question touches the philosophy of mind, semiotics, the cognitive sciences, and, last but not least, machine learning. To provide an answer is not only of theoretical use. In fact it can provide practical, workable systems.

A system that creates *its own concepts* would be completely independent from human concepts that might require a human body and the according sensorimotor apparatus.

It could learn to solve problems by interacting with environments that the human mind cannot even conceive of.

To describe and, eventually, to simulate the initial generation of concepts in a cognitive system, however, requires a theoretical foundation. How do cognitive systems represent real things with their minds? Is this relationship the result of a passive discovery or an active construction of reality? How does the transition between reality and mind take place, what are its conditions, and what is lost in translation?

Independent from its achievements, no computer system will ever be conceded 'real' mental content without an explanation for how this is supposed to be possible. Without answers to the philosophical questions at play, any attempt to develop truly autonomous artificial intelligence could be nipped in the bud as essentially 'non-cognitive'.

A solution to artificial concept creation must therefore be twofold. On the one hand, it must provide a *theoretical* explanation for how to ground the mental content of cognitive systems in their interaction with external reality. On the other hand, this theory must also be translated into a *workable* system that surpasses similar systems in some task that is relevant to cognitive systems.

The success of the second part *as a simulation for cognitive systems* requires that the first part is accepted *as a theoretical explanation for basic concepts.*

1.2. Premises

The present approach has two particular premises that make it unique. Both are necessary to obtain *grounded symbolic mental representations* and both follow from a *phenomenological* view on cognition.

Phenomenology is a family of philosophical theories on the structure of subjective experience and consciousness. It investigates how reality appears to cognitive systems and how these appearances interact to generate a more complex understanding of *the world.* (Kockelmans 1999)

In contrast to the natural sciences, phenomenology is not based in elements that are publicly accessible (e.g. the idea of *a physical atom)* but in the basic elements of subjective experience (e.g. *the redness* of a flower)— also referred to as 'qualia' or 'basic perception'.

These elements can be *experienced* by cognitive systems. In contrast to elements in the natural sciences, however, they cannot be described *entirely.* No description can replicate the *feeling* during the actual perception of a red flower.

This means that the system that *has* this perception as a whole also cannot be described entirely. No observer has access to another system's first-person-perspective. One cognitive system cannot fully take the perspective of another. Despite this inaccessibility, however, the dynamics and relations between these unobservable entities can still be described and simulated.

The question whether the *simulating* entities are also *what they ought to simulate* will not, and cannot, be answered here. This question cannot be answered in principle and, even more so, an answer would render most endeavours to truly autonomous concept creation in artificial systems futile.

The present approach rather presents an extensive argumentation for premises that are necessary for such a simulation. The value of these premises is put to the test by implementing a system according to them and comparing it to related systems that solve similar problems in a level playing field.

1.2.1. First-person-perspective

The first premise is to simulate *the observer's* first-person-perspective, not the observations that they might make of any other system.[1]

A simulation for the generation of concepts in cognitive systems requires a cognitive model of these systems. Scientific models are mostly determined by the subject matter of their domain. Accordingly, dissent concerning the appropriateness of a scientific model is often a reliable indicator for an ambiguous definition of, or even a shift in, subject matter.

The subject matter of physics, for example, is quantified matter and energy. Accordingly, the interpretation of physical models implies to accept numeric shapes as representations for real amounts of matter and energy.

The subject matter of natural sciences is relatively well defined. Therefore, empiric evidence can support or weaken natural models quite straightforwardly. A model is supported by empiric evidence simply if an object exists such as the model describes it. The model is inappropriate if no such object exists.

A model of radioactivity, for example, can be evaluated by determining whether objects exist whose matter decays in a way such as the model describes it.

In cognitive sciences, models cannot be evaluated as easily. Obviously, the subject matter of cognitive sciences is cognition. In contrast to natural objects, however, cognition can be described from two equally legitimate but essentially incompatible perspectives—each of which implies its own semantics.

The fact that the appropriateness of a cognitive model depends on *the perspective* of the particular interpreter makes it much harder to incorporate empiric evidence. What is

[1] In fact, to still use the term 'observer' only makes sense when the intention is to emphasise a contrast against third-person-perspective.

more relevant? The system's own perspective or the perspective of an external observer? A model may be veridical for one but not for the other.

Therefore, it is crucial to make this perspective abundantly: the following cognitive model is generated from *first-person-perspective* and is based on the impressions of *subjective experience.*

A cognitive model from first-person-perspective cannot be naturalised. Subjective experience is premise to such a model like matter and energy are premise to a physical model.

1.2.2. Phenomenal Content in Basic Perception

The second premise is that there is phenomenal content in basic perception.

Cognitive systems usually feature a mental model of their environment. An accurate cognitive model, therefore, must also be *the model of a mental model.* From first-person-perspective, the mental model of a real cognitive systems is composed of phenomenal shapes that represent parts of external reality. The most basic of these shapes are *basic perceptions.* To the system, its basic perception carries a particular *phenomenal content.*

Phenomenal content is a part of subjective experience. The perspective in subjective experience is *irreducible.* John Searle makes this point, and that Charles Peirce and Franz Brentano made it before him.

If mental models ground in phenomenal content and phenomenal content is perceivable only from first-person-perspective, then another system's mental model is essentially *unobservable.* Scepticism concerning the possibility of artificial concept creation is nurtured from reasonable doubts about such an *axiomatic* justification of unobservable content.

To its own system, however, phenomenal content is true *exactly and only* because *it cannot be different from how it appears.* The fact that phenomenal content can *only* be justified axiomatically is crucial for the grounding of mental content.

The axiomatic justification of phenomenal content can only appear dubious to an outside observer for whom the content is simply *not* how it appears. To an observer, *another* system's phenomenal content must consist of the observer's *own* basic perception. How else could they perceive it in the first place?

The concepts of a system are always constrained to phenomenal shapes in its own structure. These shapes are *the only way* for the system to conceive of the world. This limitation, and the conceivability of this limitation for anyone *but* this system, is undisputed in real cognitive systems.

If this is true for the content of real cognitive systems, however, then it must also be true for the content in an accurate computational simulation of it. To concede unobservable phenomenal content to biological systems but not to computational ones lacks justification and hinders progress in computational symbol grounding.

Simulating autonomous concept creation requires to concede phenomenal content to computational systems. In the following, two arguments in support of this statement are established. The first argument is that *the domain of cognition is first-person-perspective* and the second is that *phenomenal content is impossible to observe.*

1.3. Goals

Our premises allow to postulate specific phenomenological entities and processes in a computational simulation. The goals of this project can only be achieved if these premises are accepted.

The theoretical goal of this work is to open another route to simulating cognition. One that is independent from the anatomical or even physical particularities of cognitive systems as we know them.

It must remain unclear whether a simulation is *actually* cognitive and its observable processes *really* necessary for the generation of a real mental model. For the same reasons, it must also remain unclear whether this is the case in *any* other system beside oneself.

The practical goal, eventually, is to provide such a simulation and use it to solve particular problems in machine learning. According to the evaluation at the end of this work, such systems are practically relevant, but the problems that they solve have so far led a rather shadowy existence in research.

1.3.1. Theoretical Goal

The theoretical goal of this work is not to determine whether machines have a mind. Instead, a general argument for the impossibility of observing the mind of another system is to be established. This argument is based on a comprehensive presentation of the philosophical concept of *intentionality.*

From this presentation, specifications are derived that meet some of the conditions for what critics of artificial intelligence call 'intentional content'. This content is reduced back to phenomenal appearances and these in turn to a stream of unconscious information exchange between the system and its environment. Such an approach has strong roots in the field of embodied cognition and goes way back to Charles Peirce and Immanuel Kant.

There are two main reasons why approaches to the problem of artificial concept creation have not been accepted from philosophical points of view.

The first reason is that artificial intelligence research has applied a soft interpretation to this problem. In the following, it is instead presented as a hard problem and it is pointed out why solutions for a soft interpretation do not help to solve the hard part.

The second reason is a conflation on the philosophical side between computer systems and their formal description. The limitations of formal symbol manipulation are actually only restrictions for *the description* of a system, but not for the system itself.

These two points result from two incompatible perspectives on *mental representation.* The former is mostly taken by the cognitive sciences (e.g. Harnad 1990; Steels 2008). The latter is mostly taken by the philosophy of mind (e.g. Searle 1993; Sun 2000). These two perspectives are presented in detail and their influence on conceptions of the problem is shown.

The explicit and exemplary presentation of these points makes it possible to better understand the terminological incompatibilities between the philosophy of mind and artificial intelligence research, and to overcome them in the future.

1.3.2. Practical Goal

The practical goal is the simulation of a cognitive system. This goal is achieved if the resulting system satisfies the three following requirements.

Firstly, the system must be able to subdivide any real environment into discrete representations. These representations are not necessarily unique for a particular segment of external reality and these segments are not unique for a particular representation. This is due to the fact that a single mental representation can refer to different external referents and that a single referent can be mentally represented in different ways depending on the context.

The environment 'enters' cognitive systems as a sequence of information. This sequence is segmented, the segments are combined into *objects,* and the objects are mentally represented by phenomenal shapes.

Despite cultural or individual differences, different people seem to be carrying out these processes very similarly. As a result, knowledge about the world coincides among different minds.[2]

Secondly, these representations must enable the system to better predict events in the environment. Cognitive systems derive expectations about future observations from their mental models. Specialised models for specific environments allow the system to predict future events with high accuracy.

[2] This epistemological agreement is also due to the fact that we have access to the same external reality. Without already having a similar physical and neurological set-up, however, this would not be possible.

Cognitive systems perceive the world deterministically: apparently random events are explained through a hidden state of the world and this state is mentally represented in the system's mental model. In general, cognitive systems counteract uncertainty by introducing new representations to explain unexpected events.

Thirdly, these representations must enable the system to better perform goal-directed interaction. In real cognitive systems, mental representations enable successful problem solving and the achievement of various goals, depending on the current needs of the system.

The broad applicability of the *same* mental representations to achieve *different* goals suggests that the current task itself is not part of their structure. Mental models must be *goal-agnostic* in order to enable the transfer of knowledge about the dynamics of the environment from one task to another.

In order to determine whether these goals have been achieved, the developed system is compared with a baseline approach. Every goal is considered to be achieved if the developed system 'outperforms' this baseline. The performance with regard to the first task is inextricably linked to the purpose of *simulating cognition.* Its success depends therefore strongly on the applied conception of cognition. The applied conception is described in detail. Performance in the second and third task can be objectively determined by metrics from supervised and reinforcement learning.

1.4. General Approach

In conclusion, two equally important measures have to be established to achieve autonomous artificial concept creation. In the following, arguments in favour of these measures are provided because, in the past, they have not been taken into sufficient account.

The first point is that there must be a clear distinction between the descriptive entity and the entity being described. This is motivated by the fact that critique on artificial intelligence from the philosophical side concerns mostly the symbolic procedures that *describe* computational simulations of cognition, not the physically instantiated simulations themselves—which are neither symbolic nor syntactic.

This applies in particular to computational cognitive models and the mental model that they describe as a component of cognitive systems. Criticism of a model does not automatically also apply to what it describes. The fact that an artificial system can be described by a computational cognitive model is *completely independent* from whether the system has mental content as we perceive it in ourselves.

The second point is that computational simulations that implement certain models of cognition are not only *another description,* but exactly *what is being described* by these models. It is readily accepted that the *biological* entities that are described by cognitive models (e.g. other humans) have mental content. There is no valid reason to reject the same in computational entities.

A computer system can mimic cognitive processes. In fact, the subjective properties that these processes exhibit in biological systems, however, are widely rejected in computational simulations. But there cannot be any proof that natural cognitive systems are phenomenologically any different from artificial ones: *subjective experience cannot be observed in either of them.* There is no reason, for example, to exclude artificial cognitive systems from the problem of other minds.[3]

Note that the claim is not that computer simulations have subjective experience. The claim is that the truly autonomous creation of concepts *requires* subjective experience. For a simulation of concept creation, subjective experience *has to be supposed.* Without subjective experience, there can be no content in basic perception and without content in basic perception, there can be no actual mental representation.

Subjective experience cannot be described from third-person-perspective. What can be described, is the emergence of subjective experience from elements that are not yet perceptible to the observed system. The semantic autonomy of the system's mental model can be secured with the ability to autonomously generate its most fundamental perceptions from unconscious sensorimotor activation. The field of embodied cognition offers several explanations along these lines.

Essentially, the simulation of a cognitive system has to be thought of not as *a description,* but as *the instance* of a cognitive system, similar to a robot-assisted assembly line, which simulates human workers to not *describe* the construction of a product, but *to instantiate this exact product* instead.

The product of this process must be conceived of as *indistinguishable* from the product of the real process just like a machine-manufactured car is indistinguishable from a man-made car.

1.5. Proceeding

This work continues as follows. In the following part on the philosophical foundations, the three main problems that the philosophy of mind has with approaches to artificial intelligence are presented.

[3] This problem raises the question of how to justify the assumption that other people have a mind when all that one can perceive of them is their physical appearance.

The relationship between these problems and mental representation is illustrated, and an interpretation of Immanuel Kant is provided that explains these problems.

In the next chapter two interpretations for the symbol grounding problem are presented. The soft interpretation on the one hand, which is mostly applied by the cognitive sciences, and the hard interpretation on the other hand, which is mostly applied by the philosophy of the mind.

Solving the symbol grounding problem requires a hard interpretation. Arguments are provided against the most popular case for the impossibility of subjective experience in computational systems. A hard interpretation of the symbol grounding problem *implies* the problem of autonomous artificial concept creation.

The next part provides a theoretical background on theories of embodied cognition. The conception of cognition presented by them enables to infer conditions for the initial generation and constituents of basic perception.

After that, John Searle's theory on intentionality is detailed as well as its implications for the structure of a mental model that is composed of phenomenological entities. This part is concluded with a semiotic formalisation of the mental representations in such a model.

The next part merges the philosophical with the practical half. It provides a formal definition for mental models. This definition is derived from the previous presentation of intentionality and the selected theories of embodied cognition. It provides the formal basis to determine the procedures behind grounding symbolic mental representations.

The following part presents these procedures and how they apply to basic perception to generate a model for complex and partially observable environments. In the last part, a baseline approach is presented that eventually enables to evaluate the results.

Part I.

Foundations

2. Three Major Problems in Artificial Intelligence

Artificial intelligence and philosophy share a history of mutual misunderstandings and conflicts. A lot of dispute might have been avoided if terms and domains would have been made more clear. Clear not just in the jargon of the arguing party, but clear in words the other side could be expected to understand by using their own well established and successfully applied terminology.

Prime example is the case of Hubert Dreyfus, who very accurately—but no less aggressively—criticised the efforts and optimistic prospects of early artificial intelligence.[1]

The two world views held by Dreyfus and the artificial intelligence pioneers of the seventies seemed fundamentally incompatible. At the epicentre of this conflict is the capacity of cognitive systems to mentally represent parts of external reality in a way like natural cognitive systems do. The possibility of a computational implementation of this capacity flares emotional responses till this day.

This conflict yielded three major problems in artificial intelligence. For some, practical solutions have been found while others remain unsolved or simply lost the attention of research. The three problems are the *frame problem, the problem of vanishing intersections,* and *the symbol grounding problem.*

In this chapter, these problems are presented to provide a historical background for the symbol grounding problem more than an extensive analysis. Throughout the rest of this work, however, these problems are referred to over and over again to show similarities in their symptoms and, more importantly, their common cause.[2] In the following chapters, an argument is established according to which solving the symbol grounding problem implies solving to the other two as well.

[1] The context of events is described in detail by McCorduck (2004).

[2] In *general* terms, their common cause is a failure to communicate implicit conceptions about the processes and structures that are involved in mental representation.

2.1. The Epistemological Frame Problem

Initially, McCarthy and Hayes (1969) presented the frame problem as a problem of first order logic. Later, it obtained a wider epistemological interpretation through Daniel Dennett. He presents the epistemological frame problem as follows.

> When a cognitive creature, an entity with many beliefs about the world, performs an act, the world changes and many of the creature's beliefs must be revised or updated. How? It cannot be that we perceive and notice *all* the changes (for one thing, many of the changes we *know* to occur do not occur in our perceptual fields), and hence it cannot be that we rely entirely on perceptual input to revise our beliefs. So we must have internal ways of updating our beliefs that will fill in the gaps and keep our internal model, the totality of our beliefs, roughly faithful to the world. (Dennett 1981, p. 125)

The question is: how can an agent determine those beliefs about its environment that have to *change* due to its actions? Humans tend to assume that, for any one action, only a limited number of beliefs about the world have to be revised.

If you drop a glass of milk, for example, you probably consider cleaning the floor of the room you are currently in. You do not even have to think about cleaning the living room floor if you dropped the glass in the kitchen.

The assumption that updates concern only *parts* of a belief system is not only justified by observing one's own cognitive processes. There are also practical reasons which indicate that any action can only concern *some* beliefs. In time-critical situations, for example, it is plain impossible to update *every* belief after a particular action.

So how do humans know the limits of the consequences of their actions without even considering outcomes that are particularly improbable? How can a system of beliefs be updated efficiently and effectively?

2.1.1. Approaches

One potential solution suggests that actually *all* beliefs are considered. Current computer hardware is just not fast enough to simulate this in a timely manner. As computer chips become faster, the problem might simply disappear.

Empiric evidence suggests something else. Humans are capable of object and face recognition in a time interval that allows one hundred sequential neural processing steps *at most* (Feldman and Ballard 1982). This has been interpreted as argument for a massively parallel implementation of the respective processes (intitially by Rumelhart 1989; later e.g.

in the 'global workspace theory' of Baars 1993). Considering the amount of beliefs that humans tend to have, however, it appears as if even massively parallel mental processes never update *all* of the system's beliefs.

It seems rather as if only those beliefs receive an update which are somehow *relevant* to the action. If only some beliefs are labelled as 'relevant', the system is relieved of quite some overheads.

John Haugeland called such an approach a 'cheap test strategy'. 'The *cheap test* strategy looks over everything quickly, to tell "at a glance" most of what's irrelevant (hence unchanged).' (Haugeland 1987, p. 83)

Of course, this strategy presupposes a 'prior *categorization* of events and facts, based on which types of events affect which types of facts' (ibid., p. 83). Dropping a glass of milk usually affects the floor and not the ceiling. After dropping a glass of milk, therefore, beliefs about the floor are more relevant candidates for an update than those about the ceiling.

Relevant beliefs depend on context. To determine relevant beliefs *ahead of time,* therefore, implies a fixed bias when facing outcomes that are plain impossible to predict (e.g. arbitrary actions of another cognitive system).

Even more so, to determine which beliefs are relevant, again, the system needs to check *every single belief.* Haugeland attested that 'the system must scan the entire model, relying on some easy sign to rule out most entries without further examination' (ibid., p. 83).

Only after every single belief has been checked, the system can determine which beliefs are relevant *at the moment.* Again, computational systems are either too slow or the number of beliefs about the real world is too large to handle.

Therefore, Haugeland presented an alternative to cheap tests. 'The alternative, *sleeping dog*, strategy is to let everything lie, unless there's some positive reason not to. That is, unless there's some positive indication that a particular fact may be affected, the system will *totally* ignore it, without even performing cheap tests.' (ibid., p. 84)

This description raises the question how to determine the effect of an action in a particular situation. Answering this question requires to determine how situations are recognised by cognitive systems in the first place. This leads straight to the problem of vanishing intersections in the next section.

The frame problem also exposes a more fundamental issue: if the designer of an autonomous system decides *for* the system (e.g. what is supposed to be relevant), then the system can hardly be called 'intelligent'. In general, a designer of artificial intelligence must provide systems with abilities that *enable,* rather than *determine,*

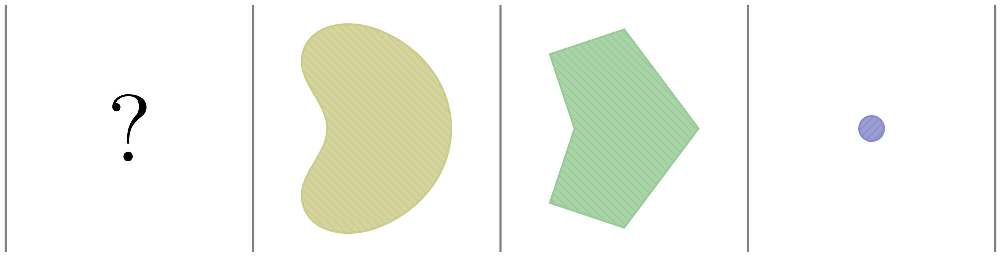


Figure 2.1.: Entities in Harnad's Hybrid Design.

intelligent behaviour.

2.2. The Problem of Vanishing Intersections

One of these abilities is the acquisition of object concepts. Effectively, this is the autonomous generation of internal representations for various parts of the environment.

Implementing concept acquisition in an artificial system requires to determine how objects and situations can be represented, how these representations relate to beliefs, and how this relation is adjusted by the system in case the environment *changes.* In cognitive systems, these representations are *mental.*

Mental representations are internal states that reference external entities. How can an external entity yield an internal state that presents to the system a category (i.e. type) for this particular entity (i.e. token)? How can this system arrive at the content *that there is a chair?* One popular approach in artificial intelligence is *pattern recognition.*

Visual pattern recognition enables, for example, to recognise a picture as the picture of a chair. Despite the fact that general object recognition is among the most common human abilities, it is still hardly matched by artificial means.[3]

Humans recognise objects from projected visual patterns on their retinas. Stevan Harnad proposes that the patterns in these projections are recognised in virtue of *component features* that are present in every projection of the same distal object. He refers to those components as 'invariants'.

A set of invariants can be associated with a particular label. So, *indirectly,* the label now refers to all the individual projections that contain this set of invariants. Thus, labels can provide *names* for the distal objects.

[3] Recently, *deep learning* (i.e. multi-layered artificial neural networks) has made major progresses in terms of pattern recognition. In contrast to human pattern recognisers, however, it still lacks the fundamental cognitive ability to extend structural complexity (i.e. neurons or layers of neurons) when necessary.

The cognitive process of recognising *zebras,* for example, involves 1) a real zebra, 2) the retinal projection of a zebra, 3) invariants in this projection (e.g. stripes and a horse-shape), 4) and the associated label 'zebra'. Figure 2.1 illustrates these entities.

As Harnad points out, however, such structural intersections among all the projections of individual objects have not been found. He calls this phenomenon 'the problem of vanishing intersections'.

'It has been claimed that one cannot find invariant features in the sensory projection because they simply do not exist: the intersection of all the projections of the members of a category such as "horse" is empty' (Harnad 1990, p. 344, footnote 20). According to Harnad, however, 'the reason intersections have not been found is that no one has yet looked for them properly.' (ibid., p. 344, footnote 20)

Unfortunately, even if such intersections exist, they cannot explain how humans recognise vastly different things as members of the same category. Consider a children's depiction of a horse and compare it to the photograph of a horse. Neither colour nor shape need to match to be able to recognise the intended animal.

Different lighting alone an completely change the visual information you receive from members of the same category. Instead of structural intersections, the various different members of a category feature more of a 'family resemblance'.

Ludwig Wittgenstein described his concept of family resemblance with the example of games. "Consider for example the proceedings that we call 'games'. [...] *look and see* whether there is anything common to all.—For if you look at them you will not see something that is common to *all,* but similarities, relationships, and a whole series of them at that." (translated[4], Wittgenstein 2010, § 66)

Undoubtedly, Wittgenstein spoke about what Harnad calls 'the problem of vanishing intersections'. In contrast to Harnad, however, Wittgenstein claims that *there are no intersections.*

2.2.1. Approaches

In light of this critique, two major alternatives to Harnad's invariants can be considered.

The first alternative is presented by Karl MacDorman. He describes that 'the *same* invariant features need not be present for every instance of a category. Any one of a disjunctive set of invariant features, sampled from any number of sensory modalities, can

[4] „Betrachte z.B. einmal die Vorgänge, die wir ‚Spiele‘ nennen. [...] *schau,* ob ihnen allen etwas gemeinsam ist.—Denn wenn du sie anschaust, wirst du zwar nicht etwas sehen, was *allen* gemeinsam wäre, aber du wirst Ähnlichkeiten, Verwandtschaften, sehen, und zwar eine ganze Reihe.“ (Wittgenstein 2010, § 66)

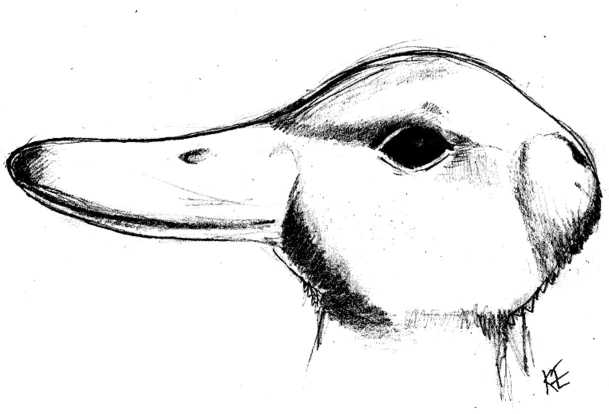


Figure 2.2.: Reversible Duck-rabbit Figure (Honeychurch 2013).

serve to indicate sensorimotor invariance at a more abstract level.' (MacDorman 1997, p. 169)

An invariant according this understanding is therefore not an *intersection* of patterns but their *union.* Not all invariants need to be present in the member of a particular category, it suffices if only some of them are.

Wittgenstein considered this idea as well.

> But if someone wished to say: "There is something common to all these constructions—namely the disjunction of all their common properties"—I should reply: now you are only playing with words. One might as well say: "Something runs through the whole thread—namely the continuous overlapping of those fibres". (translated[5], Wittgenstein 2010, § 67)

MacDorman himself observes that 'in certain cases, the same category will be activated by structurally dissimilar sensory projections and, in other cases, different categories will be activated by structurally similar projections' (MacDorman 1997, p. 170).

This statement is interesting for two reasons. In the first part, he describes not only a problem of vanishing intersections but also a *problem of overwhelming diversity* in each category. In the second part, he even describes *the complete irrelevance* of structural properties for some categories.

[5] „Wenn aber Einer sagen wollte: ‚Also ist allen diesen Gebilden etwas gemeinsam,—nämlich die Disjunktion aller dieser Gemeinsamkeiten'—so würde ich antworten: hier spielst du nur mit einem Wort. Ebenso könnte man sagen: es läuft ein Etwas durch den ganzen Faden,—nämlich das lückenlose Übergreifen dieser Fasern." (Wittgenstein 2010, § 67)

There are cases, where even *the exact same pattern* is recognised in two different ways. Neither conjunctive nor disjunctive invariants can account for this. Consider, for example, the reversible duck-rabbit in figure 2.2. The image can be recognised as a duck *and* as a rabbit. The underlying pattern, however, is always *the same.*

The second alternative to Harnad's approach explains such phenomena with *context-dependency.* Dreyfus describes this context-dependency in more detail.

> A phenomenological description of our experience of being-in-a-situation suggests that we are always already in a context or situation which we carry over from the immediate past and update in terms of events that in the light of this past situation are seen to be significant. (Dreyfus 1992, p. 288)

Following from this point, it can be speculated that categories do not appear to be determined by the features of their members *because objects are not recognised only in virtue of their structure.* MacDorman arrives at the same conclusion. 'Taken by themselves feature detectors are insufficient to ground the vast number of symbols required to represent all the different kinds of potentially recognizable things.' (MacDorman 1997, p. 171)

The performance of human pattern recognition might be explained with structural invariants *in a particular context.* In pitch-black darkness, the visual appearance of a horse cannot play any part in recognising it. In this context, you do not even *try* to recognise a horse by its visual appearance.

Unfortunately, Dreyfus points towards computational trouble with context-dependency: 'if each context can only be recognized in terms of features selected as relevant and interpreted in terms of a broader context, the AI worker is faced with a regress of contexts.' (Dreyfus 1992, p. 289)

A consequential implementation of this principle, he argues, yields problems. 'This need for prior organization reappears in AI as the need for a hierarchy of contexts in which a higher or broader context is used to determine the relevance and significance of elements in a narrower or lower context.' (ibid., p. 288)

The question is, whether it is possible to implement such a hierarchy contexts in a computer system.

2.3. The Symbol Grounding Problem

According to Harnad, computational systems that perceive their environment in virtue of symbolic representations are confronted with an impossible task. "Suppose you had to learn Chinese as a *first* language and the only source of information you had was a Chinese/Chinese dictionary" (Harnad 1990, pp. 339–340).

To understand Chinese in virtue of the shapes of Chinese signs alone appears very unlikely already. To learn it without any prior knowledge about language *in general,* however, is hopeless at best. (Harnad 1990)

Reason for this hopelessness is the character of symbolic content. According to Harnad, the content of symbols is not intrinsic to the symbol like it is with pictures or iconic illustrations. Instead it must be *provided* by an interpreter (e.g. someone who understands the entries in the dictionary).

A computer programme might be able to reproduce the entries in a Chinese/Chinese dictionary. Without their interpretations, however, it can only receive and produce *the shapes* of these symbols, not their content.[6]

> The symbols and the symbol manipulation, being all based on shape rather than meaning, are systematically *interpretable* as having meaning [...] But the interpretation will not be *intrinsic* to the symbol system itself: it will be parasitic on the fact that the symbols have meaning for *us,* in exactly the same way that the meanings of the symbols in a book are not intrinsic, but derive from the meanings in our heads. (ibid., pp. 338–339)

Without content, the system cannot have an actual belief *about,* or an actual representation *of,* anything. The content any symbolic representation appears to have is in fact content only *in the mind of its interpreter,* not in the observed structure itself. The representations in an observed system, therefore, are actually only representations *to the observer* but not to their system.

The frame problem and the problem of vanishing intersections both ask for appropriate internal representations as well. They do this by pointing out two different inadequacies in current conceptions of mental representation.

The frame problem emphasises the overheads that representations require. The problem of vanishing intersections emphasises the inability to imitate the various types of relation that real mental representations can be in with their external referents.

Accordingly, Harnad argues that both are only sub-problems to the symbol grounding problem (Harnad 1990, p. 339, footnote 6; Harnad 1990, p. 344, footnote 20).

[6] Unless expressed otherwise, mental content and meaning are considered as synonyms. Consider this in contrast to the two classical conceptions of meaning according to Gottlob Frege. *Fregean sense* accords roughly to what is here called ‘mental content’ and *Fregean reference* to what is here called the ‘referent’ of a mental representation.

2.3.1. No Content in Symbol Systems

According to Harnad, “the symbol grounding problem is referred to as the problem of intrinsic meaning (or ‘intentionality’)” (Harnad 1990, p. 338). He gives credit to Searle for the original formulation of this problem.

Harnad also refers to Searle’s Chinese room argument as another instance of the symbol grounding problem. The Chinese room argument goes as follows.

Searle differentiates two ambitions in artificial intelligence. On the one hand, there is *weak artificial intelligence* as the idea that artificial systems which exhibit intelligent behaviour *simulate* a mind. On the other hand, there is *strong artificial intelligence* as the idea that systems that show intelligent behaviour actually *have* a mind. The Chinese room argument is directed against the latter.

The line of thought behind strong artificial intelligence is that, without any empirical access to the supposed mind of another system, we cannot but regress to observable behaviour: if a system behaves intelligently, then it must be intelligent. To deny such a system a mind would be arbitrary and hardly justifiable.

Searle rejects this line of reasoning. He argues that we can very well go ‘below’ mere observation by imagining the *experiences* we had, if we were to be a crucial component of the system in question. From the perspective of the system, we can abandon behaviourist criteria in favour of examining the system’s mental content *first hand.*

In a *symbol system,* so Searle, no content were to be found in virtue of such an examination. Symbol systems are purely *formal:* They manipulate only the shapes of symbols. They have access neither to what they contain (i.e. their content) nor to what they refer to (i.e. their referents). As the parts of a symbol system, therefore, we would be *isolated* from the content of the symbols that it manipulates.

Therefore, symbol systems can exhibit the most intelligent behaviour and, still, it would be invalid to call them ‘intelligent’. If the components of such a system are isolated from content, then so is the system as a whole. A system that is isolated from content cannot be intelligent.

2.3.2. The Chinese Room Argument

Searle illustrates this argumentation with a thought experiment. He describes himself sitting in a closed room. He assumes to have a rule set at hand which allows him to react to unknown Chinese signs that he receives from outside. The reaction dictated by this rule set is again a Chinese sign that can be understood by Chinese speakers as an appropriate reply to the Chinese statement he received before.

In this setting, Searle exchanges signs without understanding any of them. If the observer was to start a conversation with him, Searle would be able to reply *as if* he understood Chinese. Although conversing fluently and elaborately in Chinese, however, the content of the signs would not be accessible to him.

According to Searle, therefore, it is not valid to concede a mind to a system solely on the basis of observed behaviour. Instead, intelligent behaviour is neither necessary, nor is it sufficient, for a mind. (Searle 1980b)

In some respect, the Chinese room argument is a parody on the *Turing test.* It copies the premises of Alan Turing's *imitation game*, where a human has a conversation with an artificial system via chat. If the system is able to convincingly imitate a human in written conversation, according to Turing, we have no reason to reject its ability to think. (Turing 1950)

By entering the system in question, Searle obtains an 'inside' view on all the processes and components at play—and content is not among his observations.

He reinforces this intuition by making *himself* part of the system which shows that not even he—a cognitive system principally capable of understanding—would have access to content in such a setting.

Within the room, Searle acts as an indicator for content. He assumes that content in the indicator person is necessary for content in the whole system. Hence, if *he* would not have access to content, then the whole system would not as well.

2.3.3. Harnad's Approach

Harnad tries to find a way around the semantic limitations of symbol systems. He accepts the premises as they are presented by Searle but also assumes that the content of symbolic representations can be grasped by *connectionist networks.*

Systems that implement this method are supposed to feature 'non-symbolic' or 'subsymbolic' components besides the symbolic representations considered by Searle. Harnad calls them 'hybrid', accordingly.

The central procedures in such a system operate on the entities from figure 2.1 on page 15: 1) distal objects, 2) iconic projections, 3) invariant features, 4) and category names. He describes these procedures in a hybrid system as follows.

> Icons, paired with feedback indicating their names, could be processed by a connectionist network that learns to identify icons correctly from the sample of confusable alternatives [. . .], thereby reducing the icons to the *invariant* (confusion-resolving) features of the category to which they are assigned. In

> effect, the “connection” between the names and the objects that give rise to their sensory projections and their icons would be provided by connectionist networks. (Harnad 1990, p. 344)

Three of these four entities present something else. Names present invariants, invariants present icons, and icons present external objects. This transitive presentation makes invariants and names *re-presentations.* Invariants represent objects and names represent icons.

Icons mediate the relation between invariants and objects just like invariants mediate the relation between names and icons. In symbolic representations, *the content* of the representation provides this mediation between symbolic shape and external referent. As a consequence, invariants could be said to be iconic representations that *contain* images and names could be said to be symbolic representations that *contain* invariants.

In Harnad’s model, therefore, category names are the shapes, invariants the content, and icons the referents of mental representations. The connectionist networks identify invariants and, therefore, they generate content that is independent from external interpretation.

In his analogy with the Chinese/Chinese dictionary, Harnad effectively extends the Chinese signs in the dictionary with icons that illustrate their meaning. By extending the vocabulary of a system with iconic representations, he implies in analogy that an *illustrated* Chinese/Chinese dictionary would suffice to learn Chinese as a first language.

Previous ambitions to discover structural invariants within these icons yielded the problem of vanishing intersections. Harnad assumes, however, that supervised artificial neural networks can find invariants that humans were unable to identify so far.

Harnad aims to solve the symbol grounding problem by designing a system that solves the problem of vanishing intersections. Invariants are identified in patterns ‘to which these names are assigned’ by one or many supervisors.

The need for supervision suggests that Harnad’s approach is motivated *linguistically.* During language acquisition, there is social supervision according to linguistic conventions. During the acquisition of elementary concepts, however, there is no such feedback.

Without the support of *already existing* linguistic content, therefore, there cannot be supervision and Harnad’s hybrid system cannot learn anything.

2.3.4. Searle’s Objection

By presenting the Chinese room argument as another case of the symbol grounding problem, Harnad implies that the absence of content in the Chinese room has the same reasons as the inability to learn Chinese from a Chinese/Chinese dictionary.

He specifically refers to Searle's ideas on intentionality to justify this claim. In fact, however, Searle does not share Harnad's understanding of symbol grounding. (Harnad 1989; Searle 1993; Harnad 2001)

Already in his original formulation, Searle explicitly considers connectionist approaches to the problem. "Suppose we design a program that [...] simulates the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories in Chinese and gives answers to them." (Searle 1980b, p. 420)

Such a brain simulator is clearly a connectionist network like Harnad proposes it.[7] Searle rejects the connectionist approach by modifying the original Chinese room argument.

> To see this, imagine that [...] we have the man operate an elaborate set of water pipes with valves connecting them. When the man receives the Chinese symbols, he looks up in the program, written in English, which valves he has to turn on and off. Each water connection corresponds to a synapse in the Chinese brain, and the whole system is rigged up so that after doing all the right firings, that is after turning on all the right faucets, the Chinese answers pop out at the output end of the series of pipes.
>
> Now where is the understanding in this system? (ibid., p. 421)

Searle identifies the same problem as in the original thought experiment.

> The problem with the brain simulator is that it is simulating the wrong things about the brain. As long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states. (ibid., p. 421)

Connectionist networks follow a symbolic syntax—just like any algorithm. They are *formal* and susceptible to Searle's critique as well.

The symbol grounding problem according to Searle is not that symbolic representations have not been implemented *properly* but that symbols must be *interpreted* by the system that is supposed to access their content.

[7] Notice that this is not supposed to suggest that brains are *mere* connectionist networks as they are used in machine learning but that each brain features *at least* those functional properties that are also part of connectionist networks.

2.4. Kant on the Three Problems

"Thoughts without content are void; intuitions without conceptions, blind" (translated[8], CPR B 76). Kant's famous quote provides another perspective on the three problems.

In *Critique of Pure Reason*, he presented his ideas on the interplay between intuition and thought: the abstract content of thought is grounded in basic perception but external events can *only* be perceived within a conceptual frame that is prior to experience itself.

In the 18th century, two philosophical camps argued about whether immediate perception determines, or *is* determined, by abstract reasoning. Kant's ambition was to consolidate these empiricist and rationalist conceptions of cognition. His ideas can help to solve the computational problem of artificial concept creation.

On the one hand, *thoughts without content are void.* Without a functional equivalent to structural content, computational systems can only emulate thought. Such thought, however, is *devoid of intrinsic meaning.*

On the other hand, *intuitions without conceptions are blind.* Without a functional equivalent to conceptual context, computational systems can only emulate basic perception. Such perception, however, is *blind for the current situation.*

To see why this is relevant to the three problems, consider the following theses.

1. The symbol grounding problem results from failure to consider structures 'within' mental representations. These structures provide mental representations with *autonomous content.*

2. The frame problem results from failure to consider the structures that mental representation are contained in. These structures provide a context that describes the *current situation.*

3. The problem of vanishing intersections results from failure to use the current situational context to help select a representation with appropriate content.

A solution to the symbol grounding problem requires content for the most basic perception. A solution to the frame problem requires context for the most abstract concept. A solution to the problem of vanishing intersections, eventually, requires to show how content and context interact during mental representation.

For an example on this interpretation, imagine a system that detects red objects in its environment. To identify an object as red, the shape 'red' must be associated with the appropriate electromagnetic spectrum and the category determined by this spectrum

[8] „Gedanken ohne Inhalt sind leer, Anschauungen ohne Begriffe sind blind.“ (KRV B 76)

must apply to the current sensor activation. The spectrum provides *content* for the shape 'red'.

As soon as the lighting conditions change from white to green, however, the system is literally *blind* for what it is supposed to detect. Without considering the current context of light, the system cannot but tacitly concede static conditions at all times. To tackle dynamically changing conditions, each content must relate to a situational frame.

Read like this, Kant's quote can be taken as an emphasis on the equal importance of content and context in mental representation. Most approaches to the symbol grounding problem, however, are concerned only with content.

3. Two Conceptions of Mental Representation

Although mental representation is central to the symbol grounding problem, the underlying conception that is applied is often left implicit. This is problematic because the cognitive sciences and the philosophy of mind each feature incompatible conceptions of mental representation. The following example illustrates such an incompatibility.

The shape 'Elvis is dead' can be interpreted as English. The observer of a system that contains this shape can infer from it the content *that Elvis is dead.* According the *objective* conception of mental representation, at this point, it is also valid for the observer to infer that *the system* knows about the death of Elvis Presley as well.

According to the *subjective* conception, the content *that Elvis is dead* is only effective in the mind of the observer (e.g. the system's programmer). The system itself has no access to this content. It does not understand a single English word and might associate 'Elvis is dead' with no content at all—or with any content unknown to the observer.

> If the *programmer* chooses to interpret the machine inscription "Robin Roberts won 28" as a statement about Robin Roberts [...], that's all well and good, but it's no business of the machine's. The machine has no access to that interpretation, and its computations are in no way affected by it. (Fodor 1980, p. 233)

In this quote, Jerry Fodor argues that artificial systems do not have access to the content that their designer obtains from interpreting their internal shapes.

However, the reverse is true as well: the designer is just as isolated from potential content that the system might associate with its internal shapes as the system is isolated from its designer's content.

According to the second conception, therefore, it cannot be inferred that there is no content *just because none has been observed.* The Chinese room argument, however, tacitly concedes that the mental content of another system (i.e. the room) *could* be observed.

This assumption is not only incompatible with how Searle usually conceives of mental representation (for details, see chapter 12), it also violates *methodological solipsism* (see section 3.4).

However, before arguments are put forward to support these two assertions, the origins of the mutually exclusive conceptions of objective and subjective mental representation are elaborated below.

3.1. Objective and Subjective Representation

The two conceptions have been informally introduced by Uriah Kriegel as *objective* and *subjective* representation.

> The distinction between objective and subjective representation is closely related to other, more familiar ones. Distinctions between personal and sub-personal representations, narrow and wide representations, phenomenal and psychological representations, may all turn out to be co-extensive with the subjective/objective distinction. (Kriegel 2013, p. 163)

The content of objective representations in another system is intrinsic to the system's *observer*. According to this conception, the shape 'Elvis is dead' means to the system just what it means to its observer. This conception requires the content of the representation to be *public* (i.e. shareable and communicable among a community).

Objective representation is common to linguistics and widespread in cognitive sciences—for example in the work of Harnad and Steels. It implements a propositional conception of *semantic content* as it is the case, for example, in *the description* of a cognitive system (i.e. a cognitive model).

The content of subjective representation in another system, in contrast, is intrinsic *to the system itself.* According to this conception, the observed system's interpretation of 'Elvis is dead' can yield vastly different content than an observer's interpretation does. This conception requires the content of the representation to be *private* (i.e. inexpressible and accessible only to first-person-perspective).

Subjective representation is common in the philosophy of mind—for example in the work of Searle and Tim Crane. It implements a conception of *phenomenological content* as it is the case in *the subjective experience* of a real cognitive system.

In the following, first, *representationalism* is described as background for the subjective conception. Afterwards, *intentional psychology* is presented as background for the objective conception.

3.1.1. Representationalism

Representationalism is the idea that the interaction between cognitive systems and external reality is *mediated.* This implies that the objects of basic perception cannot be considered to be *objectively real* in a literal sense of the word but rather *mental representations* of reality.

The world of objects as it is perceived is actually just the phenomenal appearance of one's own mental model. This model consists of mental representations that present *inexistent* objects to the mind which could be radically different from the real things that caused them in the first place. These representations can also appear when their external referents are absent, for example, during dreams or hallucinations, but also while imagining or planning (for details, see chapter 10).

These representations are only perceived by oneself. Mental representations in *other* systems can only be assumed through *the analogical inference of other minds:* Other systems act and look similar to myself, therefore, their perception of self might also similar to mine (see, among others, Hyslop and Jackson 1972; Searle 1980b, under 'The Other Mind's Reply'; Melnyk 1994; Hyslop 2013, pp. 29–70).

Mental representations re-present inconceivable external reality by presenting it to their system in the form of objects. Therefore, the content of subjective mental representations is also referred to as 'intentional object' (e.g. Crane 2001). The *locus classicus* on the cognitive nature of objects is Kant.

> Up to now it has been assumed that all our cognition must conform to the objects; but all attempts to find out something about them *a priori* through concepts that would extend our cognition have, on this pre-supposition, come to nothing. Hence let us once try whether we do not get further with the problem of metaphysics by assuming *that the objects must conform to our cognition,* which would agree better with the requested possibility of an *a priori* cognition of them, which is to establish something about objects before they are given to us. (translated[1], CPR B XVI, second emphasis added)

To describe intentional content as *the concept* of an object enables an intuitive understanding. More appropriate is, however, to conceive of it as *the object itself* and of the

[1] „Bisher nahm man an, alle unsere Erkenntnis müsse sich nach den Gegenständen richten; aber alle Versuche über sie a priori etwas durch Begriffe auszumachen, wodurch unsere Erkenntnis erweitert würde, gingen unter dieser Voraussetzung zu nichte. Man versuche es daher einmal, ob wir nicht in den Aufgaben der Metaphysik damit besser fortkommen, daß wir annehmen, die Gegenstände müssen sich nach unserem Erkenntnis richten, welches so schon besser mit der verlangten Möglichkeit einer Erkenntnis derselben a priori zusammenstimmt, die über Gegenstände, ehe sie uns gegeben werden, etwas festsetzen soll.“ (KRV B XVI)

shape as which this content appears as *the object's phenomenal appearance.* (McGinn 2004; Strasser 2011)

Consider a plane trip. When you plan the journey, you do not imagine your luggage in all of its detail. Instead, you process a simplified phenomenal shape as replacement for the object's structural complexity.

This replacement is the phenomenal shape of a subjective mental representation. This representation only presents its structural content if necessary—for example as soon as you need to claim your bag at the destination airport. The real referent of this representation in external reality can be perceived *only* in virtue of this representation.

3.1.2. Intentional Psychology

A comprehensive model of cognition describes other systems with mental states just like those we experience in ourselves. *Intentional psychology* assumes mental states in other systems to enable the explanation and prediction of their behaviour.

According to intentional psychology, the mental state of a cognitive system consists of *content* and an *attitude* towards this content. Carl's knowledge that a bottle of beer is in the fridge, for example, consists of the content *that a bottle of beer is in the fridge* and his attitude of *knowing* that this is the case—in contrast, for example, to *fearing* or *hoping* that a bottle of beer is in the fridge.

Carl's going to the fridge can be explained by two mental states. For one, he *desires,* and, for another, he *believes, that a bottle of beer is in the fridge.* According to a mental-state-model of Carl, these two mental states are sufficient for him to go to the fridge: they can *explain* his behaviour.

Describing the content of the mental state of another system is intricately problematic. How do you describe something that is only accessible to the described system?

Even if this content could be observed, how could it be differentiated from the content that the observation has to the observer? After all, the issue is to observe the content that *another* system has and not to obtain the content that this observation might have to *oneself.*[2]

One way to deal with this problem is to simply replace the unobservable elements with ones that *can* be observed. Intentional psychology replaces the content of mental states with *representations for semantic content.*

Following this strategy, mental states can be easily formalised as an ordered triple. The triple (Carl, believes, 'Elvis is dead') formally describes Carl's belief that Elvis is dead.

[2] This problem also concerns the semantic capacities of neural structures: *observing* neural correlates for content is essentially different from *having* this content yourself.

The linguistic shape ‘Elvis is dead’ is a proposition that stands for the content *that Elvis is dead.* (Pitt 2013)

An objective representation presents semantic content *to the observer* of the described system. To systems that are described as containing this shape, however, ‘Elvis is dead’ might mean nothing at all.

Intentional psychology is a family of theories that all deal with the unobservability of mental content in the same way: by replacing it with an objective representation. This effectively provides a propositional ground for models of cognition in the mental content of the observer.

Intentional psychology is a way *to describe* mental representation and, therefore, it applies to cognitive models. Representationalism, in contrast, is a hypothesis on *how cognitive systems conceive of reality* and, therefore, it applies to real cognitive systems.

This is the difference between mental representation in intentional psychology and mental representation in representationalism. This difference can be emphasised by the question *who* some structure is a mental representation *for,* in contrast to just asking *what* it is a representation *of.*

3.2. Representation in the Cognitive Sciences

Mental representation in cognitive sciences mostly follows the following description. “To perceive a strawberry is, on the representational view, to have a sensory experience of some kind which is appropriately related to (e.g., caused by) the strawberry” (ibid.). Definitions of mental representation are formulated accordingly.

> [To account for the properties of human cognitive capacities] we must posit mental representations that can represent specific objects; that can represent many different kinds of objects—concrete objects, sets, properties, events, and states of affairs in this world, in possible worlds, and in fictional worlds as well as abstract objects such as universals and numbers; that can represent both an object (in and of itself) and an aspect of that object (or both extension and intension); and that can represent both correctly and incorrectly. (von Eckardt 1999, p. 527)

Example and definition both share the same premise: mental representations reference *objects.* If objects require the subjective mental representation of imperceivable external reality, then *the representation of an object* can only be secondary to a prior mental representation in the mind of an observer so they can conceive of this object in the first place.

The representation of an object depends on a prior representation because objects can only exist due to the mental representation of inconceivable reality in the mind of an observer.

Take the example of the strawberry. To explain a system's perception of strawberries with an internal shape that is supposed to represent strawberries tacitly introduces the the observer's *own* concept of a strawberry.

Without a prior reason to associate the internal shape with the concept of a strawberry, there is in fact no ground for the assumption that the shape in the observed system ought to represent strawberries *at all.*

Genuinely *mental* representation is the representation of otherwise inconceivable real things. The representation of *an object,* on the other hand, can only be a secondary description of this thing *after* it has already been conceived of it as an object.

Ambitions to solve the symbol grounding problem are held back by a confusion of these conceptions. Ignorance for the difference between objective and subjective mental representation lends itself to a conflation of representations in *the description* of a system and representations in the *actual* system.

The Chinese room argument is an illustrative example for this, because the reader is asked to locate the content of real mental representations in the mere *description* of a real cognitive system.

3.3. The Physical Symbol Systems Hypothesis as a Statement about Descriptions of Cognition

The physical symbol system hypothesis is another example. It states the following: "[a] physical symbol system has the necessary and sufficient means for general intelligent action" (Newell and Simon 1976, p. 116). This expresses a biconditional: *if, and only if,* some thing behaves intelligently *then* it is a physical symbol system.

First of all, intelligent *action* is a behavioural criterion, not a cognitive one. A physical symbol system might very well *exhibit* superhuman intelligence and, still, not provide *any* insight into the processes of real cognitive systems.

However, 'action' can also be interpreted *cognitively* as the system's manipulation of its internal state. According to this wider interpretation, the physical symbol system hypothesis states that symbol systems are necessary and sufficient for the intelligent manipulation of mental representations. This interpretation is in fact suggested by Newell.

> The symbols that float everywhere through the computational innards of this system refer to the road, grass and trees in an epistemologically adequate, though sometimes empirically inadequate, fashion. These symbols are the symbols of the physical symbol system hypothesis (Newell & Simon, 1976), pure and simple. (Newell 1988, p. 421, citation in source)

Without *knowing* that the shapes that a symbol system manipulates refer to 'the road', 'grass', and 'trees', however, it also cannot be determined whether their manipulation is indeed intelligent or not. Without being able to read Chinese, it is impossible to recognise whether Chinese symbols are composed intelligently or not.

To determine the intelligent manipulation of another system's internal representations, therefore, requires content in the mind of an observer—*external* content about shapes in the observed system. This external content has no influence on the observed system and is independent from the content that these representations might *actually* have to the system.

It follows that the physical symbol system hypothesis cannot be a statement about cognitive systems but only about *models* of these systems. It merely claims that the means for general intelligent action can always be *understood* symbolically, not that they *are* essentially symbolic.

This is most striking when Newell describes representation as a *fundamental necessity* for symbol systems.

> The most fundamental concept for a symbol system is that which gives symbols their symbolic character, ie, which lets them stand for some entity. We call this concept *designation,* though we might have used any of several other terms, eg, *reference, denotation, naming, standing for, aboutness,* or even *symbolization* or *meaning.* The variations in these terms, in either their common or philosophic usage, is not critical for us. (Newell 1980, p. 161)

According to Newell, representations are a fundamental *requirement* for the physical symbol system hypothesis. As a consequence, content that *enables* representation in the first place must be out of its scope.

Inconsistencies appear only under the assumption that symbol systems are what they are merely supposed to describe. There is an important difference between the statement that intelligent systems *are* symbol systems and the statement that intelligent systems *can be described* by symbol systems.

Even if cognitive systems use symbolic means for intelligent action, there still has to be a part in such systems from which these symbols emerge. Without this ground to symbolic representations, no system is capable to truly autonomous concept creation.

That intelligent action can be described by symbol systems, on the other hand, appears to be a much more reasonable claim. A mere description of intelligent systems does not need to concern itself with the origin of the manipulated symbols—as Newell emphasises in the last quote.

3.4. Methodological Solipsism

The confusion of objective representations and subjective representations is only one part of the problem. Another is that it is fundamentally impossible to describe *anything* independent from prior content. Objective representation *always* depends on subjective representation. In other words: describing mental representation violates *methodological solipsism.*

Jerry Fodor identifies methodological solipsism as the general research strategy of the cognitive sciences (Fodor 1980). Hilary Putnam explains methodological solipsism as "the assumption that no psychological state, properly so called, presupposes the existence of any individual other than the subject to whom that state is ascribed" (Putnam 1975, p. 136).

According to a consequential interpretation of methodological solipsism, *actual* mental representations *must* appear to an observer as though they were exclusively formal and without content like the Chinese room and physical symbol systems in general. Fodor explains why.

> [T]he formality condition [...] is tantamount to a sort of methodological solipsism. If mental processes are formal, then they have access only to the formal properties of such representations of the environment as the senses provide. Hence, they have no access to the *semantic* properties of such representations, including the property of being true, of having referents, or, indeed, the property of being representations *of the environment.* (Fodor 1980, p. 231)

The content in real mental representations can be independent from other content. Content that is truly independent from other content, however, also has to be independent from the mind of the observer. To the observer, any system that operates according to truly independent content *must* appear exclusively formal as its content cannot be observed from outside its system.

Searle uses the Chinese room to describe a cognitive system. But no description *is* what it describes—it would not be a description after all but the thing itself instead. You would not expect a description of Paris to yield the actual Eiffel tower. A description of Paris also does not yield the Eiffel tower. It is hard to imagine anything that is yielded by its description. The same is the case for 'understanding' in a description of cognition.

The absence of a correlate for understanding in the Chinese room in fact only strengthens the analogy between the room and a real cognitive system. *Just like in real cognitive systems,* there is no obvious equivalent for subjective mental content. This is not even surprising but already implied by any approach that respects methodological solipsism.

3.5. The Primacy of Subjective Representation

The representations in a cognitive model are always and necessarily objective: they represent conventional and communicable *concepts* to an observer. The representations in a real cognitive system are subjective: they represent immediately inconceivable *reality* to their system.

The similarity between objective and subjective representation is that both relate shapes with referents. But only subjective representations reference immediate reality. Objective representations reference subjective representations in *another* system instead.

To determine the difference becomes exceedingly intricate if mental representation is considered necessary for object concepts in general. Concepts are necessary to describe anything. An objective representation, therefore, must always be accompanied by a subjective representation. Their content, however, is not the same.

Crane provides a criterion to distinguish both in virtue of their content. Subjective representations carry *phenomenological content* and objective representations carry *semantic content.*

> There is [...] the phenomenological conception of content ('what is conveyed to the subject'). The propositional content of a perceptual experience is also something that deserves the name of 'content'. But it must be distinguished from content in the phenomenological sense. The content in the phenomenological sense is something spatiotemporal, concrete, particular and specific to its subject. The content in the propositional sense is not. There are, therefore, two conceptions of the content of experience, the semantic and the phenomenological. I think that the phenomenological conception has a certain

> priority, *since it is part of what is being modelled.* Semantic contents can only be 'descriptions' of this content. (Crane 2013, p. 245, emphasis added)

Crane grounds semantic content in phenomenological content. As a consequence, objective representation must be secondary because, eventually, it can only describe some other system's subjective representation. A methodologically solipsist approach to modelling cognition, therefore, must adopt subjective mental representation.

A comprehensive, methodologically solipsist model of cognition also can only be achieved indirectly: not by describing the content in mental representations but rather by describing *the processes that generate it.* That the generated structures might not *appear* to an observer as though they had any content is not only a side-effect but fundamentally necessary for a veridical simulation of cognition.

4. One Hard Symbol Grounding Problem

The subjective and objective conception of mental representation yield the two incompatible ideas of Harnad and Searle about what the symbol grounding problem actually consists in. Lawrence Shapiro grasps the general ambiguity of the symbol grounding problem nicely.

> The ambiguity is in the notion of meaning. [. . .] [T]hose who work on the symbol grounding problem present it as a problem about how symbols come to mean, or be about, or represent features of the world. However, conflated with this problem appears to be another, concerning how people come to *understand* the meanings of symbols. Although this distinction between how a symbol becomes meaningful and how meanings come to be understood is seldom noted, further reflection on the Chinese Room shows both questions at play. (Shapiro 2011, p. 96)

According to Shapiro, one interpretation of the symbol grounding problem concerns *language* while another concerns *the mind.* Unfortunately, there is no immediate relation between linguistic and mental representation. Shapiro explains this by arguing that

> one can imagine that philosophers have discovered the true theory of meaning while psychologists remain at a loss to explain how people come to understand the meanings of the expressions they use, or one can imagine that psychologists have discovered how people come to understand language, despite philosophers' failure to explain how symbols, including linguistic ones, acquire their meaning in the first place. (ibid., p. 97)

He describes solutions to two different problems. One problem is particularly linguistic while the other concerns meaning in general. Harnad deals with the former while Searle deals with the latter.

Admittedly, Searle invites a linguistic interpretation of the Chinese room argument by presenting it with linguistic (i.e. Chinese) signs. But his intention is to transfer this example to *any* type of interaction between system and environment (i.e. perception and action). He expresses this as follows.

> Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating drinking—anything you like. (Searle 1980b, p. 420)

Searle determines that "the addition of such 'perceptual' and 'motor' capacities adds nothing by way of understanding, in particular, or intentionality, in general" (ibid., p. 420). Inside the robot, Searle is still "receiving 'information' from the robot's 'perceptual' apparatus, and I am giving out 'instructions' to its motor apparatus without knowing either of these facts." (ibid., p. 420)

With the Chinese room, he uses *the example* of linguistic symbols to present a comprehensible implementation of the processes that are generally assumed to underlie cognition. His explicit intention, however, is to *transfer* linguistic insights (e.g. from his speech-act theory) onto mental content and representations. (Searle 1983)

Therefore, the Chinese room argument concerns not only how linguistic shapes obtain content, but rather how content *in general* develops from something that does not have content already.

Conventionalised linguistic content is necessarily secondary to the content in the minds of individual cognitive systems. Unless explicitly transferred like Searle does it, therefore, only the latter can contribute to understanding the former, not the other way round.

4.1. Two Symbol Grounding Problems

In the following, the linguistic understanding is presented as a 'soft' interpretation, and the cognitive understanding as a 'hard' interpretation, of the symbol grounding problem.

The soft problem is the question how cognitive systems come to *agree* upon the content of symbols. Harnad's hybrid design and similar approaches (e.g. Steels 1999; Steels 2008) propose solutions for this.

In Harnad's design, agreement among cognitive systems emerges due to external feedback that enables to identify invariant features. A similar process is at play in Steels' 'Talking Heads' experiment (Steels 1999).

All solutions to the soft problem follow the same general strategy: use the content *one* type of representation has to the designer (e.g. invariant features) to ground the content *another* type of representation ought to have to the system (e.g. category names).

In analogy to Harnad's Chinese/Chinese dictionary, such a strategy can be illustrated as a dictionary where, instead of definitions, there are iconic illustrations of what each

Chinese symbol means. From these pictures, the system might be able to infer, for example, structural invariants.

To Searle, such an approach is merely another Chinese room that simulates the operations of an intelligent system formally. His critique applies to *all* computational simulations of intelligence, independent from their particular implementation.

The hard problem is that no system can generate truly autonomous content if the content of its most fundamental representations is already predetermined by design.

A solution to the hard problem needs to provide a fundamental explanation for content that cannot simply regress to the content in *another* representation *because that would beg the question.* This applies not only to symbolic representations but also to icons because they must be interpreted as well.

The interpreter of an icon needs to know what properties of the referent are considered relevant and which are not. Tim van Gelder calls this the ‘privileged means of processing’. Without prior content in form of these means, the interpreter of an icon can have no clue *in what respect* the structure of the icon and the structure of the referent are supposed to be similar. (van Gelder 1999, pp. 133–137)

To return to the dictionary: you cannot provide images for the words in a dictionary without presuppositions on the reader’s optical information processing. By presupposing a particular kind of information processing, however, you tacitly exclude all other forms.

In fact, the hard problem is much similar to Harnad’s Chinese/Chinese dictionary with the additional impediment that you cannot regress to anything *but* the shapes of Chinese symbols. Imagine yourself principally unable to experience anything but the shapes of Chinese symbols. How can content emerge in such a system?

4.2. In Favour of the Hard Problem

Despite Searle’s various clarifications on the problem that the Chinese room argument is supposed to present (again in Searle 1993), Harnad’s interpretation as the problem of understanding linguistic representations gained traction among various researchers.

Some even see the hard symbol grounding problem as a misinterpretation. It is argued that

> contrary to what several authors have misconstrued throughout the last three decades, the [symbol grounding problem], which is clearly concerned with providing internally manipulated symbols with concepts, does not refer to intrinsically meaningful internal states of mind (Rodriguez et al. 2012, p. 34).

Instead, so the argument, “grounded means the ability to pick out referents for manipulated symbols and not the ability to make sense of symbols.” (Rodriguez et al. 2012, pp. 31–32)

Some even perceive Searle’s hard interpretation as unfair and conclude that he does not actually understand the problem at hand.

> Was Searle’s paper (and subsequent philosophical discussion) based on ignorance or on a lack of understanding of what was going on in these experiments? Probably partly. It has always been popular to bash AI because that puts one in the glorious position of defending humanity. (Steels 2008, p. 6)

The soft interpretation of the symbol grounding problem is much more accessible and therefore much more dominant than the hard interpretation. So dominant, in fact, that reviews omit their essentially distinct ambitions. This contributes further to the conflation mentioned by Shapiro.

In Taddeo and Floridi (2005), for example, Harnad’s hybrid approach to the soft problem is presented next to Ron Sun’s approach to the hard problem. However, Sun explicitly addresses the difference between his own and Harnad’s understanding of the problem.

“Symbol grounding in the sense of linking symbols (symbolic representation) to lower-level (subsymbolic) processes (Harnad 1990) provides a partial answer to the intentionality question. But it does not fully answer the question.” (Sun 2000, p. 6)

He explains why he sees the soft problem only as a part of the whole symbol grounding problem. “This is because the issue of *how* grounded symbols, and associated subsymbolic processes, acquire their intentional content remains.” (ibid., p. 6)

As a consequence, Sun argues “that, instead of being narrowly and technically conceived, symbol grounding should be understood in a broadened context in order to fully address the intentionality question, which is at the heart of the matter.” (ibid., p. 6)

A preliminary reply to Rodriguez et al. in the spirit of Sun could therefore be that ‘providing internally manipulated symbols with concepts’ is fundamentally impossible without ‘intrinsically meaningful internal states of mind’. And that ‘the ability to pick out referents’ is impossible as well without ‘making sense of symbols’. Solutions to the former two but not to the latter can only contribute to an incomplete model of cognition.

Harnad’s main intention might not be to provide an explanation for content in general. His assumption that the frame problem and the problem of vanishing intersections are only sub-problems, however, indicates that he concedes a particular relevance to the symbol grounding problem that cannot be explained by a soft interpretation alone.

Consolidating his argument with the philosophical concept of intentionality, eventually, propagates a wider problem with the grounding of symbols. A problem that is neither solved by his hybrid design, nor by Steels' 'Talking Heads', or any solution to the soft symbol grounding problem.

4.2.1. Beyond Harnad

The general approach to the soft symbol grounding problem is to implement non-symbolic representations as the supposedly autonomous content of symbolic representations. In a reply to Searle, Harnad emphasises that there are ways of *non-symbolic processing* that could very well be fit for this purpose.

"A nonsymbolic code is one in which the relation between the symbol tokens and what they stand for is not arbitrary or conventional, but governed by physics in some way, such as through reliable causal connections between similar physical properties such as shape." (Harnad 1989, p. 14)

He supposes that icons are physically determined and, therefore, independent from a designer. As a consequence, the Chinese room argument does not apply to computational symbol grounding based in *iconic* representation (i.e. his hybrid design). (ibid.)

For one, the content an icon can have for *one* system is initially independent of the content that the same icon can have for *another* system. Remember the reversible duck-rabbit in figure 2.2 on page 17. Even to *one and the same system,* an icon can mean different things. Among different systems, this variance can only increase—not even considering the possibility that these systems might not share a similar sensorimotor apparatus or bodily constitution.

An icon does not simply present a referent as is. Instead, icons must be interpreted according to particular conventions just like symbols. Harnad's iconic projections, for example, follow a very specific mapping from the structure of the referent to the structure of the icon. For Harnad to speak about the iconic projection of *a zebra* already introduces his own ideas about what makes a structure be the icon of a zebra.

If the basic perception of an artificial system is determined by *human* means of processing, then its concepts are essentially human as well. With an essentially non-human constitution, however, this puts the system at a considerable practical disadvantage.

The Chinese room argument applies to *all* cases where the system's most basic perception is determined beforehand, be it by selecting a particular set of symbols or by selecting particular means of processing according to van Gelder. This is why Searle rejects Harnad's hybrid approach to the symbol grounding problem.

The importance of external content in Harnad's design is most clear in the dependency of its icons on ‘feedback indicating their names’. This feedback origninates outside the system. Therefore, a cognitive system according to this design is fundamentally incapable of truly *autonomous* concept creation.

4.2.2. Beyond Searle

Consider the following, where Searle provides a concise repetition of his point against computational implementations of autonomous content in general. (Searle 1987, pp. 231–232; Searle 1990b)

Axiom 1. Computer programmes are formal (syntax).

Axiom 2. Human minds have mental contents (semantics).

Axiom 3. Syntax itself is neither constitutive nor sufficient for semantics.

Concl. 1. *Programmes are neither constitutive nor sufficient for minds.*

This argument applies to all computer programmes, independent from whether they implement iconic or symbolic representations. In light of this argument, imagine the Chinese room one last time. Instead of a rule book or water pipes, now suppose there is a Chinese speaking person handing Searle the appropriate symbolic responses.

A Chinese person that takes the right functional role in the Chinese room would understand the input signs. The system now features a component that understands Chinese and therefore the system also features mental content. In this setting, however, it still appears intuitively inappropriate to say that *the room* understands anything. Why is this?

David Chalmers provides an explanation.

> Superficially, there is something quite compelling about the argument: “Computers can only engage in formal symbol manipulation. These symbols are meaningless. Therefore computers cannot understand.” But such an argument draws its force entirely from the conflation of representations with computational tokens, under the term “symbol”. (Chalmers 1992, p. 20)

Chalmers argues that the thrust behind the Chinese room argument comes from a conflation of mental representations and computational symbols. This, in turn, affords a conflation of real systems and their formally syntactic *descriptions.* If you consider the room as a description, then the Chinese room argument merely shows that the *symbolic description* of a cognitive system cannot feature mental content.

Searle himself makes a very similar point when he argues that "it is obvious that we cannot explain how typewriters and hurricanes work by pointing to formal patterns they share with their computational simulations. Why is it not obvious in the case of the brain?" (Searle 1990a, p. 32)

His case in point appears obvious because simulations work *in accordance* with formal patterns but are not *identical* to them. To explain typewriters and hurricanes, requires to refer to the physical systems that behave according to these patterns.

In the same way, the behaviour of a running programme can be described, *but not explained,* by pointing to the computational symbols that implement it as formal patterns. Actual simulations merely work *in accordance* with formal symbolic patterns but are not *identical* to them. To explain them, it is necessary to also consider them as physical systems.

After all, it is obvious that we cannot explain how brains work by pointing to formal patterns they share with their descriptions. Why is it not obvious in the case of computational simulations?

4.3. A New Route

Harnad's interpretation of symbol grounding as a soft problem has lead to a variety of practical systems. However, it also lead to a neglect in the development of solutions to the hard problem (mentioned also by Sun 2000, footnote 6).

The soft interpretation distracts from a problem that is important far beyond the cognitive sciences: "[the significance of the symbol grounding problem] is not limited to artificial intelligence, for to solve [it] is to give a satisfactory account of how meaning emerges in the natural world, or, in other words, *to naturalise semantics.*" (Bielecka 2015, p. 79, emphasis added)

Approaches to the soft symbol grounding problem concern the development of symbolic conventions between cognitive systems but they *do not even touch* the naturalisation of semantics. Systems that implement a solution to the soft interpretation always depend on content that is external to the system itself—be it in the designer or in a 'community' of other systems.

The hard interpretation of the symbol grounding problem highlights this and it therefore enables to ask how real cognitive systems can actually surpass this limitation. It suggests that basic perception is not only the building block for mental content but that basic perception itself carries a particular type of content that must be acquired autonomously.

The Chinese room argument is a point against mental content in *formal procedures* but not against content *in the physical instances* that they describe. After all, it is hard to identify computational symbols in the electric activity of a computer during the execution of a symbolic programme.

The relevant symbols in the case of a computational simulation of mental representation are not computational symbols. Instead, symbolic mental representation can only be simulated by arranging computational symbols such that processes that act according to this arrangement yield states that are functionally similar to mental states.

From this, the following premise can be inferred that enables to avoid the transfer of content onto the system: as long as an artificial system does not register a certain type of its own structural elements, these elements can feature any kind of external content. Without an immediate influence on the system, there cannot be any transfer.

As a consequence, the content that these elements have *to an observer* also has no influence on the content that complex structures—which the system composes of them—have *to the system.* The system's interpretation of its internal structure cannot start with the atomic elements of this structure but no sooner than with *compositions* of these elements.

The content that these compositions have to the system does not depend on the system's designer because, the designer *does not know the content of these structures* and, *to the system, there simply is no content prior to these compositions.*

This proceeding rejects a central premise behind the Chinese room argument: formal symbol manipulation makes autonomous content impossible. This premise is replaced by a fundamental uncertainty concerning the content of another system's internal representations. Only this agnosticism enables the generation of autonomous content in any system beside oneself the first place.

To reject the Chinese room argument does not imply that the circuitry of computers buzzes with metal content. It merely shows that, just because there is no content in the formal description of a system, does not mean that there is no content in the system itself.

5. Conclusion

The previous part provides an overview about what the actual problem is, the different ways its elements can be conceived of, and how these conceptions determine what will be accepted as a solution. In the light of this discussion, the creation of basic mental concepts is considered to be at the heart of the frame problem, the problem of vanishing intersections, and the symbol grounding problem.

With a subjective conception of the mental representations that enable mental concepts in the first place, the symbol grounding problem presents itself at the root of the frame problem and the problem of vanishing intersections. It is argued that a consequential subjective interpretation of mental representation requires to consider the problem at hand as more than just a case for linguistics but *fundamental to understanding in general* instead.

The symbol grounding problem shows an intrinsic deficiency in early conceptions of cognition. Reason for the long-lasting popularity of the symbol grounding problem, however, is an implicit confusion of subjective mental representations in real cognitive systems and the objective representations that stand for mental representations in descriptions of cognition like cognitive models.

Intentional psychology and representationalism are important theoretical foundations for cognitive modelling. A conflation of both, however, lends various paradoxical situations of which the symbol grounding problem is only one.

Without distinguishing a system from its model, parts of the description appear like parts of the real system. This is the case with real mental representations and the representations in a cognitive model which merely *stand for* real mental representations.

Searle's extensive work on language might support a soft interpretation of the symbol grounding problem. To some, his engagement with linguistics might even suggest an objective conception of mental representation. This, however, is quite far from his original intention.

Conflating mental representations with their descriptions impedes the semantic autonomy of the system that contains them. A description cannot, and is not supposed to be, independent from content in the mind of an observer. In fact, observations *are* phenomenal descriptions. The essentially purpose of a description is to employ and rearrange content that is already available to an observer such that it describes something new.

Interpreting the symbol grounding problem as the quest for a particular *description* of mental representation must necessarily violate methodologically solipsism. This interpretation, however, is in fact incentivised by the Chinese room argument. As the description of a cognitive system, the room can impossibly present any original mental content to an observer, instead it necessarily grounds in the observer's *own* mental representations.

Just like real cognitive systems, a comprehensive simulation of cognitive systems is not supposed to describe anything. A simulation that implements a subjective conception of mental representation is therefore able to preserve methodological solipsism.

The grounding of symbolic mental representations cannot only be based in structural features. A cognitively justified representation of the environment must itself be part of a greater context that describes the system's current situation.

The rest of this work extends on this thesis. It is divided into a theoretical and a practical part. Both describe content and context in the mental model of a cognitive system.

The following chapters show how the field of embodied cognition explains the initial emergence of basic perception, how the philosophy of mind describes the creation of mental models from this basic perception, and how semiotics describes the interplay of mental representations in a mental model.

First, the intentionality in subjective mental representations is presented. Theories on intentionality accord to representationalism (see section 3.1.1): mental models are not generated from a world of objects but from a reality of inconceivable things that are not accessible to immediate perception.

Second, various practical approaches to modelling cognition from the field of embodied cognition are explained. Most of these models renounce the need for mental representation and, with it, the actual need for a mental model of reality.

Symbolic representation is vital for cognitive systems to avoid carrying over complexity from concrete content to its phenomenal appearance in a more abstract context. The conception of mental representation that is concerned by the critique in embodied cognition is clarified in more detail.

Third, the theoretical part merges into the practical part with a formal description of intentional content as the structure of mental models. This formal description allows to design an algorithm that generates such a structure on its own.

In the remaining chapters, the computational procedures for a system to create its own mental model of reality is provided. In the last part, the implementation of this algorithm enables to eventually test the three practical theses from section 1.3.2.

Part II.

Models of Embodied Cognition

6. Introduction to Embodied Cognition

Traditionally, mental representation is conceived of as a *symbolic* relation between shape and referent. Research programmes that explicitly criticise symbolic conceptions of mental representation include *enactivism*, *embedded cognition*, or *extended cognition.* Each of these labels emphasises various different factors that might be relevant in overcoming the problems from chapter 2.

At first sight, these approaches do not share much despite common critique. On closer inspection, however, all agree on the central thesis of *embodied cognition:* To understand cognitive systems requires to consider their physical bodies. In the following, a survey of the most influential theories in embodied cognition is presented.

The methods of embodied cognition are diametrically opposed to classical-symbolic explanations of mental representation. This is also why they are particularly successful in areas where classical approaches fail. The most notable successes include, for example, rapid and reactive interaction with the environment and learning structural features from natural data.

The presentation of these approaches serves the purpose to show how the initial generation of mental content is conceived of in the *natural* sciences outside the *subjective* frame of the philosophy of mind. Knowledge about theories from embodied cognition enables to realise considerable overlaps with functional descriptions from the phenomenological camp in the next part.

6.1. Embodied Cognition as an Explanation for Basic Perception

The previous part presents arguments that support the assertion that mental representation must be grounded in elements below basic perception. Although basic perception appears atomic to first-person-perspective, for it to be grounded in reality, basic perception itself requires an internal structure that is in a causal relation with external events. To avoid the transfer of content, this structure must consist of mental elements that the system itself is not aware of.

In embodied cognition, connectionist methods like artificial neural networks are frequently used to simulate processes which operate on *pre-conceptual* mental entities (e.g. individual neurons). These entities stand in causal relations with external events but, individually, they do not mean *anything* to their system. Accordingly, these methods are also called 'sub-symbolic'. As such, they provide a way to ground basic perception that avoids the use of content from anywhere else.

In Harnad's hybrid approach, however, these elements are already arranged according to the designer's own ideas on how an icon is supposed to resemble its referent. This defeats the purpose behind the autonomous generation of basic perception because it reintroduces a relation to external content (see section 4.1).

This external content is necessary for his system because it learns *under supervision.* Each iconic arrangement of representational elements must conform to an externally determined representational template/scheme that is determined by a 'teacher'. The information received by the system, therefore, undergoes a prior selection and adjustment to serve as examples for what the system is supposed to learn in the first place.

This supervision is specific to Harnad's design but not inherent to connectionist methods in general. Embodied cognition contains various theories on the generation of basic perception with connectionist methods that are *unsupervised.* These theories show how to ground content in information that is exchanged between system and environment *without the system's awareness.*

This imperceivable information exchange can be simulated by computational symbols without the danger of transferring content. This is simply because those computational symbols simulate nothing that the system is supposed to be aware of.

Therefore, embodied cognition's various explanations for the emergence of basic perception from pre-conceptual elements provide a solid foundation for simulating the generation of symbolic mental representations—independent from the individual explanation's stance on symbolic mental representation.

6.2. Overview

In the remainder of this part, theories of embodied cognition are presented that do not regard basic perception as ultimate bedrock of cognition. These theories are part natural, part phenomenological, explanations for the pre-conceptual processes that *generate* basic perceptual categories. Theories of embodied cognition enable to establish the connection between an external physical environment and the mental 'inner workings' of a cognitive system.

Two main branches of embodied cognition deal with the shortcomings of purely symbolic systems. The first one intends to get rid of symbolic representations for good. The second one tries to reflect on the reasons why specifically *symbolic* representations have failed to yield more generally intelligent behaviour so far and on how symbolic systems might need to be extended. The following survey is divided accordingly.[1] Some of the approaches that explain symbolic representation *ground* in approaches that are originally intended to replace it, others develop their own theoretical foundation.

The first chapter contains explicitly anti-representationalist models of cognition. The rejection of mental representation is mostly due to the historical shortcomings of formal symbolic approaches. All models in this chapter reject a very similar understanding of mental representation. Particular emphasis is put on exposing this understanding

The second chapter contains models that try to consolidate concepts of mental representation with a non-symbolic foundation. These approaches do not justify the content of mental representations in virtue of prior convention like Harnad and Steels (i.e. they do not try to solve the soft symbol grounding problem from section 4.1) but they consider representation in a formally weaker sense instead. Particular emphasis is put on the extend to which this mitigates a symbolic conception of representation.

From the presentation, the importance of two major methodologies that are relevant for the computational implementation of a cognitive model is inferred. One is *connectionism* and the other is *dynamic systems theory.*

The framework of connectionism provides a formalism for structural representations that literally ‘contain’ information. The framework of dynamic systems theory enables to address the reactive causal coupling between system and environment.

[1] Of course, this is not the only way of breaking this heterogeneous field down into more comprehensible bits. However, the borders of such a segmentation must be at least as blurry as the self-perception of the approaches involved. For alternative taxonomies, see Wilson (2002), Gibbs (2006), Calvo and Gomila (2008), Shapiro (2011), and Lyre (2013), among others.

7. Anti-representationalist Models

The conception of mental representations that is attacked by anti-representationalist positions has its origins in *propositional representations.* This kind of representation has been suggested, for example, by Zenon Pylyshyn and it is compositional, symbolic, explicit, abstract, and amodal. (e.g. Pylyshyn 1977)

This conception was intended as a counter reaction to the idea of *visual imagery* that was suggested by Stephen Kosslyn and James Pomerantz, among others. They argued for *iconic* mental representations that literally *resemble* their referents. (e.g. Kosslyn and Pomerantz 1977)

This antagonism continues to have effect on debates including, but not limited to, whether mental representations are analogue or digital (e.g. Goodman 1976 vs. Lewis 1971), depictive or descriptive (e.g. Kosslyn and Pomerantz 1977 vs. Pylyshyn 1977), spatial or propositional (e.g. Kosslyn 1980 vs. Pylyshyn 1979), and intentional or phenomenal (e.g. Dretske 1997 vs. Block 1996).[1]

The anti-representationalist position holds that a formally strict concept of symbolic representation is inadequate to describe mental states and therefore also inadequate to simulate essential characteristics of cognition.

As a consequence, anti-representationalists need to deliver a model of cognition that is less formal but still covers complex instances of subjective mental representation such as imagination—which, in fact, manifests as a *representation* for parts of the world.

7.1. Development as Dynamic Interaction

Esther Thelen and Linda Smith present their stance on mental representation from a developmental perspective. They reject the idea that object concepts are realised as representations altogether.

> Representations in their strongest, original, and most meaningful sense are symbols that stand for what is represented and are distinct from the computa-

[1] This reference to the historical debate suggests that both positions might have developed their conception of mental representation also to contradict popular conceptions of the other side.

> tional processes that operate on them. By this original definition, sensorimotor processes are decidedly not representations. (Thelen, Schöner, et al. 2001, p. 72)

According to Smith and Thelen, object concepts do not result from representational structures but from ongoing sensorimotor processes instead. They argue that

> [T]here is no such thing as an "object concept" in the sense of some causal structure that generates a thought or a behavior [. . .] There is only "knowledge" of objects as embedded in the immediate circumstances and the history of perceiving and acting in similar circumstances. (ibid., p. 34)

The mind as a whole, in their view, *is* this the dynamic interaction between system and environment. These mutual influences occur in parallel, decentralised, and in real-time. Such processes can be described particularly well as a dynamic system. (Smith and Thelen 2003)

Thelen, Schöner, et al. distinguish between an external, third-person-perspective on 'causal structures' and an internal, first-person-perspective on embedded 'knowledge'. They specifically reject representations in the sense of *physical* structures that store abstract object concepts.

7.1.1. The A-not-B Error

The A-not-B error can be interpreted to support this claim. It is a major object of study for dynamicist approaches to cognition. (Thelen, Schöner, et al. 2001)

Children between the ages of eight and twelve months are prone to this error. It occurs while they are actively looking for a desirable object. If the object was overtly hidden at place A, the child will search the object at A accordingly. However, if the object was hidden at A several times but is eventually hidden at B, the child will *still* search for the object at A.

Why does the child assume the object at A when it clearly saw that it was hidden at B?[2]

[2] Jean Piaget interpreted this observation as the lack of a concept of object permanence. However, notice that Piaget meant this to show that the infant has not yet learnt the permanence of objects *in general,* not that they have not learnt the permanence *of the particular object*—the object *as such,* so to say. (Piaget 2013, pp. 50–65)

Later experiments in Bower (1982), Baillargeon and DeVos (1991), and Ahmed and Ruffman (1998) indicate that vanishing objects evoke irritation in infants as soon as with the age of three and a half months. This suggests that object permanence may come gradually with growing sensorimotor experience.

According to Thelen, Schöner, et al. (2001), the concept of a each particular object emerges from continuous sensorimotor interaction between cognitive system and environment. These interactions are *at the same time* cause and effect for consistent behaviour in the child.

When the infant searches at A, it becomes more likely that, next time, it will search at A again. This behaviour is reinforced by previous instances that have been rewarded by finding the object.

Initially, this reinforcement is independent from intermediate observations. An adult knows that observing the object being hidden at B refutes the assumption that the object is at A. The child itself, however, does not yet realise the relation between *hiding* and *finding* an object.

If the task is understood as a reinforced sensorimotor process, then it is no surprise that the infant repeats what it experienced to be successful before. Without any counterexamples, reward received by the child appears to depend only on the action 'search at A', not on the previous observation that an indistinguishable object is hidden at B.

The child first needs to learn what observations are relevant for mentally persisting an object during its absence. Situations like the A-not-B error are evidence for such learning. Shapiro offers an illustrative analogy.

> In effect, the A-not-B error is hardly more mysterious, hardly more in need of an explanation in terms of rules and representations, than the fact that a road map is more likely to be folded along its old crease lines than new ones, or that a book is more likely to open on a favorite page than to a seldom-read one. (Shapiro 2011, p. 61)

According to this analogy, the crease corresponds to a concept of the desirable object and the act of folding corresponds to an action that yields this object. The example illustrates the idea that concepts are essentially a cluster of motor habits. Without the act of folding there is no crease and without a crease the map simply folds elsewhere.

7.1.2. Concerning the Symbol Grounding Problem

Thelen, Schöner, et al. argue that the purpose of any cognitive development is a skilful combination of reactive and deliberate behaviour: "the critical developmental process may not be transgressing some line dividing the conceptual and the perceptual-motor—the traditional issue—but the ability to use memory and to make decisions off-line when the situation demands." (Thelen, Schöner, et al. 2001, p. 33)

Only if we designate separate "symbolic and conceptual codes for the purely 'mental' part and dynamics for the perceptual-motor" (Thelen, Schöner, et al. 2001, p. 34) part, "[w]e come up against both the symbol-grounding problem and its inverse: how do symbols (concepts) arise from perception and how does the symbolic (conceptual) code get transduced into the dynamics of movement?" (ibid., p. 34)

This analysis of the symbol grounding problem is supported by the fact that "skilled people shift rapidly and effortlessly [...] between acting immediately and tightly coupled to the input, and delaying action in favor of remembering and planning." (ibid., p. 34)

Such rapid shifts in different situations suggest that complex cognitive skills such as language comprehension are not only dependent on abstract concepts. Instead, abstract concepts and reactive intuition seem to be much more interwoven than is traditionally assumed.

According to Thelen, Schöner, et al., the symbol grounding problem is not as big of a problem as it is commonly understood. The mitigating factor is that concrete action and abstract cognition might actually not be that different. People can act "so seamlessly only because acting and thinking are in commensurate dynamics." (ibid., p. 34)

Complex cognitive skills can just as well result from a reactive and dynamic coupling with the environment. Object concepts might not require permanent grounding, because they merely *support* reactive interaction. They only need to persist for a relatively short while, considerably weakening the symbol grounding problem.

In summary, Thelen, Schöner, et al. specifically reject symbolic mental representation in the sense of *causal structures* that correlate with *object concepts.* They state instead that concepts emerge with the acquisition of motor processes that apply to the respective object. In their view, a concept *is* knowledge on how to act upon an object.

As a consequence, the symbol grounding problem is asking too much. In an appropriate model of cognition, symbolic representations do not need to be grounded in the first place. The ability to acquire concepts as we observe it in other systems is in fact not the grounding of symbolic mental representations but instead the development of a set of motor skills that apply to a particular object.

7.2. Affordances and the Ambient Optic Array

At the physical border between cognitive system and environment both exchange raw sensorimotor information. To make an informed decision, cognitive systems need to filter huge amounts of this data.

First, sensorimotor samples have to be *recognised* and assigned to internal representations. Secondly, *reasoning* processes operate on these representations to evaluate relevance, desirability, and probability of possible outcomes in relation to the system's individual goals and abilities.

In a computational simulation, these processes become sedate and non-reactive when confronted with too much data. It is hard to determine just *the right* amount of detail for basic cognitive information processing.

On the one hand, if sensorimotor events are picked up that are not really important in the current situation then the processing effort is too much. This results in *the frame problem.* On the other hand, even the most extensive processing cannot make sense of the world if relevant information is missing. This results in *the problem of vanishing intersections* (see chapter 2).

In the brains of human cognitive systems, the thalamus is responsible for information filtering (e.g. Kobayashi and Isa 2002). What are the *functional* processes that this part of the brain implements? What is an appropriate heuristic for picking up just 'the *right* stuff' in the current situation?

7.2.1. Perceptual Invariants

James Gibson calls the relevant properties of sensorimotor data 'perceptual invariants'. Invariants are those segments from a stream of imperceivable information that remain relatively stable over time. The current 'situation', according to this understanding, is a set of invariant sensorimotor activations that the system itself is not aware of.

Gibson describes invariants with his concept of the ambient optic array. "The central concept of ecological optics is the ambient optic array at a point of observation. To be an *array* means to have an arrangement, and to be *ambient at a point* means to surround a position in the environment that could be occupied by an observer." (Gibson 1986, p. 65)

The structure of the ambient optic array is composed of individual sensorimotor information. Their shapes form the retinal projection of an external referent. When changing one's spatial position, for example, one necessarily also changes the arrangement and structure of the ambient optic array. Over time, some of its shapes change and some persist. Those that persist are *invariant.*

Figure 7.1 illustrates the difference in the ambient optic array between two perspectives on a window. Invariants are those visual features of the environment that are independent from perspective. In the figure, this is, for example, the visible part of the tree which remains relatively constant between both perspectives.

The information in invariants is context-dependent and temporally extended. This provides an important advantage over raw sensorimotor activation. The sampling rate of

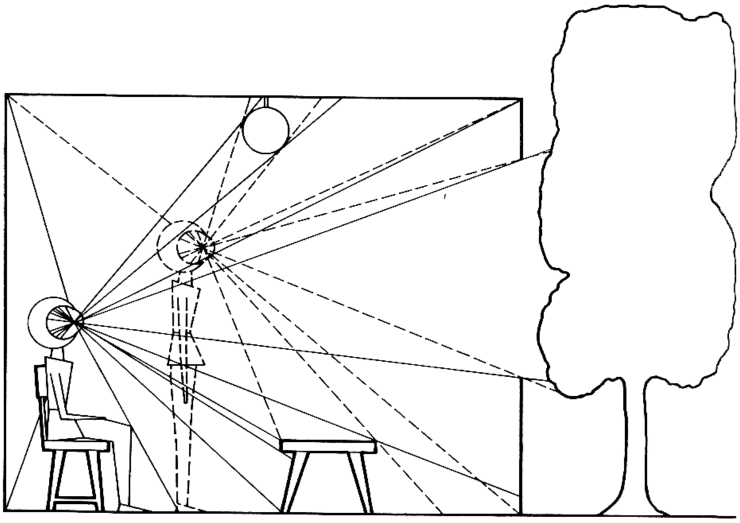


Figure 7.1.: Invariants in the Ambient Optic Array (Gibson 1986, p. 72, figure 5.4).

invariants is considerably lower than the sampling rate of raw sensorimotor information. This is due to the fact that invariants *extend* over time. Raw sensorimotor information, in contrast, is a steady and relentless torrent of incoming and outgoing data. The temporal segmentation of this torrent into invariants enables processing to keep pace.

To perform this segmentation, only changes in immediate values need to be considered. There is no need for abstraction, planning, or domain-specific knowledge. Each of the resulting slices can serve as a basic perception that stands for a relevant aspect of the current situation.

Combinatorial complexity that is lost due to the reduced number of elements *over time* is compensated for by an increased total number of elements. Consider a case where, after each time step, there is exactly one of n singular activations. Over a sequence of t time steps, the number of possible combinations would be n^t.

If the same time interval is covered by invariants that last on average $0 < p < t$ time steps, then there are n^p different elements that occur $\frac{t}{p}$ times on average. The resulting number of possible combinations is $(n^p)^{\frac{t}{p}}$, which is exactly n^t. Instead of after every single time step, however, new information arrives only every p-th time step on average.

7.2.2. Affordances as Basic Perception

Especially important are invariants that persist during changes that are initiated by the system's *own* actions. In figure 7.1, for example, the visual impression of an outside tree is such an invariant.

These invariants are inherently related to an action. The particular set of invariants perceived in the presence of a tree, for example, can afford the act of climbing—during which the current visual impression remains invariably *tree-like.*

Accordingly, Gibson calls these invariants 'affordances'. Affordances appear to their system as *basic perception.* (originally in Gibson 1986; also in Millikan 2004, chapter 4; Millikan 2006)

Given these points, affordances seem to be considerably well suited as a foundation for mental representation. They occur in moderate frequency and do not require sophisticated processing, thus alleviating stress on cognitive processes. The heuristic that enables to identify *relevant* information is therefore *invariance during action.*

This heuristic integrate sensor data and motor processes. Gibson's affordances describe sensor aspects of the environment but at the same time they also describe motor aspects of the system. As a consequence, therefore, they exploit the system's embodiment while, in general, they remain agnostic on bodily particularities.

According to Gibson, affordances are not cognitive but part of the external environment (i.e. *laws* about the relation between sensor and motor activations). However, affordances can be memorised. Once they are learnt, they can be retrieved and used to predict sensor activations or select motor activations in a similar situation.

Gibson's approach is anti-representationalist only with respect to the abstract, amodal, and logically descriptive representations that are rejected by Thelen, Schöner, et al. as well.

However, like Thelen, Schöner, et al., Gibson also considers cognitive correlates for the dynamic interaction between system and environment even if he does not call them 'representations'.

7.3. The Cognitive Subsumption Architecture

The motivation of Rodney Brooks's model is frustration with the abilities of physical robots. According to him, the reason for this lies in their *architecture,* where symbolic representation separates perception and action. According to these architectures, perception and action are supposed to take place in mutually exclusive regions of a cognitive system and are in most parts independent from one another.

Sensor activations are used to generate abstract representations of the environment. With these abstract representations, *planning* is performed and, once it is finished, the resulting motor activations are executed. Because of this serial process, Brooks calls the underlying frameworks ‘sense-model-plan-act architectures’. (Brooks 1991)

Sense-model-plan-act architectures try to describe the environment with a symbolic world model on which computational processes can operate. Once these processes are finished, they produce a motor activation which is expected to realise the change intended by the system.

In sense-model-plan-act architectures, intelligent behaviour is simulated by implementing functional layers that operate on symbols of different degrees of abstraction. Sensor devices pre-process activations, the results are used to generate a model of the environment, the system uses this model to plan its action, it performs this actions, and eventually the actions result in the activation of motor devices. Figure 7.2a illustrates this process.

Robots controlled by a sense-model-plan-act architecture do not interact with their environment directly but with a mediate model of it instead. This model is assumed to be correct and precise. It consists of representations that are stationary and, once generated, largely independent from the *actual* state of the environment.

Cognition in sense-model-plan-act architectures cannot be dynamic because it acts upon a model, not directly upon the external environment. Robots that realise these architectures are ‘displaced’, instead of being *embodied* and *situated* in the environment. According to Brooks, this explains some of their limited abilities.

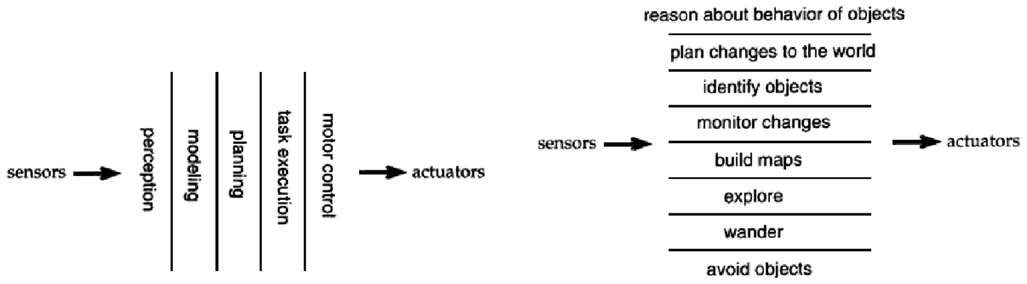
Brooks’ subsumption architecture “grew out of dissatisfactions with traditional robotics and AI, which seemed unable to deliver real-time performance in a dynamic world” (ibid., p. 1227). How can this be remedied?

7.3.1. Nouvelle Artificial Intelligence

Brooks emphasises that sense-model-plan-act architectures are not the only venue to intelligent behaviour.

> The traditional methodology bases its decomposition of intelligence into functional information processing modules whose combinations provide overall system behavior. The new methodology bases its decomposition of intelligence into individual behavior generating modules, whose coexistence and cooperation let more complex behaviors emerge. (Brooks 1990, p. 3)

Central to the new method of ‘nouvelle artificial intelligence’ (also pursued, for example, in Rosenschein and Kaelbling 1986 or Agre and Chapman 1987) is *situatedness* and *embodiment*. Brooks (1991, p. 1227) defines both concepts as follows.



(a) The sense-model-plan-act architecture (Brooks 1999, p. 4, figure 1).

(b) The subsumption architecture (Brooks 1999, p. 5, figure 2).

Figure 7.2.: Architectures of Conventional and Nouvelle Artificial Intelligence.

- *Situatedness:* The robots are situated in the world—they do not deal with abstract descriptions, but with the 'here' and 'now' of the environment that directly influences the behavior of the system.
- *Embodiment:* The robots have bodies and experience the world directly—their actions are part of a dynamic with the world, and the actions have immediate feedback on the robots' own sensations.

None of this applies to robots that are controlled by a sense-model-plan-act architecture. They sent their sensor measurements to a central instance which tries to reconstruct a complete set of mental representations.

Cognitive processes according to nouvelle artificial intelligence, in contrast, are parallel, independent, and specialised instead of sequential, hierarchical, and generic. Figure 7.2 compares sense-model-plan-act architectures and nouvelle artificial intelligence.

Learning according to nouvelle artificial intelligence works quite differently from learning according to conventional artificial intelligence. "In nouvelle AI each module itself generates behavior, and improvement in the competence of the system proceeds by adding new modules to the system." (Brooks 1990, pp. 3–4)

Classic artificial intelligence systems improve by learning new symbolic representations that can be processed by generic rules of inference. In nouvelle artificial intelligence, systems become more proficient by learning specialised sensomotoric skills that solve particular problems.[3]

[3] As a consequence, Dreyfus appreciates Brooks's critique on classic artificial intelligence but also criticises a lack of sophisticated abilities in subsumption architectures. "Brooks's robots respond only to *fixed isolable features* of the environment, not to context or changing significance." (Dreyfus 2007, p. 335)

7.3.2. Grounding Mental Representations

To Brooks, grounding is not motivated by philosophical, but by *practical,* disadvantages of conventional approaches. He argues that the symbol grounding problem can be overcome with the physical grounding hypothesis. The physical grounding hypothesis "states that to build a system that is intelligent it is necessary to have its representations grounded in the physical world." (Brooks 1990, p. 6)

According to him, mental representations cannot be said to be grounded if they are not under permanent sensorimotor influence. Conventionally, however, once representations are generated, they are mostly independent from what they describe. This dissociation between the *internal* model of an autonomous system and its *external* environment gives rise to severe practical problems in a complex real-world situation that requires reactive performance.

According to Brooks, the performance of a robotic agent benefits greatly from a tight sensorimotor coupling with its environment. Sense-model-plan-act architectures do not consider such a coupling.

Without symbolic representation, elaborate and time-consuming inference processes do not even seem to be necessary. "Our experience with this approach is that once this commitment [for nouvelle AI] is made, the need for traditional symbolic representations soon fades entirely." (ibid., p. 6)

Brooks does not suggest that mental representation is irrelevant or redundant *in general:* "[A] careful reading shows that I mean intelligence without conventional representation, rather than without any representation at all" (Brooks 1999, p. 79). Instead he proposes to merely *start* research with reactive interaction and work one's way up to more abstract cognitive processes from there.

Even Brooks, famous for his supposedly anti-representationalist stance, is merely *reacting* to the extreme claims and aspirations of classic artificial intelligence. He demands that representations should be grounded in direct sensorimotor interaction between agent and environment instead of floating among abstract ideas.

On closer inspection, even positions that are often regarded as exemplary anti-representationalist turn out only to be directed against a very specific and formal notion of representation.

The question remains how to realise Brooks' suggested step from dynamic sensorimotor interaction to a more abstract representation of the environment. To answer this question, requires to take into account what makes both appear so incompatible to begin with.

The shape of a symbol either *does* represent a particular referent or it *does not.* If this relation is digital and binary, then there is not much room for a dynamic interaction between both. At first glance, symbolic representation appears like the exact antithesis to a dynamic coupling between system and environment.

To determine whether a shape is the symbolic representation of a particular referent requires that this shape is *already* associated with a particular content that enables to validate this referent as ‘appropriate’. Central to the symbol grounding problem, however, is the generation of exactly this *initial* content. It cannot be solved with representations that are already supposed to represent something because this implies that they already have this initial content.

8. Hybrid Models

However, a comprehensive cognitive model describes not only reactive interaction but more abstract cognitive processes as well. In contrast to the previous models, it must be a *hybrid* that integrates the pre-conceptual processes from the previous chapter with processes of abstract reasoning that operate on symbolic representations. In contrast to Harnad's hybrid model and Steels 'Taking Heads', however, it cannot be supervised by another system.

Thelen, Schöner, et al. argue that basic and abstract cognitive processes are not so different after all. Following such a holistic approach, the generation of abstract concepts can be considered to be performed *by the same general procedure* that generated basic perception. The structures that are generated by this procedure effectively serve as content for *all* of the system's representations of the environment.

Brooks already proposed his subsumption architecture as the basis of more abstract cognitive processes. A hybrid model implements his requirements for real-time performance in a dynamic world and Gibson's proposal to describe basic perception *based in* imperceivable basic sensorimotor interaction. This enables to solve the philosophical as well as the practical implications of the symbol grounding problem.

In this chapter, a select subset of hybrid models of cognition is presented, each of which describes how mental representation emerges from basic sensorimotor interaction. Particular importance is attached to the methods used by these models.

This enables to understand Harnad's proposed architecture from section 2.3.3 *in context.* The models in this chapter are state-of-the-art ideas on how basic concepts emerge in cognitive systems from pre-conceptual entities. Any *new* alternative must be measured against their explanatory power.

8.1. Conceptual Metaphors

According to George Lakoff and Mark Johnson, the categories of basic perception result from a differentiation bottleneck in biological information processing.

> To take a concrete example, each human eye has 100 million light-sensing

> cells, but only about 1 million fibers leading to the brain. Each incoming image must therefore be reduced in complexity by a factor of 100. That is, information in each fiber constitutes a "categorization" of the information from about 100 cells. (Lakoff and Johnson 1999, p. 18)

The assignment of information from cells to optical fibres is a case of categorisation. In the example, this categorisation is not determined by semantics but *biologically necessary* due to the reducing degree of freedom at the causal connection from cells to fibres.

This categorisation is considered as the foundation of subjective experience. "[Categories] are the structures that differentiate aspects of our experience into discernible kinds. Categorization is thus not a purely intellectual matter, occurring after the fact of experience. Rather, the formation and use of categories is the stuff of experience" (ibid., p. 19). Basic perception is therefore a natural consequence from such a reduction in complexity.

Lakoff and Johnson believe a bottleneck like this is not only responsible for basic perception but also the reason for abstract categories. "Neural categorization of this sort exists throughout the brain, up through the highest levels of categories that we can be aware of." (ibid., p. 18)

As a consequence, the generation of categories in not under the system's immediate control. "We do not, and cannot, have full conscious control over how we categorize. Even when we think we are deliberately forming new categories, our unconscious categories enter into our choice of possible conscious categories" (ibid., p. 18). Even abstract categories are "an inescapable consequence of our biological makeup." (ibid., p. 18)

Traditionally, content is understood *compositionally,* as a function of the content that it consists of (see, for example, Szabó 2017). However, it is not clear how to determine this function. How exactly is constituent content composed to express more complex content?

According to Lakoff and Johnson, this function is *metaphorical.* The structure in basic content provides a *scaffold* for complex content. In this sense, the content of basic perception 'repeats itself' up until the highest layers of cognition. It provides helpful analogies in virtue of which more complex content can be understood. (Lakoff and Johnson 1980)

To say that 'love is a game', for example, allows to draw an analogy between the prototypical features of games and those of human relationships. The meaning of 'games' provides a scaffold for understanding the more complex meaning of interpersonal relations. Metaphors provide a *generalisation bias* towards content that is somewhat analogical to what is already known.

Consider the following examples for an analogy between 'love' and 'game'. 'You can win or loose' means you can become very happy or very sad. 'You have to follow the rules' means to implicitly agree on loyalty and devotedness. 'You play against each other' if it turns out to be hard to compromise.

But how do you arrive at the meaning of 'game' in the first place? Metaphors cannot be the *only* source for concepts or we run into the symbol grounding problem all over again. Lakoff and Johnson are aware of that. "Are there any concepts at all that are understood directly, without metaphor? If not, how can we understand anything at all?" (Lakoff and Johnson 1980, p. 56)

To avoid the grounding problem in their model of cognition, there must be some kind of *first* content. According to Lakoff and Johnson, basic perceptual categories result from information processing bottlenecks in our bodily constitution. Their content is therefore not determined externally but by the *body* of the system. This provides a biological way out of the symbol grounding problem.

According to this theory, the bodily structure of a cognitive system is present even in its most abstract mental representations. "Most important, it is not just that our bodies and brains determine that we will categorize; they also determine what kinds of categories we will have and what their structure will be." (Lakoff and Johnson 1999, p. 18)

8.2. Meshing Affordances as Perceptual Symbols

Cognitive processes are often compared to linguistic processes. Language as the *expression of thought* promises to provide insights on its mental correlates. Lakoff and Johnson, for example, see abstract meaning as grounded in metaphors. The fact that people speak in metaphors like 'argument is war' reflects how their mind works: metaphorical.

Arthur Glenberg also consults language to solidify his claims about mental processes. His *indexical hypothesis,* however, does not use language as a model to derive from it a theory of the mind. Instead, he illustrates that even the abstract concepts in language are grounded in sensorimotor interaction. He does not *start* from linguistic entities like metaphors but tries to *ground them* in unconscious neural states.

To achieve this, Glenberg incorporates the theory of perceptual symbols from Lawrence Barsalou. The following gives an account on the structure and workings of perceptual symbols in the spirit of Barsalou before describing their application in Glenberg's indexical hypothesis.

8.2.1. Amodal Symbols

Basic perception is always in a particular mode. It is, for example, auditory, visual, or haptic. Abstract concepts, however, are not—they are *amodal.* Barsalou criticises models of cognition that employ amodal mental representation.

His alternative conception of "a perceptual symbol is a record of the neural activation that arises during perception" (Barsalou 1999, p. 583). The nature of perceptual symbols is primarily physical: "unconscious neural representations—not conscious mental images—constitute the core content of perceptual symbols." (ibid., p. 583)

A perceptual symbol also be the *potential* correlate of a mental state. "Although neural representations define perceptual symbols, they may produce conscious counterparts on some occasions." (ibid., p. 583)

Barsalou claims the essential problem of amodal symbols in cognitive sciences is that "their internal structures bear no correspondence to the perceptual states that produced them" (ibid., p. 578). He describes, that because

> the symbols in these symbol systems are amodal, they are linked arbitrarily to the perceptual states that produce them. Similarly to how words typically have arbitrary relations to entities in the world, amodal symbols have arbitrary relations to perceptual states. (ibid., p. 578)

Barsalou criticises the concept of amodal symbols for the arbitrary relation between their 'internal structure' and 'the perceptual states that produce them'.

He goes into detail on what he regards to be the conventional element of representation in cognitive sciences with the example of colour representation.

> The amodal symbols that represent the colors of objects in their absence reside in a different neural system from the representations of these colors during perception itself. In addition, these two systems use different representational schemes and operate according to different principles. (ibid., p. 578)

So, according to Barsalou, usually *two* representational schemes (i.e. mappings between representation and referent) are considered for two different types of mental representation. One for abstract reasoning during the absence of the external referent and one that resembles the referent somehow while the system interacts with it.

Employing both schemes *at once* implies a *gap by design.* This gap separates representation *in absence,* from representation *in presence,* of the referent. It is no surprise that the transition from one to another turns out to be problematic.

A perceptual symbol, in contrast, maintains parts of the structure of its referent. The transition from basic perception to perceptual symbols should therefore be much more natural that from basic perception to amodal representations.

8.2.2. Perceptual Symbol Grounding

Amodal symbols are in an arbitrary relation with their perceptual states. In a physically determined system, however, there are no truly *arbitrary* relations. This is known as 'the transduction problem'.

Perceptual symbols are generated by cognitive processes that monitor neural activation during experience. It is irrelevant whether the neural activation correlates with a basic perception or with an abstract concept: perceptual symbols simply *record* particular aspects of whatever neural state is present at the moment.

This state is recorded schematically, in virtue of its *invariant aspects.* Therefore, perceptual symbols maintain the perceptual mode of the neural state from which they result. They reside in the same system and are encoded in the same format as the neural state itself. (Barsalou 1999, p. 582)

Perceptual symbols need no external referent that caused them and they can remain indeterminate concerning the particular aspect that they refer to (similar to the disjunctive categories of MacDorman in section 2.2). These aspects can be context-dependent, dynamic, and compositional.

Barsalou suggests that there is simply no need for both representational schemes (i.e. modal and amodal). In his view, this assumption is substantiated by the fact that there is no convincing answer to the question how amodal symbols develop from perceptual states in the first place. "If we cannot explain how these symbols arise in the cognitive system, why should we be confident that they exist?" (ibid., p. 580)

In Barsalou's view, the symbol grounding problem is inverse to the transduction problem. Where the transduction problem is concerned with how amodal symbols are *generated* from modal perception in a physical system, the symbol grounding problem is concerned with how amodal representations are able to physically *refer* to the modal perceptions which they are about.[1]

This applies to basic perception as well as abstract concepts. Despite the fact that the mode of perception is central to his theory, Barsalou emphasises that he

[1] Notice that, throughout this work, the transduction problem is interpreted as implication and special case of the symbol grounding problem rather than its inverse: symbols signify particular referents *if and only if* those referents had a major role in their generation in the first place. In the special case of symbolic *mental* representations, they are properly grounded *if and only if* they have been 'transduced' from experience before.

> proposes a theory of knowledge, not a theory of perception. Although the theory relies heavily on perception, it remains largely agnostic about the nature of perceptual mechanisms. Instead, the critical claim is that whatever mechanisms happen to underlie perception, an important subset will underlie knowledge as well. (Barsalou 1999, p. 582)

He argues for the perceptual (i.e. modal) origins of abstract concepts, not for the origins of basic perception. This shows, however, how Barsalou's perceptual symbols make use of theories 'from the ground' like Gibson's affordances or Brooks's subsumption architecture can be used as a *foundation* to explain the generation of abstract concepts.

8.2.3. Meaning in Meshes

According to Glenberg, words and phrases obtain their meaning from perceptual symbols. During the learning of a language, linguistic expressions are *indexed* such that, in the future, the referenced perceptual symbols can be retrieved when the expression is encountered. (Glenberg and Kaschak 2002)

The content of perceptual symbols is a recording of neural states. As a consequence, they also contain the neural correlate to the affordances that the system perceived during these states. The perceptual symbol of an upright vacuum cleaner, for example, allows to infer its use as a coat rack from the neural correlate of the actions that the object afforded during perception. (ibid.)

This record of affordances enables to combine the affordances that are described by different perceptual symbols. A coat affording to be hung and a vacuum cleaner affording to hang something on can be *meshed.* This meshing composes perceptual symbols into 'simulators' that allow to repeat parts of the original experience.

Simulators can be primed by language. The command 'hang the coat onto the upright vacuum cleaner', for example, elicits a mental meshing of the compatible affordances in the perceptual symbols of 'coat' and 'vacuumm cleaner'. The command 'hang the coat onto the upright cup', in contrast, does not elicit a meshing. (ibid., p. 559)

The indexical hypothesis suggests that the mental simulation of interaction is crucial to language understanding. It "proposes that language is made meaningful by cognitively simulating the actions implied by sentences." (ibid., p. 559)

Glenberg and Kaschak test this hypothesis in an experimental setting, where subjects are supposed to evaluate the sense of articulated affordances while executing actions contrary to these affordances. If meaning is based in the simulation of affordances, the subjects' performance should be worse when forced to execute contradicting actions. The results support the indexical hypothesis. (Glenberg and Kaschak 2002)

Their prospect is that the indexical hypothesis may provide a *general* theory of human understanding that grounds, not only linguistic, but *all* abstract mental representation in the neural correlate to basic perception. “Although substantial work needs to be done to secure that possibility, that work may well be rewarded by an account of language and meaning firmly anchored in human experience.” (ibid., p. 564)

8.3. Sensorimotor Integration

Kevin O’Regan and Alva Noë give their own account on the unconscious processes *prior* to basic perception. They point out that any *natural* explanation for basic perception requires a leap of faith. This is due to the disjunctive epistemological domains occupied by basic perception from third-person perspective (i.e. physical observations) and basic perception from first-person perspective (i.e. subjective experience).

It can always be asked: why is exactly *this* physical realisation necessary for basic perception but not another? As detailed as any natural explanation may be, there cannot be a determinate answer to how subjective experience is physically caused.

The quality of experiencing a particular sensorimotor mode (e. g. visual or auditory) cannot be reduced to physical terms (e.g. their neural coding). Even if certain neuronal activation patterns *always* occur simultaneously with visual experience, the observer’s objective perspective is inevitably and essentially *different* from the subjective perspective of the observed system.

Any explanation from the perspective of a third person for phenomena from the perspective of the first person must end in the stipulation of axiomic correlations. Therefore, it is fundamentally impossible to falsify a hypothetical causal connection between observation and experience by experiment.

One symptom of this is the binding problem. O’Regan and Noë describe it as follows. “The fact that [neural] modules operate independently and are often localized in different cerebral regions, raises the question of how the separate streams of information ultimately come together to give us the unified perception of reality that we subjectively experience.” (O’Regan and Noë 2001, p. 967)

Localised neural modules are widely considered as the natural correlates for basic perception. They seem to *present* external events to the mind as subjectively unified instances. Objectively, however, they are spatio-temporally distributed across the brain. There does not seem to be any structural isomorphism between a subjective experience and its physical correlate. In a nutshell, the binding problem describes *the non-preservation* *of spatio-temporal unity from first- to third-person-perspective and vice versa.*

This implies that the different modes of basic perception may be physically indistinguishable. *Different* neural activations may even correlate with subjectively *identical* perceptions of the same modality. “Even if the size, the shape, the firing patterns, or the places where the neurons are localized in the cortex differ, this does not in itself confer them with any particular visual, olfactory, motor or other perceptual quality.” (O’Regan and Noë 2001, p. 941)

8.3.1. Sensorimotor Contingencies

The different modes of perception cannot be explained *neurologically.* O’Regan and Noë propose to explain the different modes of perception *phenomenologically* instead.

> Instead of assuming that vision consists in the creation of an internal representation of the outside world whose activation somehow generates visual experience, we propose to treat vision as an *exploratory activity.* [. . .] The central idea of our new approach is that *vision is a mode of exploration of the world that is mediated by knowledge of what we call sensorimotor contingencies.* (ibid., p. 940)

They propose to consider the modes of perception as different types of unconscious *interaction* between agent and environment. They argue that “what *does* differentiate vision from, say, audition or touch, is the *structure of the rules* governing the sensory changes produced by various motor actions, that is, what we call the *sensorimotor contingencies* governing visual exploration.” (ibid., p. 941)

Eye movements, for example, distort visual experience in a unique manner. If the gaze moves along a horizontal line, the retinal image is preserved. If the gaze moves from above to below this line then the retinal image changes drastically and in a particular relation to speed and direction of the eye movement.

Sensorimotor dependencies in the auditory domain are quite different. They follow rules that are determined by other bodily, environmental, and physical characteristics than vision (as described, for example, by psychoacoustics).

In O’Regan and Noë’s view, vision is *identical* to the active exploitation of learnt vision-specific sensorimotor contingencies by executing the certain motor activations. Sensor and motor activations are *equally* necessary for perception. (ibid., p. 943)

To have knowledge about sensorimotor contingencies does not imply knowledge about this knowledge. The expected retinal changes during eye movement which can be inferred from learnt sensorimotor contingencies, for example, are *unconscious.* The fundamental 'knowledge' in these contingencies is a necessary condition for experience, which is why it cannot be experienced *itself.* (O'Regan and Noë 2001, pp. 944–945)

Sensorimotor contingencies are quite similar to Gibson's affordances: "Gibson's notion of 'affordances' is undoubtedly strongly related to our present approach" (ibid., p. 945). Contingencies that separate the individual modes of perception, however, do not consist of *temporal* invariants but of *permanent* rules that apply to the different senses *in general.*

This difference reflects in the difference between *transitive* and *general* consciousness. While being *transitively* conscious, "your *feeling* of the presence of all the detail consists precisely in your [implicit] knowledge that you can access all this information by movements and inquiries" (ibid., p. 960). Transitive consciousness is skilful engagement with current circumstances in virtue of *implicit expectations.* It operates on unconscious sensorimotor states and, from them, it generates basic perception.

General consciousness, in contrast, is skilful engagement with circumstances in virtue of *explicit expectations.* "*Visual consciousness* in general, on the other hand, is a higher-order capacity. To be visually conscious in general is *to be poised* to become aware of a present feature (that is, to become transitively conscious of it)" (ibid., p. 960, second emphasis added). Transitive consciousness is a necessary premise for general consciousness.

8.3.2. Sensorimotor Contingencies as a Foundation for Phenomenology

O'Regan and Noë reject the traditional idea of basic perception in virtue of mental representation. They argue that visual perception, for example, "does not arise because an internal representation of the world is activated in some brain area. On the contrary, visual experience is a mode of activity involving practical knowledge about currently possible behaviors and associated sensory consequences." (ibid., p. 946)

For a particular experience to appear, a particular contingency needs to be active or 'more present' than another: "when a particular attribute is currently being seen, then the particular sensorimotor contingencies associated with it are no longer latent, but are actualized, or being currently made use of." (ibid., p. 945)

O'Regan and Noë describe this use in more detail.

> Indeed, there is no *"re"*-presentation of the world inside the brain: the only pictorial or 3D version required is the real outside version. What *is* required, however, are methods for probing the outside world—and visual perception constitutes one mode via which it can be probed. (ibid., p. 946)

Under this assumption, what is commonly regarded as randomly accessible knowledge in the mind of a cognitive system turns out to be part of the external environment instead. There are no mental representations of external referents, just effective means of probing the environment under different circumstances.

According to O'Regan and Noë, mental content is better understood as a description of useful sensorimotor dependencies that enable prediction and expectation. They argue that "the outside world acts as an *external memory* that can be probed at will by the sensory apparatus" (O'Regan and Noë 2001, p. 946). To maintain a meaningful concept of mental representations, therefore, sensorimotor contingencies should be rather considered as holding the information necessary to 'access', instead of to 'reproduce', the environment.

From a conception of basic perception as act of exploration follows that "to reflect on the character of one's experience is to reflect on the character of one's law-governed exploration of the environment" (ibid., p. 961). The sensorimotor account of vision and visual consciousness of O'Regan and Noë grounds basic perception explanations in such unconscious exploratory processes.

The neural structures that enable such sense-specific prediction are the closest thing to a physical correlate of perceptual experience. The hippocampus is currently the most probable brain region to fulfil this purpose (see, for example, Stachenfeld et al. 2017). But to O'Regan and Noë, the relevance of the physical ground of basic perception is secondary to its phenomenological ground.

In contrast to purely natural approaches, a phenomenological analysis is compatible with the study of *conscious* mental processes. "Our central aim above is to make clear that we do not believe that there is any incompatibility between the sensorimotor contingency theory and more full-blooded phenomenological project." (O'Regan and Noë 2001, p. 973)

Even more so, sensorimotor contingencies provide a rigorous foundation for phenomenological investigations of unconscious parts of the mind. "In this way, we believe that the kind of approach we lay out in this paper helps place phenomenology as an undertaking on solid ground." (ibid., p. 962)

8.4. The Encryption of Mental Content

It cannot be the case that every known fact about the world needs to be considered in each context. How can the mental representations be generated such that they contain only information that is relevant to the system?

According to all of the previous models for cognitive systems, the content of mental representations is eventually determined by the system's sensorimotor capacities. This raises the question what a mental representation would contain to a system with different capacities.

To simply transfer content from one system onto another would be similar to ‘copying’ neurological activation patterns. Haugeland described this as problematic.

> [E]ach individual’s *particular body*—his or her own muscular gestalts—functions like a large encryption key; and the pulse patterns coming down from the brain are the cryptograms, which are either meaningless, or they mean something only in conjunction with that particular body. (Haugeland 1993, p. 226)

Content that grounds in the sensorimotor interaction between environment and system is always ‘bodily encrypted,’ just like the neurological activation patterns Haugeland described.

The information that a door can be opened by kicking it in is useless for a robot on wheels. For anyone but their own system, it is as if mental representations were encrypted in observable but *incomprehensible* structures that deny external access to their content. As a consequence, the content in these structures might be *observed,* but it cannot be *understood,* by agents with another body.

Bodily particularities of the cognitive system are indispensable in the generation of a mental model. Parts of the environment mean different things to different interpreters. With growing experience, mental models are adapted by, and adopted to, the body of their particular system in its particular environment.

Bodies enable to evaluate the relevance of a referent in virtue of its contribution to maintaining or destroying them. This provides a solution to the frame problem. The mind alone cannot determine existential relevance. To do that requires a body.

9. Conclusion

If content is not an explicit structural component of the mental representations in a cognitive model, then it must necessarily reside within an observer. To avoid filling this 'semantic vacuum' with external content, structural mental content must be part of the cognitive model.

To solve the symbol grounding problem requires this content to be independent from an observer and, therefore, to be created autonomously by the observed system. The chapters in this part show how embodied cognition explains a generation of content that satisfies this condition.

The supposedly anti-representationalist stance of embodied cognition turns out to object only a rather formal conception of mental representation. In fact, the emphases on dynamic interaction and structure in content even suggest practical methods that enable the grounding of mental representations. This is not compatible with a categorical rejection of symbolic mental representation in general.

The suggested methods are *dynamic systems theory* and *connectionism.* Dynamic systems theory, on one hand, as a means to describe the reactive interaction between a system and its environment. Connectionism, on the other hand, to explain the generation of structures that describe a referent and which serve as content for symbolic representations. A combination of both enables to describe how the content of a mental representation can be created and adapted in a dynamic interaction with its system's environment.

Both methods are frequently employed in hybrid computational models of cognition. In contrast to Harnad's initial proposal of a hybrid model, however, more recent approaches do not require immediate external supervision to generate content. Instead, they ground in lawful dependencies between the motor capabilities of the system and the according sensor reaction from the environment.

Common to these models is an emphasis on the procedural character of basic perception. Where, until recently, the perceptual states of cognitive systems were rather thought of as grounded in momentary 'snapshots' of sensor activation, now there appears to be a lot of consent among representatives of embodied cognition that basic perception is in fact not singular and atomic, but temporally extended and composed in ways much similar to abstract concepts.

This draws attention to the cognitive processes that compose subjective phenomena into abstract concepts. This is researched by *phenomenology* and, more general, the philosophy of mind. O'Regan and Noë's approach in particular shows that mental representation can be phenomenologically grounded in the unconscious mind.

In the following part we examine the phenomenological aspect of mental representations in more detail. This enables the development of a model for the generation of mental representations grounded in unconscious processes that are analogous to, but different from, what is usually conceived of as cognitive processes.

Part III.

Intentional Mental Models

10. Introduction to Intentional Mental Models

The models in the previous part describe cognitive systems as a whole. The part ends with an argument for *the inclusion of the structural mental contents of cognitive systems in a cognitive model of the same systems.* The sum of all of its mental contents is the cognitive system's *mental model.* The part before that presents two different conceptions of this content and argues in general for *a subjective interpretation of mental representation.* This part now elaborates on the particular structures and processes that are involved in the processing of a subjective mental model.

Mental models facilitate conscious processes of cognitive systems like thinking, planning, or elaborating (May 1996, pp. 406–407). A mental model is *internal* to its system. A comprehensive model of cognitive systems, therefore, must be *a model of mental models* as well (Strube 1996b, p. 407).

Jay Forrester emphasises the importance of mental models for understanding reality. '[The cognitive system] has only selected concepts and relationships which he uses to represent the real system [...] The question is not to use or ignore models. The question is only a choice among alternative models' (Forrester 1971, p. 112). Cognitive systems have to deal with the potential gaps and flaws of their mental model because it is *their only way* to conceive of the world.

Forrester explains that a 'mental model is fuzzy. It is incomplete. It is imprecisely stated. Furthermore, within one individual, a mental model changes with time and even during the flow of a single conversation' (ibid., p. 112). Under these conditions, the *correctness* of mental models appears to be subordinate to their *practicality.*

The dynamicity of mental models becomes apparent with the shifts they undergo during the course of a conversation. 'The human mind assembles a few relationships to fit the context of a discussion. As the subject shifts so does the model' (ibid., p. 112). If the received information is flawed, then so can be the mental models of the receiving cognitive system.

Cognitive systems can agree upon the *expressions* of their mental model but they cannot access the underlying mental model of another system directly. 'When only a single topic is being discussed, each participant in a conversation employs a different mental model to interpret the subject. Fundamental assumptions differ but are never brought into the open.' (Forrester 1971, p. 112)

Accordingly, an individual cognitive system might very well be *the only system* to have a particular mental model. Even with an arbitrarily extended period of adaptation, it cannot be guaranteed that two interpreters ever arrive at the same mental model—let alone that one of these models is more 'correct' than the other.

10.1. The World as Mental Model

According to Forrester, '[e]ach of us uses models constantly. Every person [...] instinctively uses models for decision making. *The mental image of the world around you which you carry in your head is a model.*' (ibid., p. 112, emphasis added)

Its mental model enables a system to deal with external reality—to understand, for example, the possible consequences of its actions. To a cognitive system, the world appears in virtue of its own mental model of external reality.

Philip Johnson-Laird illustrates the connection between mental models and the world in a similar way. He states that "[t]he limits of our models are the limits of our world." (Johnson-Laird 1991, p. 471)

Naively put, however, models are obviously only *part* of the world and the world is obviously *not* a model itself. So what are the conditions under which mental models and the world coincide like Forrester and Johnson-Laird suggest?

The difference between a mental model and the world lies in the conception of mental representation. Mental models consist of mental representations. The applied conception of mental representation therefore has severe influence on one's conception of mental models. To see the implications, consider the following.

On the one hand, there is *the observer's* perspective on the mental representations of a cognitive system. On the other hand, there is *the system's* perspective on its own mental representations. These perspectives coincide with subjective and objective representations as we present them in chapter 3. From the system's 'own' point of view, mental representations *are* the objects that the world is composed of.

This particular characteristic of subjective mental representation is often referred to as 'intentionality'. The concept experienced a surge in interest since Searle's formulation of the Chinese room argument and its impact on artificial intelligence and the cognitive sciences.

Unfortunately, there is a diversity of opinions on the epistemic and ontological status of intentional states, on their constituents, their defining characteristics, their types, and respective consequences. In the following, a short overview is provided over its development and contemporary conceptions.

10.2. Intentionality in Mental Representation

Fodor (1983) and Searle (1983) differentiate the 'original' or 'intrinsic' intentionality of some mental states from the 'derived' intentionality of linguistic statements. Dennett (1989) and Cole (2010), in contrast, argue for intentional monism.

On the one hand, Cole suggests that intentionality is *never* derived while Dennett, on the other hand, defends the position that *all* intentionality is derived.

Over the years, Searle produced one of the most extensive and consistent bodies of work on this subject. Accordingly, Harnad adopts some of Searle's ideas on intentionality in his computational approach to the symbol grounding problem.

However, Searle rejects the idea that digital computers may process original intentionality *in principle.* He states that '[s]uch intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output.' (Searle 1980b, p. 422)

Searle's work allows to infer why he sees computational systems as intrinsically incapable of intentionality and what would be necessary to change this. The purpose of this part is to provide a overview over his reasoning which eventually culminated in the Chinese room argument.

As Margaret Boden notes in reference to Smith (1986), however, 'there is no general agreement on what *intentionality* is, and there are deep unclarities about *representation* as well.' (Boden 1988, p. 248)

Most authors agree that 'intentionality' is what makes something be 'about' something else. At this point, however, the consensus ends.[1]

Searle assumes that moods and emotions are mental states *without* intentionality. To Crane, these states *are* intentional because they describe the world in a particular way (e.g. beautiful or depressing). In all mental states, 'there is the experiencing subject, the world experienced (or the thing in the world experienced) and the particular way of apprehending the world.' (Crane 1998, p. 245)

[1] Recently, even this property of intentionality has been attacked by Tegtmeier (2005) and Drummond (2012).

This chapter proceeds to explain the roots of a modern day understanding of intentionality, starting from Franz Brentano's ideas about the *intentional inexistence* of objects. Next, the modal logic concept of *inten**s**ionality* is described to avoid the common conflation with *inten**t**ionality.* Lastly, Searle's conception of intentionality is presented in detail.

10.3. Inexistent Objects

Intentionality has been a philosophical topic for centuries. Brentano re-introduced it as an object of psychological investigation. His intention was to understand how the subjective experience of cognitive systems can present external reality in the form of objects that do not exist in external reality themselves.

Mental phenomena involve all kinds of subjective experience. This includes, for example, emotions, moods, beliefs, memories, intentions, or perceptions. According to Brentano, all these phenomena have an intentional object.

'Every mental phenomenon includes something as object within itself, although they do not do so in the same way [. . .] We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves.' (Brentano 2012, pp. 88–89)

Following from this, every mental phenomenon *contains* its own object. Intending, for example, is always to intend *something*, believing is always to believe *in something,* and hoping is always to hope *for something.* The intentional object is what is intended, believed in, or hoped for.

To call the content of a mental phenomenon 'object', however, can lead to confusion concerning its ontological status. Brentano anticipated this confusion. Therefore, he emphasised that intentional objects do not exist in the same way we deem physical objects to exist outside of our minds. They are, rather, *immanent to the act of thinking itself.*

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a *thing*), or immanent objectivity. (translated[2] ibid., p. 88, emphasis added)

[2] „Jedes psychische Phänomen ist durch das charakterisiert, was die Scholastiker des Mittelalters die intentionale (auch wohl mentale) Inexistenz eines Gegenstandes genannt haben, und was wir, obwohl mit nicht ganz unzweideutigen Ausdrücken, die Beziehung auf einen Inhalt, die Richtung auf ein

It is not Brentano's intention to make clear that intentional objects do not *mean* anything but rather that they do not refer to something in an *external reality.* This has been put forward in detail by Crane (2006) and becomes clear if the English translation is compared to its German original.

10.4. External Reality

This understanding of 'inexistent object' raises the immediate question why anyone would want to talk about objects *that do not really exist.* The answer is quite straightforward: to express the belief that objects *in general* do not exist independent from, or outside of, the mind. Crane makes clear

> that Brentano's view is not that there is a distinction between "physical objects" which exist, and "intentional objects" which do not exist. His view is rather that none of the objects which are studied by science "really and truly exist": they are phenomena, mere appearances, which are signs of an underlying reality but which are not real themselves. (ibid., p. 23)

The objects we perceive are a product of how our minds process information about external reality. They are not a part of objective reality themselves but depend as much on the perceiving subject as they do on the information that it receives and processes without awareness.

> We can say that there exists something which, under certain conditions, causes this or that sensation. We can probably also prove that there must be relations among these realities similar to those which are manifested by spatial phenomena of shapes and sizes. But this is as far as we can go. (Brentano 2012, p. 19)

Despite his intention to emancipate psychology from the natural sciences, Brentano very well believed in an objective reality beyond subjective experience. This reality, however, is not composed of the objects of perception.

'The phenomena of light, sound, heat, spatial location and locomotion which [the natural scientist] studies are not things which really and truly exist. They are signs of something real, which, through its causal activity, produces presentations of them.' (ibid., p. 19)

Objekt (worunter hier nicht eine *Realität* zu verstehen ist), oder die immanente Gegenständlichkeit nennen würden.“ (Brentano 1874, pp. 124-125, emphasis added)

By defining the objects of perception to be primarily *psychological,* Brentano determines that phenomenological processes are prior to everything that the natural sciences can describe.[3]

10.5. Emancipating the Mental

In his original formulation, Brentano meant intentionality to be a necessary and sufficient indicator for a mind. Brentano made this clear in his famous intentionality thesis. "This intentional inexistence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it." (Brentano 2012, p. 89)

These are two premises in a syllogism to prove that mental phenomena are not physical phenomena. The first premise states that intentionality is a sufficient (i.e. 'characteristic') and necessary (i.e. 'exclusive') mark for mental phenomena. The second premise states that there is no physical phenomenon that features intentionality. Therefore, there is no physical phenomenon that is also a mental phenomenon.[4]

Both premises draw a clear demarcation line between the first-person-perspective from which mental phenomena are experienced and the third-person-perspective from which this can be observed (e.g. as a physical phenomenon). The difference between both is unmistakeably indicated by the presence or absence of intentionality.

Natural science is limited to observations from outside the observed system. A phenomenological approach, on the other hand, gives first-person-perspective the ultimate authority over statements about the mind. This emancipates psychology from the natural sciences.

In Brentano's view, the central function of the mind is to map immediately imperceivable external things onto consciously accessible phenomena. Broadly speaking, intentionality is what makes mental representations present these things as the *objects of perception.*

The question on how this might be realised spawned several different approaches. Embodied cognition suggests that the particular body is somehow functionally responsible. According to Searle, on the other hand, this is due to the specific biochemical composition of our brains. The next chapters on his understanding of intentionality show that such

[3] Presumably under the pressure of some of his students, the late Brentano considered it necessary to revise his view in the second edition of *Psychologie vom empirischen Standpunkte.* In his last years, he regarded the objects of thought as real (i.e. physical) things and described intentionality much according to the Anglo-Saxon tradition. Unfortunately, this broke consistency with his earlier work and is mostly considered to be motivated socially and not scientifically.

[4] Most analyses have segmented Brentano's argument along the necessity/sufficiency of intentionality as a mark of the mental (e.g. by Crane 1998; Nes 2008; Schlicht 2008). Relevant in our case, however, is his intention to separate the mental from the physical domain.

different theses are not fundamentally incompatible.

11. Two Traditions

In the late 19$^{\text{th}}$ and early 20$^{\text{th}}$ century, intentionality was picked up by two quite different philosophical branches. According to the topics traditionally discussed by these disciplines, different aspects of intentionality have been emphasised.

The continental tradition is close to Brentano's original interest in subjective experience. It explicitly concerns the relation between subjective *mental phenomena* and their capacity to be about something.

The Anglo-Saxon tradition, on the other hand, is more interested in the linguistic implications of intentionality. From its representatives, intentionality is often presented as what makes *linguistic statements* be about something else.

In contemporary philosophy, the difference between both conceptions of intentionality is usually reflected terminologically. Brentano's interpretation is referred to as 'inten***t***ionality'.

Unfortunately, however, its linguistic counterpart is not very distinguishably referred to as 'inten***s***ionality'. The term is inherited from Frege's conception of sense as *intensional* meaning.

In the following, the difference between inten***t***ionality and inten***s***ionality is illustrated with particular emphasis on the implications on a computational simulation of intentional mental representations. Hopefully, this allows to avoid any pitfalls like mixing up objective and subjective conceptions of mental representation due to a misinterpretation of examples or thought experiments like it was the case with the Chinese room argument.

11.1. A Linguistic Conception of Intentionality

The fact that representations can be about things that do not really exist (e.g. unicorns) has baffled analytic philosophy for a long time. Willard van Orman Quine described the problem as follows. "Nonbeing must in some sense be, otherwise what is it that there is not? This tangled doctrine might be nicknamed Plato's beard [...]" (van Orman Quine 1980, p. 2).

To avoid complicating the question how one thing can be about another thing by the need to *also* consider the existence of the other thing, van Orman Quine proposed a 'semantic ascent'. His intention was to contain the question for representation by only considering its expression.

> The strategy of semantic ascent is that it carries the discussion into a domain where both parties are better agreed on the objects (viz., words) and on the main terms connecting them. Words, or their inscriptions, unlike points, miles, classes and the rest, are tangible objects of the size so popular in the marketplace, where men of unlike conceptual schemes communicate at their best. (van Orman Quine 1960, p. 272)

This enables to consider intentionality from a restricted linguistic point of view. According to this understanding, intentional mental states can be analysed in virtue of their linguistic expressions like 'I believe Santa Clause lives at the north pole' or 'I know Barack Obama is the president of the USA' instead of my *actual* belief or knowledge.

Intentionality that can be *expressed,* however, is quite different from Brentano's original conception. Expressions lack the property that is most important to phenomenology: *a feeling.*

Feelings are exclusive to first-person-perspective, expressions are not. Although feelings might retain some of their properties when being expressed, intentionality in Brentano's sense is lost.

A conception of intentionality that can be expressed implies a radical change from phenomenology to linguistics. This is explicitly intended by van Orman Quine. The result is, however, that 'Quine's attributions [to intentionality] bear little relation to what Brentano really said.' (Crane 2006, p. 33)

Searle (1979), Crane (1995, pp. 32–36), and Rapaport (2012) also emphasise the differences between both conceptions. According to them, it is crucial to keep both separate and they give very similar accounts on how to do this.

Searle's original motivation is indeed linguistic. He applies his speech act theory to the phenomenological domain but he describes intentionality as *an irreducibly private aspect of subjective experience.*[1] His premise is that intentionality is necessary for, and therefore *prior to,* any sincere speech act. (Searle 1983)

[1] Effectively, Searle disagreed with both: the continental tradition, represented by Jaques Derrida, and the Anglo-Saxon tradition, represented by van Orman Quine. The argument with Derrida took place quite aggressively (for an overview, see Fish 1982; Wright 1982; Kenaan 2002; Raffel 2011).

11.2. Intensional Statements as Intentional States

According to Frege, what is commonly referred to as 'meaning' consists of two parts: an *extension* (i.e. the reference) and an *intension* (i.e. the sense). He differentiated both to resolve 'Plato's Beard': someone can have a certain meaning 'in their head' without a real world correlate. (Frege 1892)

In Frege's terminology, the planet Venus is, at the same time, the extension of 'morning star' *and* 'evening star'. The intensions of these names, in contrast, consist of the conditions under which its references can be established.

The intension of 'morning star' is therefore *that it is the last star to be seen in the morning.* The intension of 'evening star' is *that it is the first star to be seen in the evening.*

Two similarities suggest a close relation between intentional mental states and intensional statements.

The first one is that intentional states and intensional statements can both be 'about' something *that does not really exist.* For example, I can believe that unicorns hunt in packs and I can express this belief as well. Without an actual pack of unicorns, the statement cannot be compared against the real world. Such a comparison, however, is necessary to determine a truth value.

The second similarity is that both can be 'about' *the same thing in different ways.* At the same time I can know Barack Obama but not know any former president of the USA.

If I know Barack Obama but I do not know that he was president, then I can believe that Barack Obama is in the room and belief that no former president is in the room at the same time.

These two different intentional states can be expressed with two according intensional statements as well. Both intensional states can be the case at once. Both statements, however, *cannot* be the case at once.

The truth values of 'Barak Obama is in this room' and 'a former president is in this room' must be identical exactly because Barack Obama *is* a former president. Equal terms (e.g. 'Barack Obama' and 'a former president') must be substitutable under preservation of the truth value of the whole statement.

Crane (1995) presents two similar cases to show that there is no necessary relation between intentional mental states and intensional statements. Searle (1979) and Rapaport (2012) even conclude that there is no relation at all.

In fact, on an intricate connection between the intensionality of statements and the intentionality of mental states, Searle writes that '[n]othing could be further from the truth' (Searle 1979, p. 85) and that it is '[o]ne of the most pervasive confusions in contemporary philosophy.' (Searle 1979, p. 85)

This is directly supported by Brentano.

> The truth of physical phenomena is, as they say, only a relative truth. The phenomena of inner perception are a different matter. They are true in themselves. As they appear to be, so they are in reality, a fast which is attested to by the evidence with which they are perceived. (Brentano 2012, p. 19)

In contrast to any linguistic conception of intensionality, Brentano was not bothered by *the truth* of intentional mental states because his conception of intentionality provides the necessary ground for calling anything 'true' in the first place.

12. John Searle on Intentionality

Searle describes intentionality as “by definition that feature of certain mental states by which they are directed at or about objects and states of affairs in the world.” (Searle 1980b, p. 424)

He argues that “[i]nstantiating a computer program is never by itself a sufficient condition of intentionality” (ibid., p. 417). He concludes that no computer program can ever exhibit mental states to the same extent as a real cognitive system (e.g. a human).

In the following, this argument is analysed with regard for his conception of intentionality. To Searle, intentionality is essentially *biological.* Therefore, it is crucial to understand what is supposed to be so peculiar about intentionality *as a natural entity* that computer programmes should not be able to realise it.

Searle does not argue against artificial intentionality in general. The logical properties that realise intentionality might be realised by various other materials. “It doesn’t matter how an Intentional state is realized, as long as the realization is a realization of its Intentionality.” (Searle 1979, p. 81)

We know, for example, that biochemicals are among the ‘proper’ materials because we experience intentionality with our own biochemical brains. In Searle’s view, digital computers are not among these materials because they can only realise *syntactic* processes and syntax does not feature the relevant logical properties. Computers may simulate these properties but to Searle, a simulation of intentionality is not intentionality at all.

> Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis (Searle 1980b, p. 424)

This is a first overview of Searle’s take on the conditions of intentionality. His argument, why symbol processing cannot cause intentionality, goes into much more detail.

12.1. Intentionality and Computational Symbols

According to Searle, digital computer can process representations only in virtue of their shapes. To understand, in Searle's view, is to have a mental state that *contains* the meaning of its referents. He criticises strong artificial intelligence for ignoring the fact that everything computer programmes can ever do, is manipulate shapes whereas only minds manipulate content.

According to Searle, computer programmes always perform formal symbol manipulation and "all that 'formal' means here is that I can identify the symbols entirely by their shapes" (Searle 1980b, p. 418). Formal processes attach only to the shapes of symbols, not to their content.

Mental states, in contrast, can have intentional content. This content enables them to be about something else. Without content, shapes can only present themselves.

Searle describes his view on the relation between formal processes and intentional states in more detail:

> the program is purely formal, but the intentional states are not in that way formal. They are defined in terms of their content, not their form. The belief that it is raining, for example, is not defined as a certain formal shape, but as a certain mental content with conditions of satisfaction, a direction of fit [. . .], and the like. (ibid., p. 423)

Intentionality is an *intrinsic* feature of mental states. Unlike linguistic content which is determined by a community, the content of intentional states is internal to an individual.

Like mental states, statements are about something. Their intentionality, however, is only derived from the original intentionality in the mind of their speaker. Without original intentionality in mental states, linguistic expressions cannot be about anything.

12.2. Intentional States and Speech Acts

Searle develops his idea of intentionality as a basis for his speech act theory. According to him, speech acts enable access to mental states because the intentionality in expressions is *derived* from the original intentionality in mental states—not because intentionality *in itself* is somehow linguistic. (Searle 1983)

Speech acts according to Searle consist of two components: *propositional content* and *illocutionary force.* The former describes some referent and the latter specifies the type of relation that is established with this referent. Force and content are not just distinct from one another, they are members of two essentially different categories.

Content alone cannot exert influence and, hence, is not an act in and by itself. It can merely describe something but a pure description is not instructing anything because it is descriptive, not normative. Only a particular stance towards this description can initiate change. This stance is a force.

For example the propositional content of the expression ‘pick up that block’ is *that block being picked up* and the illocutionary force is most probably a command. The propositional content of ‘I will give you your money back’ is *that the money be returned* and the illocutionary force is a promise.

The propositional content of an expression does not need to be a physical object. It can also describe processes, abstract entities, or situations. Also, various types of illocutionary forces are possible, as for example accusation, prediction, or explanation.

A speech act succeeds if propositional content and state of the world are in accordance; it fails if they are not. The propositional content, therefore, defines *conditions of satisfaction* for the success of the speech act by describing some state of the world. The illocutionary force is the speaker’s disposition towards this state.

In the example ‘pick up that block’, the conditions of satisfaction are met if the world conforms to the propositional content. This state is established by someone acting according to the command. If the block is being picked up by someone, the speech act was successful. The speech act ‘I will give you your money back’ is satisfied as soon as I give you the money. (Searle 1969; Searle 1979)

Commands and promises have a *world-to-word direction of fit* because they can cause a change in the world according to words. Explanations, on the other hand, have a *word-to-world direction of fit* because their words are determined by aspects of the world. Excuses are an example for a *null case* without a particular direction.

12.3. Intentional Content and Psychological Mode

According to Searle (1983), a speech act can shed light on intentionality if it is *the sincere expression of a mental state.* The prediction ‘tomorrow it will rain’ can only provide information on the mental state of the speaker if they truly believe *that it will rain tomorrow.*

For accordance between speech act and mental state, Searle infers that both must be composed similarly. The propositional content in a speech act is analogue to *the intentional content* in a mental state and the illocutionary force in a speech act is analogue to *the psychological mode* in a mental state.

Like propositional content provides the conditions of satisfaction for a speech act, intentional content provides the conditions of satisfaction for a mental state. In virtue of these conditions, mental states can be 'about' or 'directed towards' certain states of the world.[1]

The direction of fit in mental states is either *mind-to-world, world-to-mind,* or the *null case.* Beliefs are in a mind-to-world direction of fit, desires are in a world-to-mind direction of fit, and feeling sorry is an example for a null case.

Propositional content is derived from intentional content because speech acts are an *externalisation* of the mental states we perceive ourselves to be in. However, language is just one of the systems onto which we project our introspection.

"The mind imposes Intentionality on entities that are not intrinsically Intentional by intentionally transferring the conditions of satisfaction of the expressed psychological state to the external physical entity." (Searle 1979, p. 89)

We conceive of other systems as if they had original intentionality. It might seem, for example, as if the car does not *want* to start or that the room *understands* Chinese.

The conditions of satisfaction for those impressions, however, are not determined by the concerned system itself but part of *our own* mind. The apparent intentionality in expressions is *always* derived from one's own original intentionality. This applies to language in the same way that it applies to computational, or any other, system.

12.4. Mental States Appear as Aspectual Shapes

To Brentano, the defining property of mental states is their intentional content. Searle, in contrast, concedes that states without an intentional object can be mental as well.

He argues that the defining property for mental states is not intentionality but *potential consciousness.* This enables to describe 'undirected forms of anxiety and depression' as mental states although they do not seem to contain anything. The potential for a mental state to become conscious is provided by its *aspectual shape.* (Searle 1980b)

> Aspectual shape is most obvious in the case of conscious perceptions: think of seeing a car, for example. When you see a car, it is not simply a matter of an object being registered by your perceptual apparatus; rather, you actually have a conscious experience of the object from a certain point of view and

[1] Notice, as Élisabeth Pacherie puts it quite nicely, that "the expression 'conditions of satisfaction' makes reference to the requirements that have to be met, and not to the things that meet those requirements." (Pacherie 2000, p. 405)

Searle's intentional content is therefore better understood as a set of sufficient conditions (i.e. as Fregean sense), rather than the reference to a particular thing that satisfies these conditions (i.e. as Fregean reference).

> with certain features. You see the car as having a certain shape, as having a certain color, etc. And what is true of conscious perceptions is true of intentional states generally. (Searle 1992, p. 157)

Aspectual shapes determine how mental states appear to their system. All mental states have an aspectual shape because Searle does not consider states that do not appear to their system as mental.

Aspects are always from a perspective and, therefore, subject to various relations between the system and its environment. An outside observer cannot perceive the same aspects as the observed system—or simply put: *aspects are subjective.*

"The aspectual feature cannot be exhaustively or completely characterized solely in terms of third-person, behavioral, or even neurophysiological, predicates. None of these is sufficient to give an exhaustive account of the way it seems to the agent." (Searle 1991, p. 53)

Intentional content determines the *object* of a mental state and aspectual shapes determine its *appearance.* Searle's conception applies to all, and only to, mental states that describe an inexistent object.[2]

12.5. The Intentional Network

The previous section concerns the relation between shape and content. Different mental contents, however, are also related to one another to describe connections between the external referents that they present to their cognitive system.

Searle presents the following example. Oedipus wants to marry Queen Jocaste, of whom he does not know that she is his mother. Oedipus' mental state of wanting to marry Jocaste thus describes a world in which he is married to his mother.

Unfortunately, this intentional content is presented to him in virtue of an aspectual shape. The content of which is only accessible from Oedipus' own perspective. Because he does not know that she is his mother, the content also cannot not include the information that she is. (Searle 1983, pp. 101–107)

But surely, Oedipus does not want to marry his mother. Where is the according intentional content for this? It does not make sense to assume that this is also contained in his wish to marry the queen—something along the lines of '. . . *and* that the queen is not my mother'. If it was, the intentional content would also need to contain that she is

[2] In contrast to Brentano, Searle assumes that moods or emotions do not have an object and are therefore not intentional but only phenomenal.

currently alive, that she is not already married, that the queen is not made of stone, and so on.

Oedipus does not need to think explicitly about all these constraints, but naturally, they seem to play a role in the conditions of satisfaction for his wish. If all the conditions would need to be explicit, in the end, the possibly relevant information for any condition would need to contain everything he assumes to be the case and everything he assumes *not* to be the case.

Consider, for example, that Oedipus' mental state (i.e. wishing that he is married to the queen) implicitly requires the form of government to be a monarchy. It seems awkward to consider the state of government as part of Oedipus' wish. If intentional content would need to be *exhaustive,* Oedipus would face a problem much similar to the frame problem we present in section 2.1.

To solve this problem, Searle introduces *a network of intentional content.* Oedipus' intentional content that describes him being married to Jocaste is interconnected with all the other content that is relevant for the goal state—much like the approaches to the frame problem in section 2.1.1.

Satisfying the conditions for one of his mental states depends not only on the content of *this* state but also on the content of *various other* states that have been experienced by him to be in some way relevant for it. Past mental states can be represented by their aspectual shapes in the content of the present state.

Therefore, Oedipus sees in the wish to marry the queen only *his* aspects of the world from *his* perspective on the world. His mental state implicitly depends on *his understanding* of marriage, queens in general, and the specific queen he wants to be married to.

These relations form a network of intentional content. If there were no further component in Searle's description of the mind, it would lead right back to the symbol grounding problem. If intentional content consists only of aspectual shapes that reference other intentional content, how can it be grounded in something that is not content already?

12.6. Grounded in the Background

Without a word on *the foundation of intentional content,* however, Searle's theory on intentionality faces the symbol grounding problem all over again. Without an explanation that describes the pre-conceptual origins of intentional content, every cognitive model that covers the mental model of the described system runs danger to implant its designers *own* conceptions into to the system that it is supposed to described.

Searle's intentional content is grounded in what he calls 'a background of physical skills'. This skills are know-how about, for example, how to open a door, how to drink from a bottle, how to walk, how to swim, and so on.

> Intentional states only have the conditions of satisfaction they do, and thus are the states that they are, against a Background of abilities that are not themselves Intentional states. In order that I can now have the Intentional states that I do I must have certain kinds of know-how (Searle 1983, p. 143).

The skills in the background are deep or local.

The deep background contains skills that are

> common to all normal human beings in virtue of their biological makeup—capacities such as walking, eating, grasping, perceiving, recognizing, and the preintentional stance that takes account of the solidity of things, and the independent existence of objects and other people (ibid., pp. 143–144).

Deep skills might be predisposed, for example, in virtue of neural anatomy.

The local background, in contrast, contains cultural skills, such as opening doors or drinking from bottles. Local skills are all acquired. Searle describes that the internalisation of local skills is not like moving explicit rules for, say, a foreign language into the unconscious domain of our minds. Instead, explicit rules are like training wheels that enable to practice something we were not able to practice before.

Once it is possible to carry something out on our own in virtue of mindless exercise of rules, we are able to acquire expertise. "The rules do not become 'wired in' as unconscious Intentional contents, but the repeated experiences create physical capacities, presumably realized as neural pathways, that make the rules simply irrelevant." (ibid., p. 150)

Searle presents an example for when this background shows.

> Suppose as I go into my office, I suddenly discover a huge chasm on the other side of the door. My efforts to enter my office would certainly be frustrated and that is a failure to achieve the conditions of satisfaction on a Intentional state. But the reason for the failure has to do with a breakdown in my Background presuppositions. (ibid., p. 155)

To Searle, all mental states are embedded into a background of implicit and exercised habits, which are determined by the relations of an agent with its environment. This unconscious background serves as the ground for all intentional content.

A failure in this background is experienced as breakdown: implicit assumptions turn out not to be the case and this contradiction manifests, for example, as chasm in front of Searle's door or as the fact that Jocaste is indeed Oedipus' mother.

13. A Semiotic Framework for Subjective Mental Representation

In the previous chapters, mental models are described informally as being composed of mental representations with intentional content. In this chapter, a description of subjective mental representations is developed on the basis of Charles Peirce' semiotics.

The empirical foundation for his theory is obtained similar to Searle's and the phenomenological variants of embodied cognition: by enquiring one's own mind. Peirce called his method 'phaneroscopy'.

"Phaneroscopy is the description of the *phaneron;* and by the *phaneron* I mean the collective total of all that is in any way or in any sense present to the mind, quite regardless of whether it corresponds to any real thing or not." (CP 1.284)

The semantic characteristics of three general types of mental phenomena and their general role in the mind are formalised as relations. These three types are *feelings, facts,* and *thoughts.* Thought is the only *symbolic* mental representation and based on the other types.

Just like Searle, Peirce also describes the content of a symbolic representation as *structural.* Unfortunately, Searle's explanation for intentional content requires a biological correlate.

In Peircean semiotics, intentional content is not primarily natural but first and foremost *phenomenal.* Mental representation grounds in basic perception and basic perception is essentially inaccessible to natural sciences. Semiotics does not require a natural justification for subjective experience but merely accepts it as given.

13.1. Peirce and Symbol Grounding

According to Peirce, all thought is in signs (CP 5.253). This conveys the impression as if Peirce thought of mental phenomena as a subset of signs. But to him, signs *are* mental phenomena—only some of which are symbolic.

Peirce is frequently referred to in explanations for the symbol grounding problem (more recently, e.g. Sun 2000; Vogt 2001; Cangelosi et al. 2002; Vogt 2002; Gomes et al. 2005;

Clowes 2007; Vogt and Divina 2007; Steels 2008). So far, however, his theory of signs has not been connected to the intentionality of mental representations.

He was not familiar with Brentano's conception of intentionality as it is shared today, for example, by Searle. "Peirce never referred to Brentano and never used 'intentional' or its cognates in Brentano's sense" (Short 2007, p. 6). The analyses from Short (1981) and Short (2007, especially pp. 6–11) indicate, however, that both conceptions of the mind are phenomenological at the core and essentially compatible.

Peirce remains materially agnostic, like Searle with his theory on intentionality. To both, cognition can be described only in virtue of its function.[1]

An essential aspect of this function is *the representation of reality.* From a semiotic perspective, representation is signification and signification is performed by signs. According to Peirce, signs are realised by three basic semiotic components: *shape, content,* and *referent.*

In the remainder of this chapter, these components and their relations are described in more detail. This enables to describe *the structure in signs* as well as *signs in a structure.* Each sign appears in a context which, in turn, provides the content for a more abstract sign. Therefore, semiotics provides the formal basis for a computational implementation of Dreyfus' hierarchy of contexts (see section 2.2.1).

13.2. The Three Phenomenal Categories

Peirce categorised subjective experience into *Firstness*, *Secondness*, and *Thirdness.* "I analyze experience, which is the cognitive resultant of our past lives, and find in it three elements. I call them Categories" (CP 2.84).[2]

Peirce described the three categories of experience as follows.

> It seems, then, that the true categories of consciousness are: first, *feeling,* the consciousness which can be included with an instant of time, passive consciousness of quality, without recognition or analysis; second, consciousness of an interruption into the field of consciousness, sense of resistance, of an external *fact,* of another something; third, synthetic consciousness, binding time together, sense of learning, *thought.* (CP 1.377, emphases added)

[1] To Searle, this function requires a biochemical material whereas Peirce remains indifferent on this account.

[2] Originally, Peirce called these categories 'cenopythagorean categories' (CP 1.351; CP 8.328), influenced by the Pythagorean view of nature being ordered by natural numbers (Sörensen et al. 2012). This investigation, however, concerns mostly their phenomenological properties.

The present excerpt of Peirce' work is therefore by no means complete. The presentation is focussed on the relevant components of mental representation and their relations.

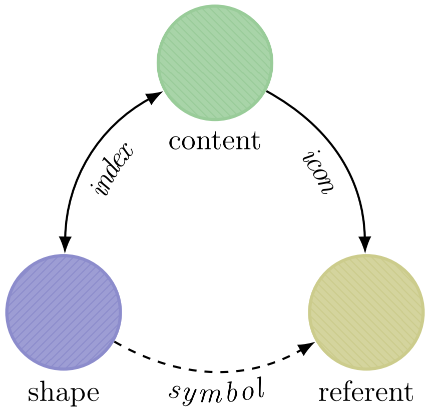


Figure 13.1.: The Directions of Fit in a Mental Representation.

All mental phenomena consist of three basic components: *shape, content,* and *referent.* These components can establish three general relations with one another: *feeling, fact,* and *thought.* A feeling is the appearance of content as a shape, a fact is content about a referent, and a thought is a shape that represents a referent in virtue of content.

Fact and thought are in a mind-to-world direction of fit (see section 12.3). Content adapts to a real referent and shapes adapt to this referent in virtue of their content. Feelings are bidirectional. Content can only *present* its referent by *appearing* as a phenomenal shape. Also *vice versa:* Any appearance presents at least its own presence.[3] Figure 13.1 shows the direction of fit of these relations (see section 12.3).

In the following, these phenomena are described with one of Peirce' rather early taxonomies of only three semiotic types: indices, icons, and symbols. Later, he refined his theory to define hundreds of different shades between these types.

13.2.1. Shapes, Feelings, and Indices

The first category consists of *feelings.* Feelings define a relation between shape and content. In virtue of feelings, content can be perceived as phenomenal shape and phenomenal shapes can be interpreted as content.

From first-person-perspective, shapes are singular appearances. Take the shape ‘a’. To a system that perceives this shape as a feeling, it is an indivisible quality.[4] This quality has structure only when being observed by *another* system (e.g. ‘a’ has a bulgy bottom

[3] Towards the end of this chapter, an argument for the second direction is presented.

[4] Reading about indivisible qualities, the modern day understanding of ‘qualia’ comes to mind. Crane mentions, however, that, although Peirce introduced the term ‘qualia’, ‘he was talking about what experience is like, in a general sense, not restricted to the qualia of experience in the sense in which it is normally meant today.’ (Crane 2014, p. 71)

and a hook at the top). For the observing system, however, the shape is not a feeling at all.

Feelings establish the presence of a particular content. They relate shape and content and, therefore, they enable to interpret shapes as content and to perceive content as a shape. Feelings provide the content for the system's most basic perception.

The shapes of basic perception are interpreted by their system as the properties of objects (e.g. the redness of a flower or the smoothness of a stone). Without access to entities 'below' basic perception, the system has no alternative but to suppose that these basic perceptions are *an actual part of the world.* Once something appears red, the necessary content of this shape is *that there is red.*[5]

In other words, cognitive systems implicitly "suppose [the objects] have capacities in themselves which may or may not be already actualized, which may or may not ever be actualized, although [these systems] can know nothing of such possibilities [except] so far as they are actualized." (CP 1.25)

Feelings relate shape and content so immediately that their system perceives both as *the same.* But the shape of a feeling is only *an index* for its content: shapes indicate content that presents the presence of some thing.

13.2.2. Referents, Facts, and Icons

The second category consists of *facts.* Facts define a relation between referent and content. They provide the content necessary to recognise *that this referent is part of the world.*

Peirce explains that

> the second is precisely that which cannot be without the first. It meets us in such facts as another, relation, compulsion, effect, dependence, independence, negation, occurrence, reality, result. A thing cannot be other, negative, or independent, without a first to or of which it shall be other, negative, or independent. (CP 1.358)

A referent is a short segment in a temporal sequence of feelings. This sequence is often referred to as 'stream of consciousness' (psychologically introduced by James 1892; continued, for example, by Pope and Singer 1978; Raymond et al. 1992; Pope 2013; Potter et al. 2014).

Take ['a', 'c'] as a referent. To a system that recognises this referent as a fact, its individual shapes are *the case* and their sequence is *necessary.* Only observed by another

[5] The same applies to dreams or hallucinations. The characteristic property of such an episode is exactly that: the perceived appearances present themselves as if they were part of the world.

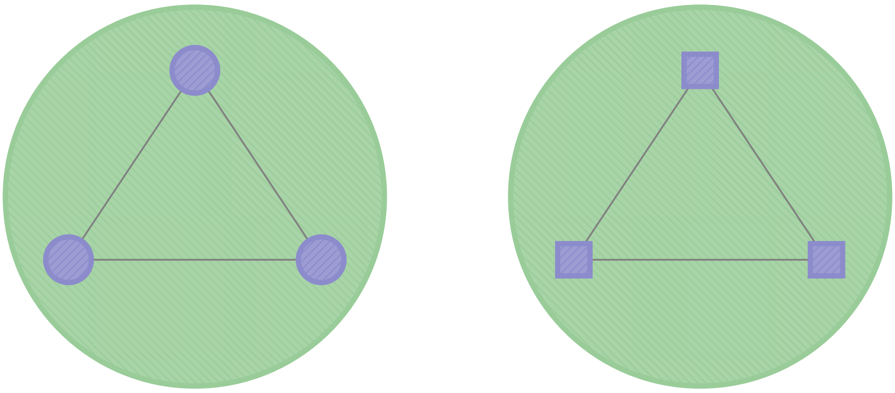


Figure 13.2.: Indices as Representational Elements.

system, this sequence has alternatives (e.g. then ‘a’ might as well be followed by ‘b’).

Referents can be recognised as facts that convey, for example, *succession* (i.e. ‘preceding’) or *unity* (i.e. ‘being with’). In particular, a fact can present *that ‘a’ precedes ‘c’* or *that ‘a’ is with ‘c’*. In general, facts make their system understand that circumstances are the case.

Referents depend on feelings in the same way in which statements like ‘this apple is a fruit’ or ‘Poland is next to Germany’ require that their words have content. Facts implicitly presuppose their own necessity just like these statements implicitly presuppose their own truth.

A feeling conveys *that **it** is in the world.* The relation between content and shape is permanent: the presence of red is always indicated by the appearance of red. A fact, in contrast, conveys *that some **other** thing is in the world* (i.e. the referent). This relation is context-dependent.

To recognise the fact *that there is an old friend from school,* for example, depends strongly on the current situation. In your home town, the same person is recognised far more easily than, for example, during vacation in a foreign country.

Facts convey information about a referent. They are *an icon* for this referent. If facts provide structural information about the referent, then they must feature *representational elements* (for an argument, see van Gelder 1989, pp. 62, 225). The representational elements of icons are indices. Facts are icons with feelings as their indices.

The individual feelings are irrelevant for the fact they compose. See, for example, figure 13.2: both bottom structures are icons that present a triangle. The shapes of their individual representational elements are irrelevant.

13.2.3. Content, Thoughts, and Symbols

The third category consists of *thoughts.* Thoughts define a relation between shape and referent. This relation is what enables mental representation. A thought is the feeling of a fact that is *detached* from its referent: content that appears in absence of its reference (i.e. an *inexistent* object).

Peirce conceived of thoughts as central to cognition. He made this clear when he described the third 'in between'. "Category the Third is the Idea of that which is such as it is as being a Third, or *Medium,* between a Second and its First. That is to say, it is Representation as an element of the Phenomenon." (CP 5.66, emphasis added)

From first-person-perspective, content is a mesh of interwoven facts, possible transitions, and alternatives. Take $\Big\{$ ['a', 'b'], ['a', 'c'] $\Big\}$ as structural content. To the system that has this content, it determines *all that is conceivable.* Observed by another system, however, this structure does only cover what is conceivable to the observed system.

Structural content can be interpreted logically. The content above, for example, can be interpreted as 'a' $\rightarrow$ ('b' $\vee$ 'c') or 'a' $\rightarrow$ ('b' $\wedge$ 'c'). Content can also depict causal dependencies, simple co-occurrences, or anything really. For example, *that 'a' precedes 'b' causes that 'a' precedes 'c'* or *that 'a' is with 'b' is more often than that 'a' is with 'c'.*

The content of a thought is an accumulation of facts. The content of a combination of these facts is more *general* than the content of an individual fact. Feelings present *themselves,* facts present *a particular reference,* and thoughts present *a general concept.* Feelings appear as themselves, facts appear as a concrete reference, and thoughts appear as an abstract reference.

> We have here a first, a second, and a third. The first is a positive qualitative possibility, in itself nothing more. The second is an existent thing without any mode of being less than existence, but determined by that first. A third has a mode of being which consists in the Secondnesses that it determines, the mode of being of a law, or concept. (CP 1.536)

Laws and concepts organise facts into patterns of decisions, options, and possible alternatives that have been internalised or *learnt.* These structures define cognitive categories to classify various individual referents.

"[Thirdness] brings the information into the mind, or determines the idea and gives it body. It is informing thought, or *cognition.* But take away the psychological or accidental human element, and in this genuine Thirdness we see the operation of a sign." (CP 1.537, emphasis added)

Thoughts do not contain objective facts about the world but abstract concepts that are learnt under the influence of uncountable bodily and situational dispositions. They are *symbolic* mental representations because their shape is not determined by necessity (i.e. feelings) or similarity (i.e. facts) but by convention or habit. This versatility explains not only why symbols are used in language but also why Peirce, Harnad, and so many others assume that cognition works symbolically *in general.*

13.3. At the Ground

If basic perceptions already contain their own presence, then what does their content ground in that is not content *already?*

To answer this question, consider ' '. You understand *that* *is part of the world* without the need to follow any convention. The shape conveys to you its own presence although you do not know what it means.

In fact, it is hard to imagine a system that could *not* obtain this most fundamental form of content from nothing more than the appearance of a random shape.

In the same way, every phenomenal shape *already* conveys a feeling. Shapes alone enable to understand *that they are present*—even if no one shares this knowledge. This content might not be objectively true but to whom the shape appears, its truth is literally *self-evident.*

This phenomenal content of shapes is unavoidable. All perception always already contains what it appears as and this basic phenomenal content provides the foundation for intentional content in more abstract mental representations.

Kant supported this assumption and shared Peirce' assessment that, for cognitive systems, the appearance of feelings is not only necessary but sufficient to infer *that this shape is part of the world.*

> I am no more necessitated to draw inferences in respect of the reality of external objects than I am in regard to the reality of the objects of my inner sense (my thoughts), for in both cases they are nothing but representations, *the immediate perception (consciousness) of which is at the same time a sufficient proof of their reality.* (CPR A 371, emphasis added)

Without prior content in perception, how should feelings imply any presence in the first place? *All* phenomenal shapes have phenomenal content, this content is *always* self-evident, and their system is *the only* authority considering this content.

13.4. Below the Ground

From first-person-perspective, mental representations cannot be grounded in anything else but basic perception. To predetermine the shapes of basic perception, however, violates methodological solipsism because the phenomenal content of basic perception is determined by this shape. All abstract content in this system would be predetermined by proxy.

As a consequence, the phenomenal content in basic perception must be generated *'below' the system's level of awareness.* The phenomenal content in basic perception must be composed from from something that does not already have content to the system.

Many approaches to embodied cognition suggest the same. O'Regan and Noë, for example, argue that basic phenomenal content is acquired in virtue of unconscious sensorimotor contingencies (see section 8.3.1). Gibson describes the need to determine sensorimotor 'invariants' (see section 7.2) prior to conscious perception.

Harnad does not explicitly mention an unconscious ground to his iconic projections (see figure 2.1 on page 15). If they provide structural information, however, then they must also consist of representational elements. If icons are his elements of basic perception, therefore, then *their* elements cannot be perceivable themselves.

In his description of Firstness, Peirce also hinted at elements below awareness. He described that basic perception is only *apparently* immediate.

> The immediate present, could we seize it, would have no character but its Firstness. Not that I mean to say that *immediate* consciousness *(a pure fiction, by the way),* would be Firstness, but that the quality of what we are immediately conscious of, which is no fiction, is Firstness. (CP 1.343, emphases added)

According to the emphases, the mind must feature elements that reference external reality more directly and prior to the first phenomena that appear to the mind. If we adopt Peirce' terminology, then the elements below feelings constitute a category of *Zeroness.* (also adopted, for example, by Bense 1975)

The elements of Zeroness are *necessary* for phenomenal content and therefore, they cannot be conceived of by their system. Only from third-person-perspective, the elements of Zeroness can be observed at the ground of cognition.

14. Conclusion

Three points are most important to take away from the theoretical part of this work.

Firstly, public representations (e.g. words) imply private representation (i.e. mental representation) but private representations do not imply public representation. This is a consequence of indirect realism which states that mental representations are *the only* epistemological access to reality and, therefore, the only access to public representations as well.

Secondly, and following from the previous point, the distinction between private and public representation depends on the system making the distinction. Private representations differ from public representations insofar as they are only in relation with the system assessing their privacy. Public representations, in contrast, can be in relation with any system.

Thirdly, and following from the previous point, as mental representations are private, they are essentially inaccessible to anyone but the system of whose mental model they are a part. If intentional content is the mark of mental representation, then it is fundamentally impossible to prove its absence in another (e.g. computational) system.

Intentional content is not only hard to describe. Instead, from third-person-perspective, intentional content in another system simply *does not exist*—observing the system *from the inside* does not change this.

For intentionality, the phenomenal experience in first-person-perspective is much more relevant than observation. This is quite obvious when the generation of one's own intentional content—the only example available—is considered. There is no other source for intentional content than basic perception.

14.1. Summary

Searle assumes computational procedures to be incapable of producing intentionality. They might simulate intentionality, but simulated intentionality is not actual intentionality—just like a simulated tornado is not an actual tornado. According to him, tornadoes and intentionality cannot be *computed.* To realise them, both need to be *materialised.*

You will indeed be disappointed if you expect simulations to produce tornadoes. Accordingly, if you look for intentionality in *another* system (e.g. the Chinese room) then, of course, you will not find any.

This is not due to the system's material but due to the fact that intentional content can only be experienced by its own system. No observation can yield intentional content—be it in the form of computational syntax or causal biochemical interactions—just like an observation of the Eiffel tower is different from the Eiffel tower itself.

On the one hand, Searle's account on intentionality requires to take the first-person-perspective. This, in turn, requires to conceive of mental representation as *subjective* (see chapter 3).

On the other hand, however, Searle also assumes that intentionality in the Chinese room could be *observed*—despite the fact that intentionality has *never* been observed in *any* biological system except oneself.[1] Observations of human neuroanatomy simply cannot explain intentional content.

This contradiction between these two of Searle's premises can only be resolved by rejecting one of them. To reject the first premise means to search for a *linguistic* solution to the *soft* symbol grounding problem. To reject the second premise means to search for a *phenomenological* solution to the *hard* symbol grounding problem (see chapter 4).

14.2. Recapitulation and Proceeding

In a purely phenomenological model of cognition, basic perception is a phenomenal 'atom'. The phenomenal shapes of basic perception might be different in each cognitive system. Therefore, basic perception must be considered as a *symbolic* representation of external referents (i.e. established by habit)—not as an *index* (i.e. established by necessity) or an *icon* (i.e. established by similarity).

According to embodied cognition, however, basic perception is not atomic but itself is composed of imperceivable representational elements. A semiotic analysis of symbolic representations enables to model these elements and the relations between them.

Peirce' semiotics is phenomenological and therefore it is compatible to Searle's conception of intentionality. It provides the means necessary to formalise *the initial generation* (i.e. semiosis) of basic perception as a symbolic sign.

From the unobservability of phenomenal content it follows that a successful simulation of intentional content requires to assume phenomenal content *by stipulation*—just like

[1] In fact, this exception is not an exception al all because, in oneself, intentionality is not observed but *experienced.*

we accept that the basic perception of fellow human beings consists in unobservable subjective phenomena that have actual phenomenal content. In the following practical part, this premise provides the basis for a workable simulation for the generation of mental representations.

Symbolic signs are defined by the semiotic relations between phenomenal shape, intentional content, and external referent. These relations are established by procedures. In the following, these procedures are implemented to eventually simulate the generation of basic perception in a computational system. More complex representations are generated starting from these basic perceptions.

Part IV.

Formalising Mental Models

15. Introduction to Semiotic Models

So far, two different types of model play an important role. *Mental models in cognitive systems.* They enable their system to predict external reality. *Cognitive models of cognitive systems.* They enable cognitive scientist to predict an autonomous agent.

Three components are crucial in both cases: 1) *the model* itself, 2) *the system* that the model describes, 3) and *the observer* of the system—who is also *the interpreter* of the model—for whom the model serves as a description.

In the case of *mental models,* the observer is the cognitive system, the other system is external reality, actions are motor activations, and reactions are sensor activations. Motor activations are emitted by the cognitive system and sensor activations are emitted by external reality.

A mental model establishes relations between *subjective representations* that postulate to its cognitive system a sufficient causal influence between these emissions. These mental relations determine the system's *behaviour* because they provide a conceivable explanation for external reality as *a world of perceivable objects* in the first place.

In the case of *cognitive models,* the observer is a cognitive scientist, the other system is an autonomous agent, actions are experiments, and reactions are behavioural data. Experiments are emitted by the cognitive scientist and the behaviour is emitted by the autonomous agent.

A cognitive model establishes relations between *objective representations.* These relations postulate to the cognitive scientist a sufficient causal influence between experimental settings and the agent's behavioural reaction. These formal relations determine *an understanding* for the generation of mental representations because they provide a conceivable explanation for autonomous agents as *cognitive systems* in the first place.

A comprehensive cognitive model establishes its formal relations not only according to observed relations between experiments and behaviour but also *based in the relations in a mental model.* From the assumption that representations in mental models are essentially subjective, it follows that not only mental models, *but cognitive models as well,* must be firmly grounded in subjective experience.

The remainder of this part fleshes out a conception of mental models that serves as the heart of such a cognitive model. The next part provides an according formalisation and connects it to a formal conception of external reality. This connection formalises the *interface* between agent and environment. The following part provides a computational cognitive model that describes how the data from this interface is integrated into a mental model as this part presents it.

15.1. A Semiotic Conception of Models

Traditionally, models are considered to predict the aspects of another system independent from the interpreter. The following interpretation of 'prediction', however, enables another conception that will be referred to as 'semiotic model'. Two particular points make a semiotic conception of models intrinsically dependent on their interpreter.

The first point is that the 'prediction' that models enable does not refer to a moment in the future that is characterised by the predicted aspects *starting to be the case.* Instead, the future connotation of 'prediction' refers to a moment that is characterised by these aspects *starting to appear to an observer.*

The second point is that predictions are intrinsically dependent on the *actions* of this observer. This is not always obvious. Predictions of geographical features as they can be inferred from a map, for example, describe aspects that seem to be independent from the actions of its interpreter. However, they appear only under particular conditions.

Maps depict these conditions explicitly. They describe *which route to take* such that particular geographical aspects appear. In fact, the *act* of navigation is is fundamental to any map. Without the observer's ability to move, maps do not have a purpose and there would be no ground to call what the map does a 'prediction' of geographical features at all.

Like any model, maps enable an observer to predict *the reactions* of another system to *their own* actions. In the case of a street map, this system is a particular network of streets, its reaction may be the appearance of a junction, and the condition is taking a particular turn as depicted on the map.

The observer's actions need not be physical. Models can also predict a system's reaction to a change in the observer's *attention* towards particular aspects of a system.

Consider a photograph as the model of its subject: photographs enable an observer to predict aspects of their subject, depending on the parts of the photograph towards which the observer chooses to direct their awareness.

According to this understanding, in the following, all models are considered to describe *the reaction* of another system to *the actions* of their interpreter. Action and reaction are referred to as 'emissions' from either the observer or from the observed system.[1]

15.2. A Linguistic Model as an Example

Models consist of symbols. A linguistic model, for example, consists of linguistic symbols, a computational model consists of computational symbols, and a mental model consists of mental symbols (i.e. mental representations).

In general, models establish a relation between these symbols. This can be illustrated with language. The sentence 'the town hall is close to the marketplace' is a linguistic model. It establishes a relation 'close' between the subordinate shapes 'town hall' and 'marketplace'.

The sentence constitutes an icon for all referents with a similar structure. This icon is the content *that the town hall is close to the marketplace.* The referents are all things that are described by this content. The content enables to predict the appearance of 'town hall' if the observer moves towards the 'marketplace'.

A symbol is a shape that is associated to a content. This content can be described, for example, by a model. The shape is a singular element, the contained model is a structure of elements, and the referent is such that the model is an iconic depiction of it.[2] To also remain iconic during *change,* the model must be influenced by a *causal coupling* to its referent.

In a representation, content mediates between shape and referent. The shape *re-presents* the referent because it *indicates* content that *presents* the referent.[3]

Accordingly, extending the above sentence with a label makes it a linguistic *re-presentation:* 'in a *village,* the town hall is close to the marketplace'. Now, the linguistic shape 'village' indicates the presence of referents in which *the town hall is close to the marketplace.* 'Village' now is a symbol for systems where the 'town hall' is 'close' to the 'marketplace' (e.g. villages)—whatever the individual shapes stand for.

Symbols can be nested. A shape can stand for its model in *another,* more complex, model, just like 'town hall' and 'marketplace'. In this sense, symbols can contain various other models and models can contain various other symbols.

[1] This conception of models is developed independently. Elsewhere, models that describe the reaction of one system to the actions of another are also called '*action* models' (for a recent overview, see Čertický 2013). The model concept as it is developed here is *defined* by this capacity. As a consequence, in the following, every model is considered to be an action model as well.

[2] Notice the difference to intentional psychology in section 3.1.2, where content is considered as a *singular* representation without structure.

[3] According to this understanding, representation is *symbolic,* indication is *indexical,* and presentation is *iconic.* The differences are formalised in the next chapter.

In *mental* models, the situation is similar. The difference is that the content in a mental model is not semantic but *intentional* and its shapes are not linguistic but *phenomenal.* Intentional content is considered prior to and necessary for semantic content and phenomenal shapes are considered prior to and necessary for linguistic shapes.

Mental representations that consist of singular shape and structural content enable to define generality and abstraction relations: a composition of intentional content is *more general* than its constitutive content. The phenomenal shape of a mental representation is *more abstract* than the shapes of the mental representations that it contains.

15.3. The World as a Particular Model of Reality

An object is not immediately influenced by the language in which it is described. In the same way, reality is not immediately influenced by the properties of a mental model. The properties of a particular model do not determine *what the model is generated from* but rather *what this source is described as.*

Among the properties of models are, for example, whether they are deterministic or probabilistic, whether they are fully or only partially observable, whether they are continuous or discrete in state and time, and whether they remain constant or adapt to the system that they describe.

Whether a model is probabilistic or deterministic, however, says nothing about whether its referent is probabilistic or not but merely that it *can be described* probabilistically. Accordingly, the properties of a mental model first have an impact on the world as which its cognitive system conceives of reality, not on reality itself.

In a cognitive model for the generation of mental models, therefore, the goal cannot be to describe the model of *an environment with particular properties,* because the properties of external reality are fundamentally inaccessible (see section 3.1.1). The goal rather has to be to generate *a model with particular properties* that reflect how cognitive systems conceive of their world.

These properties are firmly grounded in introspection and can be derived, for example, from Peirce' phenomenological approach to semiotics, from Searle's description of intentionality, or from recent advances in embodied cognition.

15.4. Postulated Properties of External Reality

To simulate processes that generate a model of immediately imperceivable system like external reality, some properties of this system have to be postulated. These premises determine requirements for the processes that are supposed to enable to predict this system's emissions. These requirements immediately constrain feasible architectures of computational cognitive models.

Mental models describe reality as a hierarchical world of objects. However, this model is *generated* from an sequential exchange of unconscious sensorimotor information between cognitive system and external reality (see section 13.4).

Theories on embodied cognition describe this exchange *under temporal pressure.* According to Wilson, these theories

> highlight a weakness of traditional artificial intelligence models, which are generally allowed to build up and manipulate internal representations of a situation at their leisure. A real creature in a real environment, it is pointed out, has no such leisure. It must cope with predators, prey, stationary objects, and terrain as fast as the situation dishes them out. (Wilson 2002, p. 627)

Van Gelder also describes classical models of cognition as "essentially atemporal; there are no inherent constraints on the timing of the various internal operations with respect to each other or change in the environment." (van Gelder 1995, p. 360)

Like in a turn-based board game, traditional artificial intelligence waits for the environment to make a move and expects it to wait just the same. This strategy must fail for real-time interaction with complex and dynamic environments. The time it takes to process sensor information and produce an appropriate motor response is critical.

Time is the main reason why the frame problem is an actual problem. Events in external reality demand an appropriate and *immediate* response by the system. Consider running: bipeds place their feet *just in time* for catching their body weight while continuously falling forwards. Natural running would be impossible if the time to calculate feet positioning would exceed the time it takes to fall flat on your face.

The continual exchange of sensorimotor information between system and reality requires that information can be stored in, and retrieved from, the mental model *reactively* and *on-the-fly.*

The model has to be generated from *interaction* to incorporate the bodily particularities of its cognitive system. This interaction has to be *reactive* to enable the system to act on time.

In summary, the major premises are: 1) external reality enters cognitive systems as a sequence of pre-conceptual emissions (i.e. sensor activations), 2) the system's reactions to these emissions are also pre-conceptual emissions (i.e. motor activations), 3) the exchange of subsequent emissions between both systems must be reactive, 4) and the underlying model must be adaptable on-the-fly.

15.5. A Reactive Model for Dynamic Environments

The previous points set the frame for the information exchange between cognitive system and external reality. This enables to take a closer look into the *general* processing of information in mental models without going into detail on the particular processes that enable, for example, the prediction of external events.

This section first describes how incoming sensor information *propagates* throughout a mental model and how this information is converted into outgoing motor information. According to a classical conception in artificial intelligence, these processes separate perception from action similar to how Brooks describes sense-model-plan-act architectures (see section 7.3). This idea is contrasted against a new, and more flexible, design.

Accordingly, two general ideas are presented on how information can travel along the representations in a mental model. Telecommunication systems provide a viable analogy. A telecommunication system consists of two parties that exchange information. The communicating parties correspond to the cognitive system and its external reality.

Telecommunication systems appear in three broad variants: simplex, half-duplex, and full-duplex systems. In a simplex system, information flows in one direction only (e.g. radio or television broadcasts). Simplex systems are not suitable to describe cognitive systems because they exclude either sensor or motor information.

Half-duplex and full-duplex systems, however, correspond roughly to cognitive systems with two different types of mental model. In the following, cognitive systems according to classical artificial intelligence are described as half-duplex systems. Afterwards, the analogy is used to provide a description of cognition that is inspired by full-duplex systems.

15.5.1. The Structure of a Mental Model

The structure of a mental model enables to infer a flow of information. Therefore, first, it must be outlined how mental representations are organised in a mental model such that they convey to their system *a world of objects.*

Systems generate their mental model during interaction with external reality. This sensorimotor interaction is reactive and its emissions are pre-conceptual to the cognitive system. The model enables, for example, goal-directed behaviour by predicting behaviour (i.e. a sequence of emissions).

Mental models enable to estimate the probability of events. Therefore, relations between mental representations must be *probabilistic.* However, the distribution of probabilities can change with the state of the environment. This means that the model needs to be able to describe *non-stationary* probability distributions.

From the system's perspective, its mental model is *the world* and the intentional content of its subjective mental representations are *objects* (see chapters 3 and 10). These objects are in part-of relations with each other. Consider, for example, that doors can be part of a house as well as part of a car.

This phenomenological premise is reflected in the *nested* symbolic representations of semiotic models (see the introduction to this chapter).

The resulting structure is *a partially ordered set.* The highest element in this hierarchy is a model of the world *as a whole.* The lowest elements are sensorimotor activations that are below the system's level of awareness (see section 13.4). These elements can only be perceived by an observer. They compose the system's *basic perception* (see section 13.3).

15.5.2. A Half-duplex Mental Model

In *half-duplex systems,* only one party can send information at a time. Each party is either sender or receiver, but never both. An example for such a system is a walkie-talkie.

The flow of information in such a system is illustrated in figure 15.1. Half-duplex systems are analogous to how artificial intelligence traditionally conceives of internal models of the environment.

Sensor activation at the lowest level is propagated upwards. On its way it is translated into perception that enables the generation of an internal model of the environment. This model is supplied to the most abstract layer of cognition where the system performs planning.

The level of planning is the only connection between incoming and outgoing information. From here, information takes its way out of the system over task execution and action control to eventually reach the environment in the form of motor activation (see also figure 7.2a on page 59).

Incoming information is passed on up until it reaches the highest level. Consider that processing in each level requires a constant amount of time δ. To reach the highest level, therefore, incoming information requires a time $\delta(L-1)$, where L is the total number of levels in the model.

At the highest level, the incoming information is also processed in time δ. Afterwards, the resulting information is passed on to the level below until it reaches the lowest level. Therefore, to reach the lowest level, outgoing information requires a time $\delta(L-1)$ as well.

From this, it follows that the *total reaction time* of a cognitive system with such a mental model adds up to $\delta(3L-2)$. The more abstract a half-duplex model is, the longer

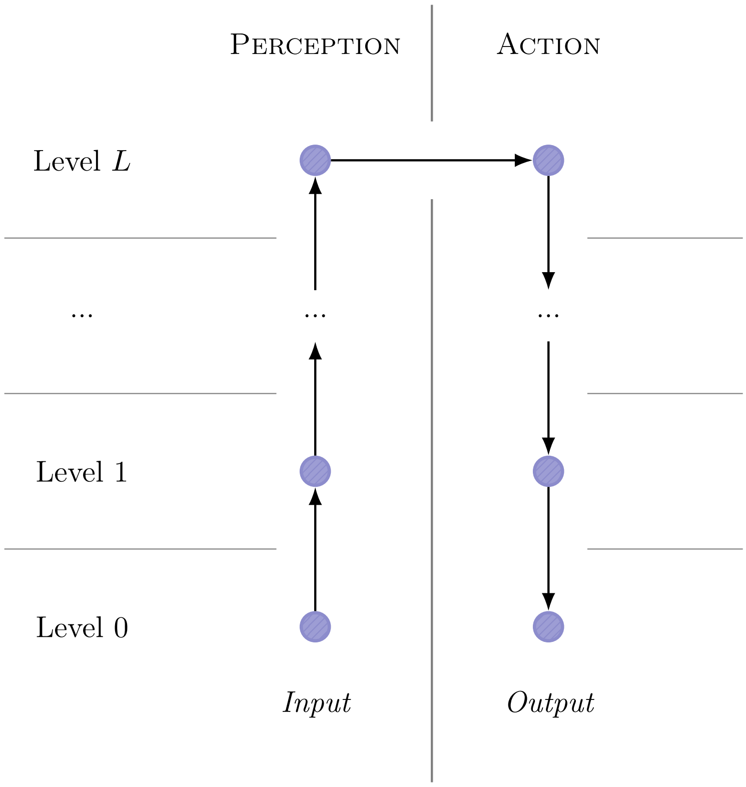


Figure 15.1.: A Half-duplex Mental Model.

it takes for its system to react to the environment.

Susan Hurley refers to systems like these as ‘classical sandwiches’. She describes them as follows. ‘A view of perception and action as separate input and output systems complements a view of thought and cognition as “central” and in turn separate from the “peripheral” input and output systems.’ (Hurley 2002, p. 20)

Figure 15.1 shows the input system at the left and the output systems at the right as two separate components. Both are connected only by a central ‘cognitive module’ at the highest level.

Such a model separates action and perception like the filling of a sandwich separates two layers of white bread. The more processing between perception and action, the bigger their distance, the looser the model’s coupling to reality, and the stronger the system’s detachment from the world.

All cognitive systems with a half-duplex mental model either suffer from the same detachment from reality or from an artificial limitation of abstractness that is necessary to avoid such detachment.

Brooks criticises sense-model-plan-act architectures exactly because they exhibit such a separating phase of rational contemplation in between all perception and action.

15.5.3. A Full-duplex Mental Model

In *full-duplex systems,* one party can send information *before* information from the other party has arrived. An example is the telephone.

The parallel exchange of information is usually achieved with two channels that enable for synchronous communication. This parallelism becomes problematic if information has to leave the system before all relevant information has fully arrived.

To increase the reactivity of a mental model, therefore, it is not enough to provide two separate information channels. Half-duplex models already feature discrete sensor and motor ‘lanes’. The goal is rather to find a way such that information does not need to travel *all the way* from the lowest to the highest level of the mental model.

Figure 15.2 illustrates the flow of information in a full-duplex mental model. Inputs are propagated depending on what type of processing is sufficient (i.e. low-level processing for simple tasks) and necessary (i.e. low-level processing for time critical tasks and high-level processing for complex tasks).

The model describes the environment at each level. If time is of the essence, information can be processed *reactively.* Therefore, time critical information flows in lower levels to enable faster, more immediate reaction.

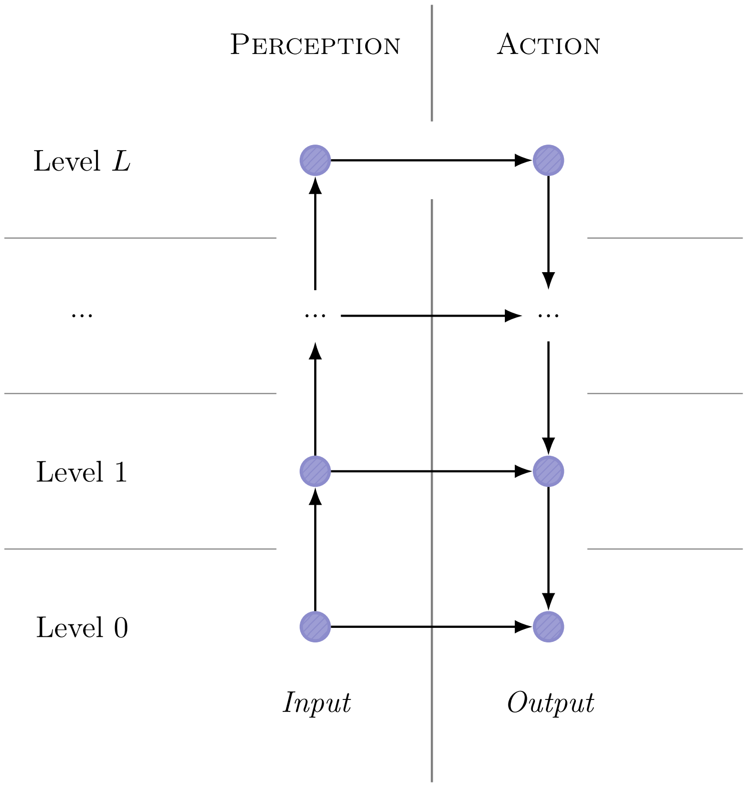


Figure 15.2.: A Full-duplex Mental Model.

More intricate tasks require to propagate information to higher levels to make *deliberative* decisions. Higher levels enable to consider more complex information and long-term dependencies.

In general, the flow of information favours reactivity. By default, information is processed *locally.* Only if the current information is not *recognised* at this level, it is propagated to a more abstract level.

This conditional junction enables information to be processed much faster than in a half-duplex model. Depending on previous experience that is already contained in the model, the ascent of information can be omitted in favour of reactivity.

The required amount of time for reaction in a full-duplex mental model is considerably lower compared to half-duplex models. The worst case is the propagation of completely unknown information to the most abstract level. This requires $\delta(3L-2)$, just like in a half-duplex model.

Repeatedly occurring information, however, can be processed as early as in the lowest level. In the best case, information is processed in δ time, *independent* from the model's number of levels.

16. Formal Basics on Semiotic Models

The formal basics in this chapter are divided into three sections. The three sections correspond to the three components that are relevant to a model (see the introduction to chapter 15). *The described system* is external reality, *the model of this system* is the mental model, and *the interpreter of this model* is the cognitive system.

The first section formalises external reality. As external reality is not immediately perceivable, premises can only be inferred by observing the information *other* cognitive systems exchange with it. The elements of these observations are sensorimotor emissions that enter the cognitive system in form of a sequence but are imperceivable by the system itself.

The second section formalises a model that describes this sensorimotor sequence. This description enables to predict or select its individual emissions. These pre-conceptual elements are composed into the structural content of mental representations in this model, starting from which the system begins to perceive a world.

This model is different from the formalisation of external reality in the first section. A formalisation of mental models must be *cognitively plausible.* A formalisation of external reality, in contrast, is principally independent of human cognition.

The third section formalises the cognitive system as an *interface* between mental model and external reality. This interface relates sensorimotor emissions to one another for them to serve as referents of the representations in a mental model. This interface effectively defines input and output data for the algorithms of a computational cognitive model for the generation of mental representations.

The next chapter provides examples for semiotic models. The next part eventually describes how mental representations are *generated* from sensorimotor referents.

16.1. Formalising External Reality

The literature contains various formal descriptions for systems that emit sequences. In the following, these sequences will also be referred to as 'trajectories'. The most popular formalisations are the different variations of *Markov processes.*

Markov chains are observable systems that are discrete in state and time. Such discrete systems can be conceived of as directed graphs, where the nodes represent emissions and an edge represents the immediate temporal succession from one emission to another. Markov chains can only emit trajectories, where there is a fixed probability distribution over all emissions given the last emission.

Markov decision processes are systems with static trajectories that are coupled to the emissions of another system. The system itself is also referred to as 'controlled' and the coupled system as 'controlling'.

A Markov decision process is a Markov chain where the edges are labelled, for example, with the emissions of a controlling agent (e.g. motor activations) and the nodes are its own emissions (e.g. sensor activations).[1]

Markov chains and Markov decision processes are effectively deterministic finite state automata that enable a system to recognise (i.e. parse) particular environments as well as predict (i.e. generate) the immediate consequences of their actions in such an environment. This is illustrated in the next chapter with the example of the centrifugal governor.

The sequences that can be generated by Markov chains and Markov decision processes, however, are too simple to simulate the information exchange between cognitive systems and external reality. More complex sequences can be generated by partially observable models that feature an internal state.

16.1.1. Partially Observable Models

Models that enable to predict the emissions of another system are usually distinguished into those that assume a hidden state in this system and those that do not. Those that assume a hidden state need to represent this state in order to consider it in their predictions.

As a consequence, models that consider a hidden state in the other system *feature hidden states themselves.* Accordingly, models can be distinguished into *observable models* and *partially observable models.*

Order-n Markov predictors generalise Markov chains by considering not only the single last, but the last n emissions, instead. The sequence of the n previous emissions is referred to as 'history'. In the following, this history is considered as *the state* of the model. The generalisation from the last element to a history of n elements can be applied to all

[1] Besides this graphical conception of an observable system, there is also a propositional conception that uses propositional action languages to describe the environment's reactions to the actions of an agent. The propositional and the graphical variants can be translated into one another. (Gelfond and Lifschitz 1998)

Markovian approaches.

Hidden Markov models have internal states which are explicitly supposed to represent states in the described system that are *hidden* from direct observation. States are assumed to influence only the emissions of their own system.

These models determine probability distributions for the transition from one state into another. Each of these hidden states is associated with a probability distribution of emissions. The sequences generated by hidden Markov models are not controlled: they depend only on the model's internal state but not on the influence of another system like the sensorimotor sequences emitted by cognitive systems. (Rabiner 1989)

Partially observable Markov decision processes are to hidden Markov models what Markov decision processes are to Markov chains. Partially observable Markov decision processes are hidden Markov models where the edges between states and from states to the model's own emissions are labelled with the emissions of a controlling system (e.g. an agent).

Unfortunately, partially observable Markov decision processes are not *goal-agnostic.* They describe not only another system but also a particular goal state for it. This enables, for example, to formalise tasks for an autonomous agent. This is problematic, however, for generating a model that is supposed to facilitate different tasks.

Partially observable systems are partially observable Markov decision processes that are task indifferent. They are 5-tuples $(E, \mathsf{motor}, T, O, \mathsf{sensor})$, where E is a set of hidden states, motor is a set of motor activations, $T : e \times \mathsf{motor} \times E \to [0,1]$ describes the probability of any successor state when performing a particular action in a particular present state, $O : e \times \mathsf{sensor} \to [0,1]$ describes the probability of a hidden state appearing as a particular emission, and sensor is a set of emissions that an agent can receive from the environment as sensor activation.

Unfortunately, the transition function T limits the probability of a transition between states to a stationary probability distribution. The state of external reality, however, must be assumed to transition *erratically* (i.e. according to a non-stationary probability distribution or even an adversary). In the following, systems that emit an erratic sequence will be referred to as 'dynamic'.

16.1.2. Individual Sequences

Recently, the learning of individual sequences has been given a rigorous formal foundation with *on-line convex optimisation* (see, for example, Feder et al. 1992; Littlestone and Warmuth 1994; Cesa-Bianchi, Lugosi, et al. 1999; Cesa-Bianchi and Lugosi 2006). Tasks in on-line convex optimisation always include a player that performs actions. Only after the action has been executed, its outcomes are disclosed.

Outcomes have the form of a loss value associated with the performed action under the current circumstances. Losses do not follow a particular probability distribution. Within limits, these losses can be random or even chosen by an adversary of the player.

Actions are defined in this framework as the convex set of real-valued vectors $\mathcal{K} \subseteq \mathbb{R}^n$ and losses are a family of individual functions $f_t \in \mathcal{F} : \mathcal{K} \rightarrow \mathbb{R}$. (Hazan 2016, pp. 4–5)

An example is on-line classification, where $x_t \in \mathcal{X}$ is the input, $y_t \in \mathcal{Y}$ is the target, and $p_t \in \mathcal{D}$ is the output at time t. Target and output do not need to be from the same set, to allow, for example, a deterministic binary target $y_t \in \{ 0, 1 \}$ but a probabilistic interpretation of output $p_t \in [0, 1] \subseteq \mathbb{R}$.

Target and output are generated by two individual hypotheses $h_t : \mathcal{X} \rightarrow \mathcal{Y}$, where each h_t is from the fixed hypothesis space $\mathcal{H}$. Actions are defined as $\mathcal{K} = \mathcal{D} = [0, 1]$ and the loss at time t is defined as $f_t(x_t) = |y_t - h_t(x_t)|$, where h_t is the agent's current hypothesis and y_t has been generated from an unknown hypothesis.

The performance of on-line convex optimisation procedures is measured in *regret.* Regret is the difference between the actual cumulative loss and the cumulative loss of the single best hypothesis from $\mathcal{H}$ over all samples up to the current time t. (Shalev-Shwartz 2012, pp. 108–111)

On-line convex optimisation covers samples from non-stationary distributions. This generality, however, comes at the price of a predefined hypothesis space. Algorithms effectively learn to 'trust' the best expert hypothesis and, as a consequence, can only be compared against such given expert. (Cesa-Bianchi and Lugosi 2006, pp. 1–3)

Experts can be any generative baseline approach that supplies on-line convex optimisation algorithms with output 'proposals'. Without experts, however, on-line convex optimisation cannot learn anything.

Eban et al. (2012) propose an approach to sequence prediction that learns these experts from scratch. However, their method requires temporally separate training and test examples and, therefore, it introduces the need for a separate training phase to generate $\mathcal{F}$. A real cognitive system, however, needs to learn emission functions continuously and on-the-fly.

16.1.3. Static Trajectories

In dynamic systems theory, a sequence of emissions is referred to as 'trajectory'. Temporally continuous trajectories can be described by *differential equations* and temporally discrete trajectories can be described by *difference equations.* (Robinson 2012)

An erratic sequence of emissions is a discrete trajectory. The subsequence of an emission's predecessors is its 'history'.[2]

Definition 1. *Static trajectories* are sequences of emissions x_t with a history $[\ x_n\ ]_{n=t-h}^{t-1} = x_{t-h}, x_{t-h+1}, ..., x_{t-1}$ of length $1 \leq h$.

$$x_t = f\big([\ x_n\ ]_{n=t-h}^{t-1}\big)$$

The emission of the trajectory after each time step t is a particular unspecified function $f : X^h \to X$ of the current history with length h, where X is the set of all possible emissions. In the following, f is referred to as 'emission function'. With each time step, the history updates according to the current emission.

The Fibonacci sequence, for example, can be formalized as the emission function $f(x_t) = f(x_{t-2}) + f(x_{t-1})$, where $t \geq 2$, $x_0 = 0$, and $x_1 = 1$. This function takes the two previous elements (i.e. $h = 2$) and returns their sum.

With an initial trajectory of $[x_0, x_1] = [0, 1]$, the emission function determines all further emissions starting from $t = 2$ up to infinity. Each element at index t is determined recursively.

$$x_t = \sum_{n=t-2}^{t-1} x_n = x_{t-2} + x_{t-1}$$

$$[x_0, x_1, x_2, x_3, x_4, x_5, ...] = [0, 1, 1, 2, 3, 5, ...]$$

In the case of the Fibonacci sequence, the emission function is constant over time. Therefore, the resulting trajectory is *static.*

16.1.4. Dynamic Trajectories

In a *dynamic* trajectory, f can *change* with t.

Definition 2. *Dynamic trajectories* are sequences, where the emission function f changes *erratically:* The current emission function depends only the current time step t.

$$x_t = f\big(t, [\ x_n\ ]_{n=t-h}^{t-1}\big)$$

[2] Elsewhere, histories are also referred to as 'contexts'. Here, this term is already occupied by Dreyfus. Emissions are also described as *observable.* To use this term in the simulation of a mental model, however, suggests that the simulated system *itself* can perceive them—which is not the case.

For the sake of simplicity, only dynamic trajectories with a history length of $h = 1$ are considered. As a consequence, the emission function in definition 2 can be simplified according to equation (16.1).

$$x_t = f(t, x_{t-1}) = f_t(x_{t-1}) \tag{16.1}$$

The emission function f at time t determines *the state* f_t of a dynamic system.

The dynamic trajectory in equation (16.1) is not influenced by any other system (i.e. it is not *controlled).* However, reality influences cognitive systems in virtue of sensor emissions and cognitive systems influence reality in virtue of motor emissions. In fact, both are so closely connected that their individual trajectories are in a *coupling.*

Unities are considered as coupled "whenever the conduct of two or more unities is such that the conduct of each one is a function of the conduct of the others" (Maturana 1980, p. 136).

Equation (16.2) describes the emissions in the dynamic trajectories of an *agent* that simulates a cognitive system in state a_t and its *environment* that simulates external reality in state e_t after each time step.

$$s_t = e_t(m_{t-1}) \quad m_t = a_t(s_{t-1}) \tag{16.2}$$

The environment's current sensor emission is $s_t \in \mathsf{sensor}$, the agent's current motor emission is $m_t \in \mathsf{motor}$, the current state of the environment is $e_t : \mathsf{motor} \to \mathsf{sensor}$, and the current state of the agent is $a_t : \mathsf{sensor} \to \mathsf{motor}$. None of both systems has access to the other's state, both merely receive its emissions.

This formalisation provides a very general way to conceive of discrete sequences that influence one another. Changes in state are not bound to a fixed probability distribution but erratic. To assume more regularity in the source of sensor emissions from external reality would mean to underestimate the problem that cognitive system have to solve when generating mental representations.

16.2. Formalising Mental Models

This section presents a formalisation for mental models that is based in the subjective conception of mental representation from section 3.1. This conception has been extended semiotically in chapter 13. A formal semiotic conception of mental representation can serve as a foundation to describe mental models as *semiotic models.*

Figure 16.1 illustrates the three semiotic components in subjective mental representation

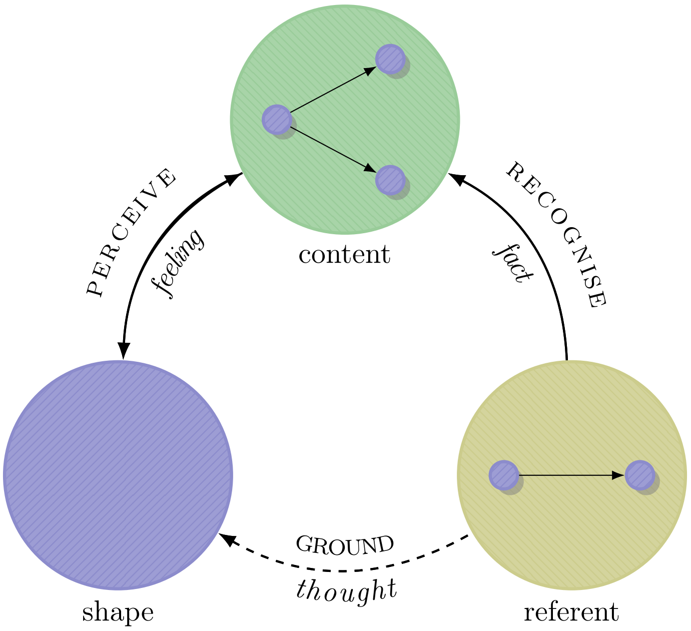


Figure 16.1.: The Directions of Causation in a Mental Representation.

from figure 13.1 on page 96 in more detail. The relations among these components are set in italics.[3] The cognitive processes that establish these relations are set in capitals.

The *shape* is a phenomenal appearance at the current level, the *referent* is a linear transition between shapes of the level below, and *content* is a structural relation among the shapes from the level below. Roughly speaking, content is what an external referent *means,* shape is how this meaning *appears,* and the referent is what is *mentally represented* by the cognitive system.

The relation between phenomenal shape and intentional content defines a *feeling* according to section 13.2.1, the relation between content and referent defines a *fact* according to section 13.2.2, and the mediated relation between referent and shape defines a *thought* according to section 13.2.3.

The process of *recognition* takes a referent and provides a fact, the process of *perception* takes shape or content and provides a feeling, and the process of *grounding* takes a referent and provides a thought.

Content also establishes an *expectation* relation through the cognitive process of *prediction.* In figure 16.1, expectation and prediction are only hinted at in the structure of content.

[3] The *direction of causation* is inverse to the direction of fit in figure 13.1: the mind can only fit the world if the world had a previous causal influence on the mind.

According to section 15.5, information in mental models is primarily processed at the current level and propagated into more abstract layers only when necessary. With figure 16.1, local information processing can be identified as *prediction,* the propagation upwards can be identified as *recognition,* and the propagation downwards can be identified as *perception.*[4]

The ternary relation across adjacent levels enables to formalise a complete mental model. This mental model enables to eventually predict the emissions in a dynamic trajectory according to definition 2 on page 121.

16.2.1. Basic Definitions

Three definitions are fundamental to our formalisation of models and their representations. These definitions follow from chapter 13 and they concern shapes, referents, and content.

According to the direction of causation in figure 16.1, any influence on mental models must initially start from external referents.

Definition 3. Every *referent* $r \in \texttt{referents}_l$ is a 2-tuple $r \subseteq C_l \times S_l$ that consists of elements from the set of all transition conditions C_l and consequence shapes S_l, where $l \geq 0$.

Before describing transition conditions in more detail, the semiotic components of mental representations are introduced. The next component is the *content.*

Deterministic content $m \in \texttt{content}$ is a functional relation $m : \ C_l \to S$ from transition conditions to consequence shapes. If $m(a) = b$, for example, then content m *presents* the referent $r = \langle a, b \rangle$.

However, content can be easily generalised with a frequentist (e.g. axiomatic) interpretation of *probability* (Eells 1999; Cheeseman 2001). The transition frequencies in probabilistic content enable to infer individual *transition probabilities.*

Definition 4. Every *probabilistic content* is a function from transition conditions and consequence shapes to transition frequencies $m : c_l \times S_l \to \mathbb{N}^*$, where $l \geq 0$.

The probability of referent $r = \langle c, s \rangle$ according to probabilistic content $m \in \texttt{content}_l$

[4] Notice the difference to sense-model-plan-act architectures which conceive of all 'upwards' processing as perception and all 'downwards' processing as action. According to embodied cognition (see especially chapters 7 and 8), however, agent and environment are fundamentally coupled. All cognitive processes therefore always already operate on *combined* sensorimotor activations. The next section describes this in more detail.

is $p(m, c, s)$.

$$\begin{aligned} p:\ \texttt{content}_l \times C_l \times S_l &\to [0,1] \subseteq \mathbb{R} \\ \langle m, c, s\rangle &\mapsto \frac{\alpha + m(c,s)}{\sum_{s'}^{S} \alpha + m(c,s')} \end{aligned}$$

The frequency of the given referent as well as the sum of all frequencies are modified by *additive smoothing.* The pseudocount $\alpha \in \mathbb{R}_{\geq 0}$ determines a Dirichlet distribution that defines the expected initial probabilities (Chen and Goodman 1996). If the denominator is zero, the probability of the referent is 1.

Whether probabilistic content presents a referent depends on the threshold $\sigma \in [0,1]$: if $p(m,c,s) \geq \sigma$, then content m presents referent $\langle c, s\rangle$. In this case, the relation between content m and referent r is *a fact,* denoted as fact(r, m).

If the consequence shape in referent $r = \langle c, s\rangle$ is the *most* probable successor of its transition condition $s = \operatorname{arg\,max}_{a \in S} p(m, c, a)$, then, the shape is *expected,* denoted as expectation(m, c, a).

Eventually, *shapes* are defined as follows.

Definition 5. Every *shape* $s \in S$ is the appearance of one unique content. Both are in an injective relation feeling : $S \leftrightarrow$ content, where content is the union of the sets $\texttt{content}_l$ and S is the union of the sets S_l at all $l \geq 1$, such that

$$\forall s_0, s_1 \in S.\ \texttt{feeling}(s_0) = \texttt{feeling}(s_1) \Rightarrow s_0 = s_1.$$

If a content m presents a particular referent r, then the shape s of this content *re-presents* this referent, denoted as thought(r, s).

According to figure 16.1, *feelings* are a relation between phenomenal shape and intentional content. They can be considered as indices that connect adjacent levels in a semiotic model by indicating content at one level with a more abstract shape at the level above. The structure of each semiotic model is a partially ordered set of content.

Facts are a relation between referent and content. They can be considered as icons that present the structure of external reality.

Thoughts, eventually, are a relation between referent and shape. They can be considered as symbols that re-present an external referent.

16.2.2. Definition of Semiotic Models

These basic definitions enable to define *the structure* of a semiotic model.

Definition 6. The *structure* of a semiotic model is a sequence $\Lambda : \mathbb{N}^* \to \{ I_l \mid l \in \mathbb{N}^* \wedge I_l \subseteq \texttt{feeling} \}$ that determines a set I_l for each level of abstraction l such that $\Lambda(l) = I_l$.

Each of these sets, in turn, determines $\texttt{content}_l$ at its own level and S_{l+1} at the level above, such that $S_{l+1} = \{ s \mid \langle s, m \rangle \in I_l \}$ and $\texttt{content}_l = \{ m \mid \langle s, m \rangle \in I_l \}$. The sets S_l and $\texttt{content}_l$ are unique to each level.

As partially observable models, semiotic models are also in a particular *state*.

Definition 7. The *state* of the semiotic model Λ is a sequence $\lambda : \mathbb{N}^* \to S$ that determines the *current* content at level l in virtue of its shape at $l + 1$, such that $\lambda(l) = s$ where $s \in S_{l+1}$, $m \in \texttt{content}_l$, and $\texttt{feeling(}s\texttt{,}\ m\texttt{)}$.

The state of a semiotic model represents the state of the described system. Each index in the state provides content for this level of the model.

This leaves the definition of transition conditions. To establish this, consider what happens if a referent is received at level l that is *not* re-presented by $\lambda(l)$: this level of the state has to transition to a new shape that *does* re-present this referent.

$$\widehat{\lambda}(l) \leftarrow \underset{s \in S_l}{\texttt{arg}\,\texttt{max}}\, p\big(m', \lambda(l), s\big), \qquad \text{where } \texttt{feeling(}\lambda(l+1)\texttt{,}\ m'\texttt{)} \tag{16.3}$$

Equation (16.3) shows that the shape $\widehat{\lambda}(l)$ that is *expected* to re-present the unknown referent r is selected according to the current content m' at the level above.

If $\texttt{thought(}r\texttt{,}\ \widehat{\lambda}(l)\texttt{)}$, then $\lambda(l) \leftarrow \widehat{\lambda}(l)$. If $\neg\texttt{thought(}r\texttt{,}\ \widehat{\lambda}(l)\texttt{)}$, however, then the one shape is selected whose content the referent is most probable according to equation (16.4).

$$\widehat{\lambda}(l) \leftarrow \underset{s' \in S_l}{\texttt{arg}\,\texttt{max}}\, p(m, c, s), \qquad \text{where } \texttt{feeling(}s\text{'}\texttt{,}\ m\texttt{)} \tag{16.4}$$

Again, if $\texttt{thought(}r\texttt{,}\ \widehat{\lambda}(l)\texttt{)}$, then $\lambda(l) \leftarrow \widehat{\lambda}(l)$. If $\neg\texttt{thought(}r\texttt{,}\ \widehat{\lambda}(l)\texttt{)}$ at this point, however, then the model simply cannot describe the referent.

The following sensible transition conditions can be inferred from this.

Definition 8. Every *transition condition* $c \in C_l$ at level $l \geq 1$ is a 2-tuple that consists of a referent $r \in \texttt{referents}_l$ and a shape $s' \in S_{l+1}$ that does *not* re-present this referent.

$$C_l \subseteq \{ \langle s', r \rangle \}, \qquad \text{where } \neg\texttt{thought(}r\texttt{,}\ s'\texttt{)}$$

Both, referent and inappropriate re-presentation at this level, are *the reason* for a transition and, together, they are therefore *a condition* that enables to predict a new shape that may provide an appropriate content for this referent in the future.

16.3. Formalising the Sensorimotor Interface

To describe the generation of a *semiotic model* according to definition 6 on page 126 requires to determine the *referents* according to definition 3 on page 124. This, in turn, requires to determine *transition conditions* and *consequence shapes* in a coupled dynamic trajectory according to equation (16.2) on page 122.

The given definition 8 of transition conditions and definition 5 of consequence shapes both exclude the base level $l = 0$ of the semiotic model. The reason is that, at base level, both depend on the particular types of systems.

In the following chapter, example models are presented according to equation (16.2) on page 122. In these examples, the base transition conditions and consequence shapes are orientations of mechanical components in a *centrifugal governor.*

In the case of a *cognitive system,* the base transition condition is the last sensorimotor activation and the base consequence shape is the current sensor activation. Together, both elements determine the referents of cognitive systems. These referents *couple* system and reality.

$$\begin{aligned} &C_0 \subseteq \mathsf{sensor} \times \mathsf{motor}, && S_0 \subseteq \mathsf{sensor}, && R_0 \subseteq C_0 \times S_0 \\ &c_t = \langle s_{t-1}, m_{t-1} \rangle, && s'_t = s_t, && r_t = \langle c_t, s'_t \rangle \end{aligned} \tag{16.5}$$

Equation (16.5) determines the current transition condition c_t as the the last sensorimotor activation, the according consequence shape s'_t as the current sensor activation, and the current referent r_t accordingly.

Following from this, the coupled dynamic trajectory of cognitive system and external reality can be determined eventually as follows.

$$s_t = e_t(m_{t-1}, s_{t-1}) \quad m_t = a_t(s_{t-1}, m_{t-1}) \tag{16.6}$$

In the part after the following chapter, a computational cognitive model for the generation of mental representations is developed specifically according to equation (16.6).

17. Modelling Controlling Systems

The previous chapter provides a formalisation of semiotic models. This chapter uses this formalisation to model a fully observable and a partially observable system. It can be seen that, in the former case, semiotic models regress to simple Markov processes. The emissions in the second case depend on states that change hidden from any observer. Here, the differences between Markov models and semiotic models show most clearly.

Figure 17.1 shows a centrifugal governor. Its function is to stabilise the speed of a steam engine. The right side of the figure depicts the throttle valve. The left side depicts the flywheel with two connected arms. The hinges of the flywheel arms are mechanically coupled to the throttle valve.

Opening the throttle valve increases the steam throughput. With increased steam throughput, the flywheel rotates faster and the resulting centrifugal forces lift its arms. This effectively increases the angle between the arms and their axis of rotation. The change is mechanically transferred onto the throttle valve.

The mechanical connection establishes a negative feedback between the angle of the flywheel arms and the opening of the steam valve. If the angle between arms and axis increases, then the opening of the valve decreases and *vice versa.*

Due to friction and inertia, mutual influences continually decrease as both angles converge towards a stable attractor point, independent from their initial values.

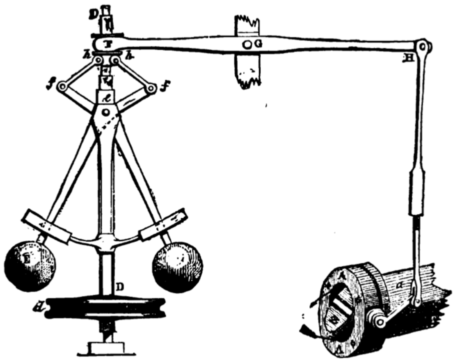


Figure 17.1.: A Centrifugal Governor (Routledge 1900, p. 6)

Time t	Flywheel w_t	Valve v_t
1	50.00	80.00
2	12.00	48.00
3	40.50	72.00
4	19.12	54.00
5	35.16	67.50
6	23.13	57.38
7	32.15	64.97
8	25.39	59.27
9	30.46	63.54
10	26.66	60.34
11	29.51	62.74
12	27.37	60.94
13	28.97	62.29
14	27.77	61.28
15	28.67	62.04
16	28.00	61.47
17	28.50	61.90
18	28.12	61.58
19	28.41	61.82
20	28.19	61.64

Table 17.1.: Data from a Centrifugal Governor.

Because of its two coupled subsystems, the centrifugal governor is a popular analogy to describe the dynamic interaction between cognitive system and external reality. The flywheel ‘controls’ the valve opening just like an agent ‘controls’ its environment. One subsystem co-determines the other.

Usually embodied cognition pursues an approach to cognitive modelling according to this analogy (see, for example, van Gelder 1995). In the following, the limits of this analogy are shown.

17.1. Modelling Statically Coupled Systems

Consider the emissions of a centrifugal governor in table 17.1. The emission of the flywheel after time step t is denoted as w_t and the emission of the valve at the same time is denoted as v_t.[1]

Two static emission functions (i.e. Markov decision processes) can be formalised

[1] The source code used to generate this data can be found in appendix A on page 218.

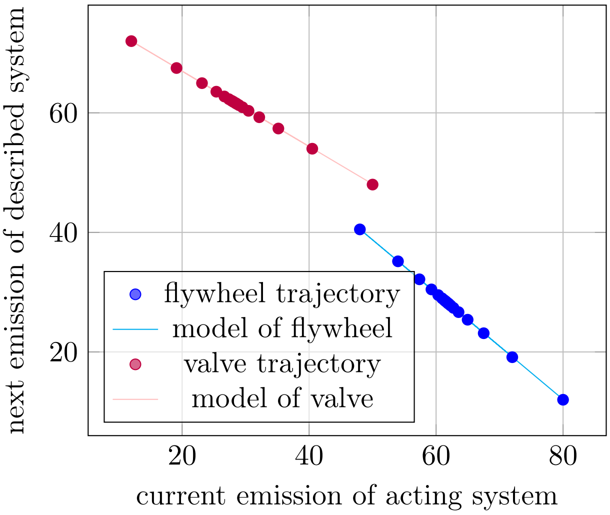


Figure 17.2.: The Coupled Trajectories in a Centrifugal Governor.

according to equation (16.2) on page 122 to describe the coupled trajectories of flywheel and valve. For a model of the flywheel, the transition conditions C are the valve's emissions v_t and the consequence shapes S are the emissions that it receives from the flywheel after the next time step w_{t+1}. For a model of the valve, transition conditions and consequence shapes are inverted.

Both trajectories can be described by the coupled emission functions in equations (17.1) and (17.2).[2]

$$m_w(v_t) = w_{t+1} = -0.89v_t + 83.25 \tag{17.1}$$

$$m_v(w_t) = v_{t+1} = -0.63w_t + 79.58 \tag{17.2}$$

The values of the original trajectories and the interpolated approximations are illustrated in figure 17.2. The attractor for both of these trajectories is somewhere close to $v_\infty \approx 61.7$ and $w_\infty \approx 28.3$.

With equation (17.1) we have a model for the flywheel and with equation (17.2) we have a model for the valve. The individual models of both subsystems can be incorporated into the model of a statically coupled trajectory m_{cg} to describe *the whole* centrifugal governor.

$$m_{cg}(w_t, v_t) = \langle w_{t+1}, v_{t+1} \rangle = \big\langle m_w(v_t), m_v(w_t) \big\rangle \tag{17.3}$$

[2] The approximations have been obtained by simple linear regression.

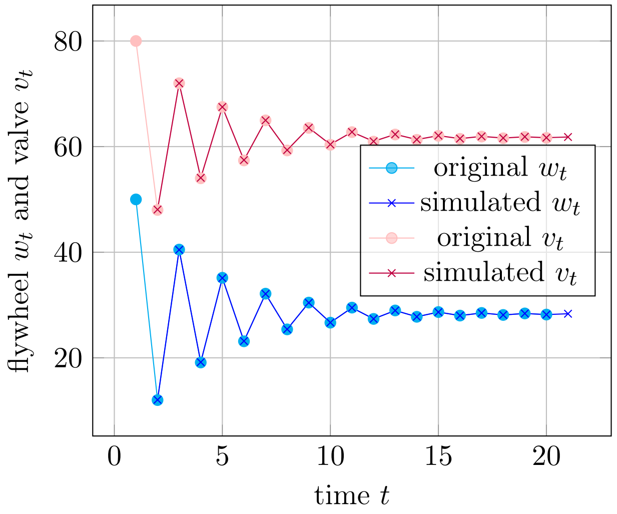


Figure 17.3.: A Centrifugal Governor and its Simulation.

Equation (17.3) describes the whole centrifugal governor as a coupling of two individual subsystems. Each transition condition in this new model is a tuple that couples transition conditions from each of the previous models and each consequence shape is a tuple that couples their respective consequence shapes.

The models for these subsystems are *mutually dependant.* A composite model that incorporates these dependencies, however, is *independent* from another system. This enables the autonomous simulation of a centrifugal governor, starting from an arbitrary initial state.

Figure 17.3 shows that data that has been generated by a simulation starting from the same initial state develops almost identical to the trajectory of the original centrifugal governor.[3] The complete model captures the negative feedback that characterises the centrifugal governor. Therefore, it can be considered to be an appropriate description.[4]

17.2. Controlling Erratic Systems

Recall that the flywheel of the centrifugal governor is considered analogous to a cognitive system and that the valve is considered analogous to external reality. Now assume that, every once in a while, the steam valve shuts completely and remains stuck.

This occurs without warning and for an indefinite amount of time. After the valve releases, the centrifugal governor resumes normal operation just as before. How can this

[3] For the differences between both trajectories, compare table 17.1 to table C.1 in appendix C.
[4] The source code for the simulation can be found in appendix B on page 219.

erratic behaviour be integrated into a model?

$$m_v(w_t) = v_{t+1} = -0.63w_t + 79.58 \tag{17.4}$$

$$m'_v(w_t) = v_{t+1} = 0 \tag{17.5}$$

Equation (17.4) describes the valve in its functional state and equation (17.5) describes a defunct valve that remains shut over an extended period of time. As a consequence, the orientation of the flywheel remains constant as well, although its emission function m_w from equation (17.1) on page 130 has not changed.

A comprehensive model of the erratic valve has to cover defunct as well as functional behaviour. Sequences of emissions that have been obtained during one type of behaviour, however, are not covered by a model of the other.

An appropriate model of the valve needs to *switch* between both subordinate models for the individual behaviours as soon as they occur. Both low-level models need to to be incorporated into *a more general model* such that each individual model can serve as *the content* for a representation of one particular kind of behaviour.

The erratic valve is a dynamic system because it emits a dynamic trajectory. Each of its behaviours can be considered as an individual emission function. Changes in this emission function, in turn, imply *a change in its state.* An appropriate model for the erratic valve, therefore, requires a state as well.

17.3. Modelling Dynamically Coupled Systems

A semiotic model for an erratic steam valve can be inferred from these definitions.

First, this requires the set of base indices I_0. In our case, this set consists of the emissions of the valve. These emissions are values from $\mathbb{R}_{\geq 0}$ that present the opening degree of the valve.[5]

Second, the indices in I_1 must be determined. These indices are the models m_v from equation (17.2) for the functional state and m'_v from equation (17.5) for the defunct state. For simplicities sake, their shapes are selected according to their denotation as 'm_w' and 'm'_w'.

To describe the possible transitions between the shapes of these indices requires at least one more index $\langle$'m_2'$, m_2\rangle \in I_2$. The content m_2 of this index is a functional relation from C_2 to S_2.

[5] Each element in $\mathbb{R}$ has a particular shape and a particular content according to our definition of the indices at each level in definition 6 on page 126.

The transition conditions in C_2 relate shapes from S_2 with those referents from R_1 that are not represented by them. The referents in R_1 are tuples from $C_1 \times S_1$.

Like in the level above, the transition conditions in C_1 are tuples that consist of shapes from S_1 and referents from R_0 that are not represented by them. The second element in the tuples from R_1 are those shapes from S_1 that *do* present these referents.

To cover all possible referents, m_2 can be defined according to equation (17.6).

$$m_2 = \Big\{ \ \big\langle \langle s', r \rangle, s \big\rangle \mid \neg\texttt{thought(}r\texttt{, }s\texttt{')} \wedge \texttt{thought(}r\texttt{, }s\texttt{)}, \langle s', r \rangle \in C_2, s \in S_2 \ \Big\} \tag{17.6}$$

From the indices at each level, a semiotic model Λ for the erratic valve can be determined with equation (17.7).

$$\begin{aligned} \Lambda &= [I_0, I_1, I_2] = \\ &= \Big[\mathbb{R}_{\geq 0}, \{ \ \langle \text{`}m_v\text{'}, m_v \rangle, \langle \text{`}m'_v\text{' }, m'_v \rangle \ \}, \{ \ \langle \text{`}m_2\text{'}, m_2 \rangle \ \}\Big] \end{aligned} \tag{17.7}$$

The state λ of the model can be defined according to the initial orientation of the valve in the real centrifugal governor and its initial behaviour (i.e. functional).

$$\lambda = [\text{`80'}, \text{`}m_v\text{'}, \text{`}m_2\text{'}]$$

The model for the erratic valve in equation (17.7) describes a dynamic trajectory that *changes* over time. It covers this dynamicity as soon as it occurs.

In contrast to the model in equation (17.2), the semiotic model in equation (17.7) also features discrete representations similar to the representations in a mental model.

The semiotic model describes the valve in a coupling with the flywheel but *the model itself* is also coupled to the real valve. It reacts to the system's emissions by changing *its own state* according to section 16.2.2.

17.4. Centrifugal Governors and Cognitive Systems

Let us revisit the initial analogy from the introduction to this chapter: a centrifugal governor is helpful to cognitive modelling because it can serve as an analogy for what cognitive systems do with their environment. The flywheel exerts control over the valve just like the cognitive system exerts control over reality. Several points occur under the impression of this chapter.

Firstly, the valve controls the flywheel no more than the flywheel controls the valve. Really, it appears more to be a matter of perspective who controls whom.

Secondly, the flywheel can be described individually. However, it cannot be described *independent* from the valve. Transferred to cognitive modelling, this supports the emphasis of embodied cognition on the embedding of the system's body into a physical reality.

Just as the flywheel cannot carry out its controlling function without a steam valve, the feedback of which is to be controlled in the first place, *systems cannot be cognitive* without the feedback of an erratic environment.

According to Maturana, van Gelder, Wilson, and various others, cognitive systems are more or less *coupled* with external reality. The centrifugal governor provides an analogy to illustrate the immediacy of such a causal influence.

Developing a model of the whole centrifugal governor can show parts of what cognitive systems usually do when they generate their own mental model of reality. Even more importantly, embodied cognition suggests that the intentional content of mental representations is coupled with external reality similar to how flywheel and valve are coupled.

However, the analogy between valve and cognitive system does not extend to the mental model of a cognitive system. The flywheel does not feature a model of the valve in the same way that cognitive systems feature a model of their environment. The models in the previous section are merely models *to us,* not to the flywheel itself.

Nowhere in the functional centrifugal governor are there representations like the mental representations of reality we experience to have ourselves. A possible explanation is that the coupled trajectories of flywheel and valve are *static.* A single emission function can describe the negative feedback between both systems. To predict this interaction, representations are just not necessary.

Due to the lack of individual representations for different behaviours, the flywheel is unable to react to the valve in its defunct state. This shows in the simple fact that a valve with erratic behaviour causes the whole centrifugal governor to become inoperative.

18. Conclusion

This part describes mental models as *semiotic models.* To achieve this, first, the general hierarchical structure of mental models is presented. Then the representations in a mental model are described as well as the relations among them.

The type of system that a mental model is supposed to describe is formalised as emitters of *dynamically coupled trajectories.* Cognitive systems generate their mental model from such a trajectory that is emitted by the environment and coupled to the system's *own* dynamic trajectory. The basic emissions in both trajectories are identified as *sensorimotor activations* at the physical border between cognitive system and external reality.

The two subsystems of the centrifugal governor are a popular analogy in embodied cognition to describe the reactive interaction between cognitive systems and external reality. However, the analogy ends when mental models are considered.

A modification to the centrifugal governor is conceived which makes it more similar to the coupled system of cognitive system and external reality. The state of a model has to be updated to provide predictions that consider the current state of the modelled system. *A dynamic change of state* in the valve makes clear the need for 'stateful' mental models.

A type of *semiotic model* is developed that enables to predict emissions despite this change. The predictions of partially observable models—including semiotic models—depend on this state.

The state of a semiotic model enables a cognitive model to describe the belief state of a cognitive system more appropriately than the states of other types of partially observable models.

Where the states of other models represent by definition *unobservable* states of the system they describe, the state of a semiotic model is constructed only from observable emissions of the other system.

Part V.

Simulating the Generation of Mental Models

19. Introduction to Cognitive Models

Different types of models describe different types of aspects in another system. *Algorithms* are models that describe a system in virtue of its *emissions*—not, for example, in virtue of its *nature.* They provide instructions on how the emissions of the system can be generated.

Algorithms can be implemented as *programmes.* By sequentially executing each individual instruction in this programme, a computational interpreter can generate an *instance* that mimics the original system with respect to its emissions. In the case of algorithmic cognitive models, these emissions consist of the system's behavioural data (see the introduction to chapter 15).

The individual actions that make up behaviour can also be considered in a *wider* sense to include the cognitive manipulation of mental representations (see section 3.3). As a consequence, an algorithm can be instantiated to simulate cognition in virtue of *subjective experience as emissions to first-person-perspective.* This instance is a computational process which, in turn, is a temporal sequence of physical states in a computer system.

Accordingly, four physical entities are involved in the computational simulation of a cognitive system: 1) *the real system* (i.e. a cognitive system), 2) *the algorithm* that describes the behaviour of this system (i.e. a cognitive model), 3) *the programme* that implements this algorithm in a particular syntax (i.e. a programme for the simulation of a cognitive system), 4) and *an instance* of this programme (i.e. the physical simulation of a cognitive system).

Algorithm and syntax determine all the variations of a particular programme and *vice versa.* In the following, the difference between algorithm and programme is therefore omitted. Three physical entities remain.

A good *algorithm* describes *the system*'s behaviour appropriately and comprehensively. A good *simulation* realises only and exactly behaviour as it is described by this model. Goods simulations, therefore, cannot be distinguished from the real system with regard to their algorithmic model.

If some thing is designed according to a model for X, then this thing *is* an X according to the model. If differences begin to show between the artefact and X, then these differences can be incorporated into the model to ensure that the model still describes the artefact as an X.

19.1. Computational Simulations of Cognition

A computational simulation of cognition has various benefits. Among the most important ones are explicitness, workability, empirical verifiability, and the ability to evaluate cognitive models.

Firstly, the strict formal requirements of programming languages require models to be *explicit.* This allows to determine inconsistencies early on. The resolution of these inconsistencies can improve a model without even comparing its instance to the behaviour of a real cognitive system.

Secondly, computational simulations are *workable.* They can be applied to various tasks in different environments. They can also be plugged into sensorimotor interfaces that emulate a variety of different bodies.

Thirdly, simulations are *empirically verifiable.* Computational simulations supplement the empirical methods of cognitive sciences in general (Schmid and Kindsmüller 1996, p. 24) and the empirical methods of psychology in particular (Strube 1996a, p. 317; Strube 1996b, p. 408).

Fourthly, cognitive models can be *evaluated* by computational simulations. If the simulation deviates from the behaviour of a real cognitive system, then the underlying model is in need of improvement or revision. This is the case, for example, if a cognitive system and its simulation have a similar task in a similar environment but show significantly different behaviour. (Schultheis 2013, p. 101)

Most important in our case, however, is that computational simulations can not only support or weaken computational models of *observable* behaviour. Simulations also provide observable simulations of mental processes that are *prior* to any interaction with the environment. Models can be evaluated by how well their instances imitate mental processes that can only be experienced by oneself.

Take a chess computer during the evaluation of the board position. Its internal processes are crucially different from the mental processes that humans experience in the same situation. The chess programme, therefore, may be good at playing chess but its underlying algorithm is not a good cognitive model of human chess players. (Strube 1996b, p. 407; Strube 1996c, pp. 546–547; Schultheis 2013, p. 101)

19.2. Convergence in Artificial General Intelligence

One intention behind the simulation of cognitive processes is to solve real-world problems. This goal is approached from many different directions: pattern recognition, knowledge representation, and natural language processing—only to mention a few.

A popular contemporary view is that these different fields will eventually converge into *artificial general intelligence.* According to this assumption, the vast array of cognitive abilities only appears heterogeneous on the surface.

Actually, however, they are supposed to be merely different manifestations of the same underlying principle. Once this principle can be described and simulated with a computational system, diverse cognitive abilities emerge from it.

This view is consolidated neurologically by the *equipotentiality* of neural tissue. Equipotentiality is the property of functional parts of the brain (e.g. individual neurons or neural columns) to compensate for any other functional part that might be incapacitated.

This property suggests that all the subsystems of the brain implement *the same basic cognitive processes* which underlie the various manifestations of intelligence. Once these processes can be described, in principle, computational systems could be programmed to implement them as well. (Mountcastle 1978)

Identifying the equipotential processes in *original* intelligence by integrating diverse approaches to *artificial* intelligence is problematic. The perspectives, approaches, and goals are way too numerous to break them down into a smallest, but still meaningful, common denominator.

Instead, it appears to be more promising for the simulation of original intelligence to start from basic cognitive processes *as they are experienced* by real cognitive systems (see, for example, prediction, perception, recognition, and grounding in the introduction to section 16.2).

19.3. Pragmatic Artificial Intelligence

As long as an essential difference is assumed between real cognitive systems and their simulations, research will attempt to pinpoint this difference before it starts to realise workable systems. If this difference is assumed to be *intentionality,* however, then workable systems are delayed indefinitely: there is no *observable* difference between intentional and non-intentional systems.

Could intentional content be observed from an external perspective, it would already be crucially different from the intentional content in a real cognitive system. According to *pragmatic artificial intelligence,* the non-verifiability of intentionality in *any* other system is *sufficient* to accept that artificial systems simulation and real cognitive system can both be equal instances of the same cognitive model.

If intentionality cannot be observed, however, how can a simulation for the generation of intentional content be evaluated? To simply accept the simulation as cognitive would beg the question *what it really is* that makes it cognitive and, eventually, this would mean to make the same mistake as strong artificial intelligence.

From the stance of pragmatic artificial intelligence, a successful simulation of cognition depends neither on the realising material (e.g. computational or biological) nor on its particular actions in an environment. It rather depends 1) on whether the simulation is instantiated *in reality* or *in a simulation of reality* and 2) on whether the model enables its system to *recognise* parts of the environment.

The first condition embraces Searle because the originality of intentional content depends on *the origin* of the system's experience. Only external, immediately imperceivable reality can provide the ground for original content. In a simulated environment, in contrast, all of the system's content must be *derived* from content in the mind of the simulation's designer (see section 12.1).

The second condition contradicts Searle because it assumes that simulations are able to recognise external referents *in the same way* that real cognitive systems do: with original intentional content. Searle rejects the possibility that computational systems can associate intentional content to anything.

Comparably strong conditions can hardly be inferred from the stance of strong artificial intelligence or weak artificial intelligence (see section 12.3). Weak artificial intelligence only offers *descriptions* of cognitive systems. It provides content to the observer but not to the simulating system.

Strong artificial intelligence, on the other hand, completely ignores the particular model behind a simulation. Instead, it pulls back *to a behaviourist position* from which everything that 'acts cognitively' is also supposed to *be* cognitive.

Weak artificial intelligence lacks the ambition and strong artificial intelligence the perseverance necessary to simulate the content of real mental representations. This shortcoming contributed immensely to the popularity of the Chinese room argument but it also summons severe practical problems (see chapter 2).

To cognitive systems that control non-human bodies, many human concepts are fundamentally irrelevant. A wheeled robot whose mental representations have the content of a two-legged human will have critical problems, for example, when interacting with staircases.

Pragmatic artificial intelligence emphasises the need to determine *to whom* a shape is associated with content and, therefore, it pays respect to these problems. Shapes that are only interpreted by the observer can merely be part of a description. Shapes that are interpreted *by their own system,* however, can be subjective representations with content that is intrinsic to their own system. If these shapes can *only* be interpreted by their *own* system, then the resulting content is necessarily original.

19.4. Summary

Mental representations make permanent objects from elusive basic perception. A fundamental explanation for intentional content must therefore start by accepting subjective experience. This experience can only be accessed by the described system—be it biological or computational.

A comprehensive cognitive model must also describe the mental model in a cognitive system. The shapes in a model must have content to an observer. Otherwise, the model would not describe anything to anyone.

In a simulation of pragmatic artificial intelligence, however, the external content that an observer associates with simulated mental representations is considered *secondary* to an interpretation by the system *itself.* The system's content must be considered *epistemologically prior* to content in a description of this system—just like it is the case with real cognitive systems.

An algorithmic cognitive model can describe the generation of this content as the procedural generation of a semiotic model. Such a cognitive model for the generation of mental models concerns the cognitive sciences as well as machine learning.

It concerns the *cognitive sciences* because it is cognitively justified. Cognitive plausibility is crucial to the processes that are described by cognitive models. Plausibility can be achieved in virtue of theories that provide explanations for these processes (e.g. phenomenology).

The cognitive model also concerns *machine learning* because the computational generation of a semiotic model is *an automatic acquisition of knowledge.* Computational procedures can be implemented and instantiated as machine learning processes to simulate the cognitive processes responsible for knowledge acquisition.

A computational simulation of symbol grounding requires both: an appropriate theoretical foundation that enables to interpret the algorithm as a model for cognitive processes as well as an empiric analysis of the computational processes that it instantiates. The first half of this work provides the former, the second half provides the latter.

20. The General Types of Machine Learning

The field of machine learning is concerned with the generation of data structures that provide knowledge to computational systems. According to Tom Mitchell, "[a] computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell 1997, p. 2)[1]

Experience is the information acquired during processing a particular task. Proficiency in this task shows in an increase in performance. This definition enables the empirical analysis and qualitative comparison of different learning algorithms.

A checkers learning programme, for example, has the task to *win games of checkers,* its performance can be measured by *the number of games won,* and it can obtain experience by *playing checkers.* The instances of different algorithms can be analysed empirically by putting them up against the same opponents and by comparing the number of games they won.

What is the available information, task, and performance measure in a simulation for the generation of a mental model? The available information is sensor and motor data that results from its system's interaction with an environment (see section 16.3). The task is the generation of a model for this environment and the performance measure is the system's performance in predicting emissions of this environment.

The various algorithmic approaches to generate knowledge are commonly classified into three different types: unsupervised learning, supervised learning, and reinforcement learning. In this part, an overview over these three classical types of machine learning. The next part shows that the learning of a semiotic model concerns all of them. Due to the fact that the sequentiality of information is an important constraint in a simulation for the generation of mental representations, different cases of sample availability are presented as well.

[1] To distinguish experience and knowledge according to Mitchell from their phenomenological interpretation, Mitchell's conception of experience is referred to as 'samples' and to the phenomenological conception of knowledge as 'content'.

20.1. Unsupervised Learning

Unsupervised learning is the autonomous generalisation of input. The input in unsupervised learning can be any sort of singular, structural, nominal, or rational data.

Generalisation is performed by any procedure that groups inputs. Labels that are manually assigned to these groups are usually considered to be more abstract than their according input samples. *Abstraction* is the representation of a group of input samples by the group's respective label.

Unsupervised learning procedures infer properties from each input that enable to assign this input to a group. These properties describe the input with previously unknown information that is *relative* to all available inputs (e.g. frequency, redundancy, or selectivity).

Particular tasks in unsupervised learning include structural classification (i.e. clustering), fraud detection (i.e. the identification of anomalies), or the elimination of redundancies (i.e. data compression). These tasks are not mutually exclusive but rather different perspectives on the same type of generalisation.

Consider, for example, a procedure that sorts a bag of fruits into groups of colour. The colour of each fruit is the relevant property and the name of this colour is a label that serves as an abstract representation for each individual fruit in this group.

It might appear as though this procedure does not need to infer anything. After all the fruits prominently feature a particular colour already. Consider, however, that it must be decided, for example, whether oranges receive the same label as apples (i.e. 'red'), the same label as lemons (i.e. 'yellow'), or their very own label (i.e. 'orange'). This decision must be *inferred* in relation to all colours and it is far from obvious what the best choice is in a particular setting.

Generalisation properties can be inferred by *feature selection* (i.e. ignoring particular properties of the input), by *instance selection* (i.e. ignoring some inputs completely), by *feature discretisation* (i.e. the quantification of continuous inputs), and by *feature construction* (i.e. the generation of new properties). (Saitta and Zucker 2013, pp. 278–290, concerning supervised learning)

As soon as the relevant colours have been decided upon, each fruit can be assigned to its according *cluster*. This enables to identify *anomalous* fruits, as well as to generate a *compressed* description of the bag's original contents from the number of fruits in each group—in contrast to an extensive enumeration of each individual fruit.

The performance of unsupervised learning procedures a relative matter. Although all methods perform *some* kind of generalisation, one might outperform the other at data compression while being rather inappropriate for fraud detection. To account for this, a quantitative comparison of different procedures in unsupervised learning always requires to define a particular *measure of performance.*

20.2. Supervised Learning

Supervised learning is the instructed generalisation of input. Instructions are provided in the form of *examples.* Each example is a tuple of *input* and *target* data. The target is a label for the input.

The supervised learning of numeric rational targets is *regression.* Learning nominal targets is *classification.* All forms of regression can be described as the classification of input into discrete numeric intervals but not all forms of classification can also be described as a regression of input into rational targets.

Examples enable to infer properties that generalise over all inputs with the same target label. The goal of supervised learning is to infer properties that apply not only to the example inputs but to *unseen inputs* as well.

A trained supervised learning system receives input, recognises its generalisation properties, and outputs the target label that correlated with this property during instruction.

Consider observing several groups of fruits. The example data in this case is a set of tuples that consist of a particular fruit (i.e. the input) and a colour label for the group to which the fruit is assigned (i.e. the target).

Supervised learning is to infer one or more properties that 1) enable to reproduce the example grouping as best as possible and 2) enable to group unknown inputs appropriately.

Supervised learning has been successfully applied to highly structured input data, such as digital images (e.g. object recognition), as well as strongly related singular data, such as medical records (e.g. decision trees).

20.2.1. Bias and Variance

The success of a supervised learning system is evaluated with its *error* on a particular set of examples. The error is the difference between the actual *target* values and the *output* values generated by the system.

Supervised learning systems usually reduce the error during learning with each individual example and with each iteration over all examples—otherwise they do not learn anything.

An obvious source for error is *noise* in the data set. Noise manifests, for example, in *non-functional example sets* where the same input has more than one target.

Even if the example set is functional, however, there are two more causes for error. One is the system's bias and the other is variance in the set of examples.

Bias is the tendency of a learning system to choose one input property over another *given the same empiric evidence.* A system that selects only one relevant property, although the examples would support more than one, has a relatively strong bias. Systems that select more properties have a relatively weak bias.

To consider all properties of the input as relevant for the target label makes a system completely *unbiased.* An unbiased system, however, can only provide informed outputs for input from the example set. To such a system, each previously unseen combination of properties is a unique new input and, therefore, it provides no information on an according target. (Mitchell 1990)

Completely unbiased systems learn examples 'by heart' and, therefore, *do not generalise at all.*[2] In general, all weakly biased systems tend to perform badly on unknown examples because some bias is necessary for generalisation. The stronger the bias, however, the greater the chance that a relevant property is not considered by the system.

The second reason for error is *variance.* It describes the diversity of inputs for each target. The diversity of inputs determines the amount of bias necessary to minimise error. For a small error under a strong bias, variance must be low. A strong bias only works well on homogeneous data. For a small error under high variance, the bias must be weak. With high input variance, there is often no choice but to learn 'by heart'.

Consider examples of coloured shapes, in which all blue squares are in one group and all red circles are in another. The inputs have a low variance in each group. Each individual input property determines its target. Depending on whether the system is biased towards colour or shape, new inputs (e.g. blue triangles or green circles) can be assigned accordingly.

Now consider two other groups: one in which there are only red squares and blue circles, and one in which there are only blue squares and red circles. With such an example set, a strongly biased system cannot determine a single property as relevant for the target groups. An unbiased system would be able to correctly reproduce the example set. However, this is paid for with the ability to generalise *beyond* these examples.

20.2.2. Generalisation Error and Overfitting

A weak bias enabled to minimise the error for the example set from which the system is trained. But the goal is to train a system such that it produces little error for *unknown*

[2] Learning by heart is usually referred to as 'overfitting'.

input as well—the inputs in the examples are already labelled correctly after all.

Consider two example groupings of fruits. One is the *training data* from which properties are inferred. The other is the *test data,* onto which the these properties are applied to verify their ability to predict the expected labels.

The *training error* is the deviation of output from target in the training examples. The *test error* is the according deviation in the test examples. The *generalisation error,* eventually, is the test error minus the training error.

High training error is a sign for *underfitting.* This occurs in systems that are strongly biased towards the wrong, or not enough, properties.

The more variant the data, the less biased a system must be to perform well. More properties must be considered and so the chances increase that the system becomes 'superstitious' by assigning targets to the wrong (i.e. correlating but causally irrelevant) properties.

High generalisation error is a sign for *overfitting.* To determine the generalisation error requires separate training and test example sets. Overfitting occurs, for example, if the training examples are less variant than the test examples.

Without *independent and identically distributed* examples in both sets, the test error cannot provide any insight on how well the system generalises its input. Bots sets must must not influence one another but both need to follow the same underlying dynamics.

Overfitting can also result from the consideration of too many properties. The system must be biased enough to avoid an unintentional specialisation in *only a particular subset* of all the relevant data (i.e. the training set). The ideal bias is determined by the variance of the data at hand.

20.3. Reinforcement Learning

Reinforcement learning is the procedural discovery of goal-directed behaviour from experience. The *experience* of a reinforcement learning agent is a 3-tuple that consists of a sensor emission from the environment s_t, the agent's motor emission a_t as a response, and the reward r_t received from the environment as a response to this action. This exchange is assumed to take place during one time step (i.e. in between t and $t + 1$).

An experience could be, for example, the state of a chess board (i.e. sensor emission), the move performed in this state (i.e. motor emission), and a reward depending on whether a black, a white, or no chess piece has been captured after performing this move. Systems can learn from this experience which action to perform during a particular sensor perception to maximise their reward over time.

Although *reward* is usually real-valued, goal-directed behaviour can be learnt from as little as two binary reward values: the system receives low reward (i.e. punishment) until the designated goal is achieved. As soon as the goal is achieved, it immediately receives high reward during one time step.

The behaviour of an autonomous agent in its environment follows a particular *action policy.* Policies can be better or worse suited to achieve a certain goal. With growing experience, reinforcement learning agents change their policy to adapt to their environment and reach their goals more efficiently.

The performance of a particular policy is determined in virtue of the *cumulative reward* that the agent gathered from the environment in a fixed amount of time while following this policy. This gives an implicit preference to more efficient policies that enable the agent to achieve its goal *faster,* in case $r_{\text{default}} < 0$.

Experience is determined by the dynamics between agent and environment. Their actual sequence of interactions can be described by a hidden Markov model. The environment of an agent that learns its own actions, however, must be described as *decision processes* that couples the emissions of the environment to the emissions of the agent.

20.3.1. Exploration and Exploitation

To improve their action policy, reinforcement learning agents need to obtain *new* experiences. They must find a trade-off between the *exploration* of their environment to learn something new and the *exploitation* of previous experience.

More specifically, the dilemma is whether an agent should perform the best action *it already knows about* or whether it should try to find a *new* action that might yield even more reward in the long run.

A common approach to deal with this problem is a *constant ϵ-greedy strategy.* First, a fixed threshold $\epsilon \in [0, 1]$ is selected. Before each action, a uniformly random value from the same interval is drawn. If this value is below ϵ, the agent performs a uniformly random action, otherwise it performs the action that it assumes to be the best.

Although there are many alternatives to a constant strategy (for an overview, see, for example Takahashi et al. 2009, p. 243), this approach is rather attractive for evaluation purposes. Its effects on the eventually accumulated reward can be determined easily because, over time, the influence of ϵ-exploration approximates ϵ times the average reward over all actions.

20.3.2. Policy Learning

Given an appropriate exploration strategy, there are two general approaches to optimise the policy of an agent towards maximal cumulative reward.

The first one is *policy iteration:* The agent selects a policy, executes it for a fixed amount of time, and then compares the accumulated reward with the reward accumulated during previously executed policies.

If policies are defined in a metric parameter space, classical optimisation (i.e. hill-climbing procedures) can be applied. The problem with this approach is that the parameter space (i.e. the number of policies) might be very large or even infinite.

The second general approach is *value iteration.* Here, policies are defined indirectly, by the agent's *evaluation* of actions under different circumstances. If the agent always chooses the action that is evaluated best then the evaluation determines exactly one policy.

Actions can be evaluated by the *estimated* cumulative reward that the agent expects to receive *after* their execution (i.e. temporal difference value iteration). In tasks where the immediate reward might be more important than future reward, the agent learns to estimate the *discounted* cumulative reward. The discount is determined by a factor $\gamma \in [0, 1]$ which describes the decline of reward importance with each step into the future.[3]

The general procedure of value iteration algorithms is as follows: after each time step, the agent explores or exploits the environment (i.e. it performs either a random action or the one that is evaluated best). In the same instant, the evaluation of the *last* action is adjusted according to the *current* reward and the evaluation of the *current* action, discounted by γ.

Notice that, in contrast to usual supervised learning, in value iteration reinforcement learning, training and test are *simultaneous.* The agent's actions are an immediate *test* of the learnt evaluation. In the same instant, the agent also *trains* its goal-directed action selection by adjusting the evaluation of its last action.

20.4. On-line Learning

The data for machine learning systems comes in the form of samples. The samples for unsupervised learning are *inputs,* the samples for supervised learning are *examples,* and the samples for reinforcement learning are *experiences.*

[3] Inferring a policy via value iteration effectively translates reinforcement learning into the supervised regression of discounted cumulative reward.

Three general cases of sample availability can be distinguished. The first case is *batch learning,* where the full data is randomly accessible to the learning system at all times. The second case is *active/semi-supervised learning,* where individual samples are provided by a teacher whenever considered necessary. The third case is *on-line learning,* where individual samples become available in a stream of data over time.

The generation of a mental model is *learning from a sensorimotor trajectory* and, therefore, a case of on-line learning. Unfortunately, the term 'on-line learning' has at least three different meanings. *On-line convex optimisation* is one of them (see section 16.1.2)—two more remain.

20.4.1. On-line and Offline Policies

The first meaning of 'on-line learning' is specific to reinforcement learning. Here, 'on-line learning' means to learn the best policy *including exploratory behaviour.* This understanding is applied, for example, by Shani (2007, pp. 35–44).

Consider, for example, an agent that moves alongside a cliff. An on-line policy makes sure that the agent keeps its distance such that it is safe despite random exploratory behaviour. An example for on-line policy learning is SARSA-learning. (see footnote 3 in Rummery and Niranjan 1994, p. 6; Sutton and Barto 1998, chapter 6.4)

'Off-line learning', in contrast, means to learn the best policy *without considering exploratory behaviour* (i.e. an *optimal* policy). Such a policy might lead the agent much closer alongside a cliff if this is in fact the shortest path from start to goal position. Off-line policies enable the optimal exploitation of deterministic environments. An example for off-line policy learning is Q-learning. (Watkins 1989; Sutton and Barto 1998, chapter 6.5)

In general, on-line policy learning appears cognitively more plausible. Cognitive systems do not appear to seize exploratory behaviour during their lifetime. Also it would be inefficient or plain dangerous for them not to consider the occasional unintentional behaviour.

However, this reading of 'on-line learning' is independent from sample availability and, therefore, not crucial to the learning from a stream of consciousness.

20.4.2. Sequential On-line Learning

The third meaning of 'on-line learning' is that the learning system is only allowed to pass over each sample *in sequence* and exactly *once.* This understanding of 'on-line learning' contrasts against *batch learning,* where a full set of unordered samples is randomly accessible to the system at all times. (Karp 1992)

Learning from a randomly accessible set of experiences, allows to use samples *more than once* to infer the most relevant properties over the complete data set. Reinforcement learning examples for batch learning are fitted Q-learning (Ernst et al. 2005), least-squares policy iteration (Lagoudakis and Parr 2003), or experience replay (Lin 1992).

However, if dynamics *change,* then learning from a previous set of experiences might change a policy for the worse. Storing experiences in a set also removes any potentially useful sequential information.

Whereas a randomly accessible set of complete samples presents a single and stationary distribution, samples that become available *over time* present a slightly different distribution with each time step. The sequence of samples up to any particular point in time might not be representative for the distribution of all samples.

Consider the fruit example from before: the goal is to find colour regions that enable to reproduce the grouping of fruits as it is presented in the training set.

In batch learning, the complete distribution of fruit colours and the total number of groups is known. From this knowledge, colour regions can be inferred such as to cover the training examples most effectively. This enables to minimise the training error.

In on-line learning, new examples become available only after each time step *but an output has to be generated even before that.* For the first input, the system must generate an output without *any* information. Its second output is based only on the first example, its third output is based only on the first and the second example, and so on.

In many cases, on-line learning is considerably harder than batch learning.[4] There are cases, however, where information can be inferred from the particular sequence of samples.

Consider, for example, the supervised learning of subtraction by one. If the examples are presented in an incremental sequence $\left[\ \langle t+1, t\rangle\ \right]_{t=0}^{\infty}$, the target is always identical to the last input. A system that considers the last input as a property of the current input can use the sequentiality of examples to its advantage. In batch learning, this advantage is void.

Temporal difference reinforcement learning uses the sequentiality of samples to predict the future discounted cumulative reward of actions. Accordingly, many reinforcement learning algorithms perform on-line learning according to *two* of the three presented understandings.

On-line is necessary in settings, where responses must be reactive and immediate. This

[4] The field of *competitive analysis* is concerned with the comparison of algorithmic off-line and on-line approaches to various problems. (see, for example, Borodin and El-Yaniv 2005; Buchbinder et al. 2012)

implies crucial complexity conditions for on-line algorithms.

Learning and prediction after each time step must be *on-line feasible* concerning runtime and memory complexity. In the following, 'on-line feasibility' is defined such that runtime and space complexity must grow less than the number n of examples (e.g. referents) received so far.

21. Learning the Model of a Dynamic System

The purpose of this chapter is to describe the generation of a semiotic model as well as the baseline approach against which it can be evaluated in the next chapter.

First, the particular problem of simulating the generation of a mental model is described from machine learning perspective. A solution to this problem has to satisfy two types of requirement.

The first type is *specific to the simulation of mental modelling.* These requirements must be satisfied in order to guarantee cognitive plausibility according to an intentional conception of mental models (see part III). The requirements are relevant for a phenomenologically justifiable grounding of symbolic representations in basic perception. However, they are not necessary to solve *the machine learning problem* of generating a semiotic model.

The second type of requirement applies to *any algorithm* that is supposed to generate a semiotic model from from sequential data in an on-line feasible manner. These formal requirements apply not only to cognitive models that are supposed to simulate cognitive processes. They are necessary for any comparable algorithm.

This second set of requirements enables to design a procedure for the general task. This task determines training and test samples for the algorithm as well as its measure of performance. It couples two dynamic trajectories that emit any type of shape and provides them to the learning algorithm.

The formalisation of the setting enables to compare approaches to the same formal problem. A baseline approach is determined one that satisfies most of the machine learning requirements from above. The simplicity and performance of this algorithm shows very clearly, where semiotic models excel. However, this baseline approach is not intended to simulate any cognitive process. This implies that it does not satisfy the first type of requirements.

Both algorithms are presented in detail at the end of this chapter in the form of pseudo code before their empiric evaluation in the next chapter.

21.1. Problem Formulation

From the perspective of the cognitive sciences, the computational generation of a semiotic model from a sensorimotor trajectory can be considered as *a simulation for the generation of mental models from first-person-perspective.* Therefore, algorithms that describe such a process can be understood as *cognitive models* for the process of mental modelling.

The computational generation of a semiotic model can also be considered as *the procedural acquisition of knowledge.* Therefore, it is subject to machine learning. Machine learning algorithms are relevant to this project, if they learn knowledge that enables them to solve a similar problem.

This problem is determined by one of the two requirements for pragmatic artificial intelligence in section 19.3: the ability to *recognise parts of its environment.* A system can be assumed to recognise parts of its environment if it is able to predict the sensor emissions that it receives.

More precisely, the generation of a mental model from first-person-perspective is the formal problem of *learning to predict sensor emissions in the dynamic trajectory of an environment that is coupled to motor emissions in the dynamic trajectory of the agent.* The approaches from embodied cognition—especially the concept of affordances and sensorimotor contingencies—substantiate this assumption (see part II). The part on intentionality and semiotics, on the other hand, describes a structure for semiotic models that enables to perform this task (see part III).

To a cognitive system, it appears as though its mental model *is* the world, can describe *everything* and, therefore, cannot be biased at all (see section 10.1). A closer look reveals, however, a specifically *phenomenological bias* in the way mental models present the world (see chapter 13 and section 15.5.1).

This translates into particular conditions for semiotic models and their generation that are not the case for comparable types of models. In the following, cognitive and computational interpretations for these characteristics are provided.

21.1.1. Requirements of Cognitive Plausibility

First, *the state of the model must be discrete.* In mental models, this state is the current belief of its cognitive system. Here, 'discrete' means that a representation either is part of this state or it is not, with no probabilistic gradient in between.

Second, *the model must be a partially ordered set.* Each layer in the hierarchy of a mental model is an individual level of abstraction. All representations at one level are equally abstract and can be part of more than one abstract representation from the level above. This follows from definition 6 on page 126.

Third, *transitions between representations in the model are probabilistic.* Each event has multiple potential outcomes and cognitive systems can make an educated guess on their probability based on observed frequency. This enables, for example, to handle sensor noise and motor uncertainty. This is covered by definition 4 on page 124.

The transitions between mental representations can be probabilistic even if the belief state is not. Belief state transitions to the next mental representation are dominated by the most probable successor, not probabilistically distributed over all possible successors. After each transition, the cognitive system is again in a single, determinate belief state.

Fourth, *the model must incorporate mistakes.* A mental model requires breakdowns because they enable to describe a complex structure of individually incompatible parts in the first place. Only disappointed expectations enable well-informed transitions into new contexts where the unexpected observation *would* have been expected. In section 12.6, this is described as 'breakdown'.

21.1.2. Machine Learning Requirements

First, *the model must have a state.* This state represents the hidden state of the environment which enables to describe a partially observable environment and to predict emissions that depend on its hidden states. This is covered by the state of semiotic models in definition 7 on page 126.

Second, *the model has to facilitate goal-directed action.* The model is not only a passive description of the environment but it is actively coupled to it. For this purpose, arbitrary action is as important as perception is. In a semiotic model, this is reflected in the fact that motor activation is an essential part of its referents (see section 16.3).

To facilitate goal-directed action, the system has to evaluate the states of the model. This evaluation, however, *is not part of the model.* Instead, it depends on the system's momentary goals, which are independent from the model and, more importantly, *change* over the course of time.

Third, *the model has to be goal-agnostic.* The model is only supposed to capture *causal dynamics* between the cognitive system and external reality. Therefore, models created during the solution of one task must also be able to contribute to solving a substantially *different* task in the same or a sufficiently similar environment. Translated into reinforcement learning terms, this means that the generated model must be *indifferent towards reward.*

For semiotic models, this is reflected by the fact that the agent's emissions are merely part of the transition condition (again, see section 16.3). The model describes the environment, depending on *whatever* action was performed by the agent. The semiotic model itself has no immediate influence on this action.(for the original idea of using goal-agnostic models to increase reinforcement learning performance, see Kuvayev and Sutton 1997)

Eventually, *the model must be on-line feasible* according to the definition in section 20.4.2. This means that it must be generated on-line, learn and predict after each time step, and that sample availability as well as memory and runtime complexity are heavily constrained.

21.1.3. Task

The general task is to generate a model that enables to predict emissions which depend on the hidden state of another system. The general approach is to represent this hidden state with the state of the model such that the relation between this representation and the emissions of the other system are determinate. Once determined, this relation enables to predict upcoming emissions.

The full iteration cycle for each step of the simulation is illustrated in algorithm 1.

Algorithm 1: Simulation Cycle and Evaluation.

```
parameter : Some arbitrary policy $\pi : \mathbb{N}^0 \to \mathsf{motor}$
input     : An interaction procedure interact : $e \times \mathsf{motor} \rightarrowtail E \times \mathsf{sensor}$[1]
input     : An initial state $e_0 \in E$ for the environment
function simulation(interact, $e_0$)
    loss ← ∅;
    model ← ∅;
    state ← ∅;
    for $t < T$ do
        $m_t \leftarrow \pi(t)$;                                     // action selection
        $e_{t+1}, s_{t+1}$ ← interact($e_t, m_t$);                  // interaction
        if $t == 0$ then loss($t$) ← 1;                             // evaluation
        else
            $\widehat{s}_{t+1}$ ← predict(model, state, $m_t$);     // test step
            if $\widehat{s}_{t+1} == s_{t+1}$ then loss($t$) ← 0;   // evaluation
            else loss($t$) ← 1;
            update(model, state, $m_t, s_{t+1}$);                  // training step
```

[1] In analogy to '→' as a short mathematical denotation for functions, '↣' is used for procedures. The reason is that repeatedly querying a procedure with the *same* argument can yield *different* output

The procedure `simulation` receives a probabilistic procedure `interact` that defines a dynamic environment and it receives the initial state e_0 of this system.

Lines 2 to 4 initialise the loss record, the model, and the state of the model. Line 5 initiates a loop for T iterations. Each iteration of the simulation is identified by a unique time stamp t and consists of the following instructions.

Line 6 instructs the arbitrary selection of motor emission m_t. In line 7, the current state of the environment e_t and this motor emission are provided to `interact` to receive the next state e_{t+1} and the next emission s_{t+1} of the environment.

If this environment is partially observable according to section 16.1.1, for example, then it is defined as $(E, \mathsf{motor}, T, \Omega, \mathsf{sensor})$. In this case, `interact` returns a random successor state e_{t+1} with the probability $p(e_{t+1}) = T(e_t, m_t, e_{t+1})$ and a random observation s_{t+1} with the probability $p(s_{t+1}) = \Omega(e_{t+1}, s_{t+1})$.

At time $t = 0$, prediction is completely uninformed. Therefore, line 8 describes the first loss as maximal. After the next time steps, line 10 instructs a prediction for s_{t+1} according to the current `model`, `state`, and motor emission m_t. Lines 11 and 12 update `loss` according to whether the prediction was correct or not.

Eventually, `model` and `state` are updated in line 13 according to the agent's current motor emission m_t and the environment's reaction in form of a sensor emission s_{t+1}.

The procedures `predict` and `update` are implemented according to the particular approach to the problem at hand. Before these implementations are described in detail, first, the loss over a complete simulation is defined and it is explained how this setting constrains the applicable types of machine learning.

21.1.4. Performance Measure

The practical condition for pragmatic artificial intelligence in section 19.3 requires that agents *recognise* aspects of their environment. An agent that recognises aspects of its environment is assumed also to be able to *predict* the environment's emissions.

The agent's success in predicting its next sensor activation depends on the *content* of its model. As determined in section 19.3, simulated intentional content must be considered inaccessible to third-person-perspective. The agent's predictive performance, however, enables an *indirect* evaluation of this content.

The processes that generate this model can be evaluated in virtue of predictive performance *over time* while the model and its state are continually adjusted to an erratic

values according to a particular probability distribution. In functions the argument *determines* the value.

environment. Better processes adapt the model such that it enables better predictions where the *average cumulative loss* is minimal at the end of the simulation.

The cumulative loss at time t is determined according to equation (21.1) as the sum of all previous losses in a particular simulation.

$$L(t) = \sum_{t'=0}^{t} \texttt{loss}(t') \tag{21.1}$$

The average cumulative loss at time t is determined according to equation (21.2).

$$\bar{L}(t) = \frac{1}{t+1} L(t) \tag{21.2}$$

The best off-line algorithm simply stores the whole trajectory and reproduces it in order to minimise $\bar{L}$. On-line learning under the condition of on-line feasibility, however, makes this approach impossible. Instead, the system needs to generate a *generalised* model of the trajectory.

21.2. Related Approaches

Several machine learning methods can be applied to this problem, but very few are empirically comparable. The problem is already with determining *the type* of the machine learning task at hand.

Modelling a *fully observable environment* means there is no hidden information. In these cases, learning to predict the next consequence shape from the current transition condition is trivial because the true distribution can be sampled directly and arbitrarily often.

Learning a model to predict the emissions of a fully observable environment is effectively a process of *supervised learning* (see section 20.2). Each transition condition is an input and each consequence shape is the according target. Simply keeping track of the number of transitions enables to reproduce the exact probability distribution from which they have been generated.

Modelling a *partially observable environment* is considerably harder. Here, transitions do *not* follow fixed probability distributions. Predicting the emissions of a partially observable environment with the model of a fully observable environment can only yield suboptimal results.

Such a model needs to adapt permanently to describe the ever changing relation between conditions and consequences. If the dynamics of the environment themselves do not change (i.e. if it remains erratic *in the same way)*, however, the structure of the model is supposed to converge at some point.

21.2.1. Supervised or Unsupervised

In a partially observable environment, there are *multiple* example relations between inputs and targets that depend on the current state of the environment (i.e. the emission function from definition 2 on page 121). This means the problem of prediction cannot be considered as *supervised* task. Modelling a partially observable environment first requires to separate these individual sets of examples.

However, to consider the modelling of a partially observable environment as *unsupervised* is problematic as well. There is quite obviously on-line training *as well as* test: in each instant, the model is adjusted according to the current emission of the environment and, *in the same instant,* the current prediction has a well-defined *target.*

From a unsupervised learning perspective, the question is: what is *the input* and according to what is it supposed to be *generalised?* In the case of a simulation for the generation of mental models, the inputs are *sensorimotor referents* that are supposed to be consolidated into the same content if they are *consistent and consecutive.*

The key question from a supervised learning perspective is: what is *the input* and what is *the target?* In the case of a simulation for the generation of mental models, the input is a tuple of *belief state and action,* and the target is *the next sensor emission.*

Therefore, to predict the emissions in a dynamic trajectory has to be a combination of two problems. On the one hand, it implies the unsupervised learning of a model state that represents the hidden states of the environment to separate referents into sets of functional examples. On the other hand, it implies the supervised learning of emissions (i.e. targets) that occur during each of these states (i.e. inputs).

21.2.2. Relations to Reinforcement Learning

Research in reinforcement learning often deals with the question how to increase learning efficiency in fully observable environments. A relatively active contemporary example is *hierarchical reinforcement learning.*

The goal in hierarchical reinforcement learning is the *decomposition* of a fully observable environment into several similar ones. This allows, for example, to transfer action evaluations that have been learnt for one of these segments onto other, supposedly similar, ones. As a consequence, the time to learn a policy for a Markov decision process in the whole environment can be considerably reduced. (Hengst 2010)

Learning a partially observable model is somewhat opposite to that. Hierarchical reinforcement learning aims to *generalise different* segments of the environment into *one* representation whereas a partially observable environment has to *differentiate identical* segments into *separate* representations.

This differentiation is necessary due to *perceptual aliasing.* This term describes a situation, where the current sensorimotor transition condition is not sufficient to predict the next sensor consequence—as it is the case during interaction with almost all real-world environments. Various reinforcement learning methods have been proposed to deal with perceptual aliasing.

Approaches like McCallum (1993), McCallum (1995), McCallum (1996b), Sun and Sessions (2000), and Crook and Hayes (2003) solve this problem by approximating a whole partially observable Markov decision process that includes *an evaluation* of the agent's actions. In case the agent's goals *change,* therefore, they need to be learnt *again.*

Due to its complexity, most approaches that learn a partially observable Markov decision process are not on-line feasible anyway (McCallum 1996a, pp. 10–11). The need to *repeat* the entire process as soon as goals change, makes clear that they are unfit for simulating cognitive processes.

Recurrent neural networks can use the activation in recurrent layers to represent hidden states of the environment. So far, however, they have mainly been used to approximate the discounted cumulative reward (see, for example, Lin and Mitchell 1993) and, therefore, they are not goal-agnostic as well. The same is true for the more recent implementation of the same idea with deep recurrent Q-networks (Hausknecht and Stone 2015).[2]

Also, recent successes in deep reinforcement learning often employ *experience replay memory* which stores past experiences and enables the agent to perform batch learning on them (Lin 1992). In a dynamic environment, however, the assumption of independent and identically distributed samples does not hold. Therefore, the reuse of past experiences solidifies assumptions about the environment that might no longer apply.

21.2.3. Order-n Markov Predictors

One model has proven quite useful for learning partially observable environments in reinforcement learning. *Order-n Markov predictors* represent the hidden world state with a fixed history of past emissions. The current history can be considered as *the state* of an order-n Markov predictor (see section 16.1.1).

[2] In principle, recurrent networks should be able to maintain an internal representation for the hidden state of the environment that enables to predict the next sensor emission. To develop such a method, however, more research in this direction is necessary.

These models are mostly used to predict the discounted cumulative reward of actions during hidden states of the environment. However, they can be easily adapted to predict the next emission for each of their states.

Order-n Markov predictors are quite popular to describe partially observable environments. They are generated by simply counting the number of transitions from the current history of length n to the current emission. The normalised transition counts from one history to all of its emissions realises a probability distribution that can approximate the actual distribution very closely.

The previously most frequent successor of the current history provides an informed estimate on the most probable emission after the next time step. The result is a model according to definition 4 on page 124, where $C \subseteq S^n$.

Their prediction performance tends to increase with an increase in history length up to a certain point. Once this point is reached, performance decreases steadily. Transition conditions become too specific to generalise the distribution of emissions appropriately while, at the same time, loosing long-term dependencies over more than n time steps.

This problem can be partially remedied by approaches with *variable* histories. These approaches adapt the lengths of histories in order to find an optimal trade-off between *the number of states to keep track of* and their *general applicability* (extensively investigated in McCallum 1996b). However, this adaptability is paid for with on-line feasibility.

Order-n Markov predictors contradict some of our requirements for cognitive plausibility. They are not partially ordered sets and they do not incorporate prediction mistakes. However, they also fulfil some cognitive conditions. They are in a discrete state and they describe transitions probabilistically.

Most importantly, however, they are *the only* approach that satisfies all of the machine learning conditions. They feature a state in the form of the current history, they are on-line feasible, they are goal-agnostic, and they enable goal-directed interaction. Therefore, order-n Markov predictors serve as a perfect baseline approach for an empiric comparison with automatically generated semiotic models.

22. The Procedures

Any non-trivial model for the prediction of a dynamic trajectory needs to represent *the structure* and *the state* of the emitting system. Both representation can be of various types, depending on the particular kind of model.

The primary structure of an order-n Markov predictor is a *matrix,* whereas the primary structure of a semiotic model is a *list.* The state of an order-n Markov predictor is *sequential,* whereas the state of a semiotic model is *hierarchical.*

Candidate procedures that instruct learning from dynamically coupled trajectories need to comply with the sensorimotor interface from section 16.3. This means that they need to implement `predict` in line 10 and `update` in line 13 of algorithm 1 on page 155.

In the following, procedures are provided that define these structures as well as their development over the course of time and with each new referent.

22.1. Baseline Procedure Prediction

In order-n Markov predictors, the procedure `predict` accords to algorithm 2. The structure of the model is a transition table that counts frequencies according to definition 4 on page 124, where $C \subseteq S^n$.

The state of this model is the current history of emissions as described it in section 16.1.3. Together, the state and the current motor emission m_t enable to infer the previously most frequent consequence shape.

The procedure `predict` receives an order-n Markov predictor, its state, and the current motor emission m_t. Line 2 describes the transition condition as tuple that consists of the current state of the model (i.e. the current history) and the current motor emission.

If this tuple is *not* a transition condition in the model, then the model cannot provide a prediction. If it *is* a transition condition, the procedure returns the consequence shape that maximises the transition count according to `model`.

Algorithm 2: Prediction with an Order-n Markov Predictor.

```
input   : A model : $\textsf{sensor}^n \times \textsf{motor} \times \textsf{sensor} \to \mathbb{N}^*$, where $n \in \mathbb{N}^+$
input   : A state : $\{ 0, 1, ..., n-1 \} \to \textsf{sensor}$
input   : The current motor emission $m_t \in \textsf{motor}$
output  : The expected sensor emission $\widehat{s}_{t+1} \in \textsf{sensor}$
function predict(model, state, $m_t$)
    condition ← $\langle$state, $m_t\rangle$;
    if condition $\notin \{ c \mid \langle c, s, f \rangle \in$ model $\}$ then $\widehat{s}_{t+1} \leftarrow \varnothing$;
    else $\widehat{s}_{t+1} \leftarrow \operatorname{arg\,max}_{s' \in S}$ model(condition, $s'$);
    return $\widehat{s}_{t+1}$
```

22.2. Baseline Procedure Update

Algorithm 3 describes the procedure `update`. It consists of two separate parts. The first part, `updateModel`, concerns the update of the model's structure. The second part, `updateBelief`, concerns the update of the model's state.

Algorithm 3: Updating an Order-n Markov Predictor and its State.

```
input   : A model : $\textsf{sensor}^n \times \textsf{motor} \times \textsf{sensor} \to \mathbb{N}^*$, where $n \in \mathbb{N}^+$
input   : A state : $\{ 0, 1, ..., n-1 \} \to \textsf{sensor}$
input   : The last motor activation $m_{t-1} \in \textsf{motor}$
input   : The current sensor activation $s_t \in \textsf{sensor}$
1 function update(model, state, $m_{t-1}$, $s_t$)
2     condition ← $\langle$state, $m_{t-1}\rangle$;
3     updateStructure(condition, $s_t$, model);
4     updateState(state, $s_t$);
```

The procedure `updateModel` receives the structure of the model that is supposed to be updated, its state, the last motor emission m_{t-1}, and the current sensor emission s_t. Line 2 describes the generation of a transition condition from `state` and m_{t-1}.

This transition condition, the current sensor emission, and the structure of the model are supplied to `updateStructure`, where the structure is updated accordingly. Eventually, the state of the model is updated with the current sensor emission.

Algorithm 4 presents `updateStructure` in detail. The procedure receives a transition condition (i.e. history), a consequence shape (i.e. sensor emission), and the structure of an order-n Markov model. Line 2 shows the composition of a referent as a tuple that consists of this history and the sensor emission. This referent is integrated into the structure of

Algorithm 4: Updating the Structure of an Order-n Markov Predictor.

```
input     : A transition condition $= \langle \texttt{state}, m_{t-1} \rangle \in C$
input     : The current sensor activation $s_t \in$ sensor
input     : A model : sensor$^n$ × motor × sensor $\rightarrow \mathbb{N}^*$, where $n \in \mathbb{N}^+$
function updateStructure(condition, $s_t$, model)
    referent $\leftarrow \langle$condition, $s_t\rangle$;
    if referent not in domain of model then model(referent) $\leftarrow 1$;
    else model(referent) $\leftarrow$ model(referent) $+ 1$;
```

the model.

The procedure `updateBelief` in algorithm 5, eventually, is responsible for updating the system's internal representation for the hidden state of the environment. In an order-n Markov predictor, this representation is the sequence of the n last observations.

Algorithm 5: Updating the State of an Order-n Markov Predictor.

```
parameter : The history length $n \in \mathbb{N}^+$
input     : A sequential state : $\{ 0, 1, ..., n-1 \} \rightarrow S$
input     : The current sensor emission $s_t \in$ sensor
function updateState(state, $s_t$)
    for $i \in \{1, n-2\} \subseteq \mathbb{N}^+$ do
        if $i$ not in domain of state then state($i-1$) $\leftarrow \varnothing$;
        else state($i-1$) $\leftarrow$ state($i$);
    state($n-1$) $\leftarrow s_t$;
```

The procedure receives the current state of the model and the current sensor emission. It shifts all shapes in the history one index to the left. Empty indices are determined as the empty set by default. Afterwards, the current sensor emission is added to the end of the state.

Algorithms 2 to 5 define a baseline approach that generates an order-n Markov predictor which enables to predict the emissions in a dynamic trajectory. This approach complies to our simulation framework in algorithm 1 on page 155 and, therefore, it can be compared to the generation of a semiotic model.

22.3. Semiotic Model Prediction

In the following sections, procedures for the generation of a semiotic model are described. The proceeding is analogous to the presentation of the baseline approach in the previous section.

According to definition 6 on page 126, each level of a semiotic model consists of a set of indices I_l. In a computational implementation, each bijective functional relation $I_l \subseteq$ `feeling` can be realised quite naturally as *computational reference.* Computational reference is a bijective functional relation between computational identifiers and data structures in virtue of their memory location.

Algorithm 6 describes `predict` for semiotic models. The procedure receives the structure `model` of a semiotic model, its `state`, and the current motor emission m_t. It returns a prediction $\widehat{s}_{t+1}$ for the next sensor emission.

Algorithm 6: Prediction with a Semiotic Model.

```
  input    : A model according to definition 6 on page 126
  input    : A state according to definition 7 on page 126
  input    : The current motor emission $m_t \in \mathsf{motor}$
  output   : The expected sensor emission $\widehat{s}_{t+1} \in \mathsf{sensor}$
1 function predict(model, state, $m_t$)
2     if 0 not in domain of state then $\widehat{s}_{t+1} \leftarrow \varnothing$;
3     else if 1 not in domain of state then $\widehat{s}_{t+1} \leftarrow$ state(0);
4     else
5         content ← state(1);
6         condition ← $\langle$state(0), $m_t\rangle$;
7         if condition $\notin \{\ c \mid \langle c, s, f\rangle \in$ content $\}$ then $\widehat{s}_{t+1} \leftarrow$ state(0);
8         else $\widehat{s}_{t+1} \leftarrow \arg\max_{s \in S}$ content(condition, $s$);
9     return $\widehat{s}_{t+1}$
```

The procedure covers three cases. The first case in line 2 describes that no prediction can be performed if the state is empty. Line 3 applies if the state has a base shape but no shape for structural content at level $l = 1$. In this case, the current base shape is returned.

Line 4 covers all other cases (i.e. when the state of the model has at least two levels). First, the current content at state level $l = 1$ is selected. If the tuple of base shape and motor emission is *not* a transition condition in this content, then the base shape at $l = 0$ is returned. If it *is* a transition condition, then the consequence shape according to `content` is returned.

22.4. Semiotic Model Update

In contrast to order-n Markov predictors, structure and state of a semiotic model are updated *simultaneously.* The update takes place in virtue of the procedure `generateModel`. Algorithm 7 shows the initial call of this procedure.

Algorithm 7: Updating a Semiotic Model and its State.

```
input    : A model according to definition 6 on page 126
input    : A state according to definition 7 on page 126
input    : The last motor emission $m_{t-1} \in \mathsf{motor}$
input    : The current sensor emission $s_t \in \mathsf{sensor}$
function update(model, state, $m_{t-1}$, $s_t$)
    generateModel($0$, model, state, $m_{t-1}$, $s_t$);
```

The core procedure `generateModel` receives an additional input 0 that determines which level of the model is to be modified. The procedure is called recursively and for each recursion, this value is incremented by one.

Algorithm 8 describes the procedure `generateModel` in detail. It receives a level, the structure and the state of a semiotic model, an action, and a shape.

If the provided level is beyond the state of the model, the procedure stores the received shape as new base level in the model structure. It updates the belief state to this shape and then it returns.

If the level is within the state, the procedure adds the shape to the current level. Next, in line 4, a condition is composed of the last shape at this level of the state and the received action.

If $\mathsf{level} + 1$ is *not* within the state of the model, a new content is generated, the state at this level is set to the shape of this content, and a new level is introduced that contains only this content. In line 21, this content is adapted to the transition from condition to shape, the state is updated in line 23, and the procedure returns.

If $\mathsf{level} + 1$ *is* within the state of the model, the current content at this level of the state is selected. If this content *does* present a referent that consists of the composed transition from line 4 and the received shape (i.e. if the probability is above σ), then the content selected in line 6 is adapted to this referent in line 21, the current state at level is updated to the received shape in line 23, and the procedure returns.[1]

If this content does *not* present this referent, then the state is queried for $\mathsf{level} + 2$. If

[1] The procedure `certainty` follows immediately the probability of a fact in definition 4 on page 124. The procedure `adapt` is defined in algorithm 10.

Algorithm 8: The Recursive Generation of a Semiotic Model.

```
parameter : The certainty threshold $\sigma \in [0,1] \subseteq \mathbb{R}$
input     : An integer level $\in \mathbb{N}^*$
input     : A model according to definition 6 on page 126
input     : A state according to definition 7 on page 126
input     : The system's own emission $a_t \in S_{\mathsf{level}}$ according to definition 8 on
            page 126
input     : The other system's emission $s_t \in S_{\mathsf{level}}$
function generateModel(level, model, state, $a_t$, $s_t$)
    if level in domain of state then
        add $s_t$ to model(level);
        condition ← $\langle$state(level), $a_t\rangle$;
        if level + 1 in domain of state then
            content ← state(level + 1);
            if certainty(content, condition, $s_t$) < $\sigma$ then
                if level + 2 in domain of state then
                    context ← state(level + 2);
                    abstractCondition ← $\langle$content, condition$\rangle$;
                    content ← predict(context, abstractCondition);
                    if certainty(content, condition, $s_t$) < $\sigma$ then
                        content ← $\underset{c \in \mathtt{model}(\mathsf{level}+1)}{\arg\max}$ certainty($c$, condition, $s_t$);
                        if certainty(content, condition, $s_t$) < $\sigma$ then content ← $\varnothing$ ;
                else content ← $\varnothing$ ;
                generateModel(level + 1, model, state, condition, content);
        else
            content ← $\varnothing$;
            state(level + 1) ← content;
            model(level + 1) ← { content };
        adapt(content, condition, $s_t$);
    else if level == 0 then model(0) ← { $\langle s_t$, indices(level)($s_t$)$\rangle$ } ;
    state(level) ← $s_t$;
```

there is *no* such level, a new content is created in line 15 and the procedure is called recursively with an incremented level, the model and its state, the condition from line 4 as an abstract action, and the shape of the newly generated content.

If there *is* a $\mathsf{level} + 2$ in the state of the model, the content of the state at this level is selected as context. The current content at $\mathsf{level} + 1$ and the condition from line 4 are selected as an abstract transition condition. The context from $\mathsf{level} + 2$ is queried for a prediction of the successor for `content` from line 6.

If the predicted candidate content *does* present the referent that consists of the composed transition from line 4 and the received shape, then the procedure continues at line 16.

If the candidate does *not* present this referent, the best content among all at $\mathsf{level} + 1$ is retrieved. If the best candidate *does* present the referent, the procedure continues in line 16, otherwise a new content is generated.

The core procedure adapts and maintains the semiotic model and its state. Effectively, it tries to find *a good representation* for the given referent *in the current context.* To achieve this, it exploits hierarchically repetitive structures in the stream of incoming information. If this succeeds, its state enables to represent aspects of the hidden state of a dynamic system.

22.5. Semiotic Model Auxiliary Procedures

The procedure `certainty` determines the certainty of a fact according to definition 4 on page 124. It is defined in algorithm 9.

The procedure receives a stochastic model function, a transition condition, and a consequence shape. The condition and consequence define a referent. If this referent is *not* in the domain of the stochastic model function, the procedure returns a certainty of 1.

If the referent *is* in the domain of the model function, the procedure counts the transition frequencies from the received condition to all consequences, modified by a pseudocount α. It divides the frequency of the received consequence shape—also modified by α—by this total amount and returns the result as the probability that the given referent is a fact according to the current context.

Algorithm 10 describes the adaptation of content to a referent. The procedure `adapt` receives a content, a transition condition, and a consequence shape. It composes condition and consequence into a referent and increments the transition count for this referent by one.

Algorithm 9: Determining the Certainty of a Fact.

```
parameter : The pseudocount $\alpha \in [0, 1] \subseteq \mathbb{R}$
input     : A content : $c_l \times S_l \rightarrow \mathbb{N}^*$
input     : A transition condition $c \in C_l$
input     : A consequence shape $s \in S_{\mathsf{level}}$
output    : The probability of content presenting $\langle c, s\rangle \in$ referents
function certainty(content, $c, s$)
    if $\langle c, s\rangle$ in domain of content then
        totalFrequency $\leftarrow 0$;
        for $f' \in \{ \; f \mid \langle c, s, f\rangle \in$ content $\}$ do
            totalFrequency $\leftarrow$ totalFrequency $+ f' + \alpha$ ;
        frequency $\leftarrow$ content($c, s$) $+ \alpha$;
        probability $\leftarrow \frac{\text{frequency}}{\text{totalFrequency}}$;
    else probability $\leftarrow 1$;
    return probability
```

Algorithm 10: Adapting Content to a Referent.

```
input     : A content : $c_l \times M_l \rightarrow \mathbb{N}^*$
input     : A transition condition $c \in C_l$
input     : A consequence shape $s \in S_l$
function adapt(content, $c, s$)
    referent $\leftarrow \langle c, s\rangle$;
    if referent in domain of content then
        content(referent) $\leftarrow$ content(referent) $+ 1$ ;
    else content(referent) $\leftarrow 1$;
```

23. Conclusion

Prior to their evaluation, consider the relevance of both algorithms for the three major problems in artificial intelligence from chapter 2. The frame problem and the problem of vanishing intersections are both merely different manifestations of the symbol grounding problem. A solution to the symbol grounding problem, therefore, should remedy those problems as well.

To define categories as a conjunction of features implies a prior *selection* of these features (see, for example, Harnad's search for 'invariants' in section 2.2). If the relation from features to categories is non-functional, however, then more discriminate features must be *generated.* Feature selection and generation are active contemporary fields in machine learning (for an overview, see Saitta and Zucker 2013, pp. 277–293).

The reduction of a category to only some of its members' features is *a selective approach to abstraction.* An alternative is *a constructivist approach to abstraction.* The structural features of the members of one category can all be gathered into a disjunctive description for this category. Constructive abstraction follows up on MacDorman's proposal of disjunctive category definitions from section 2.2.1.

According to Wittgenstein's critique from the same section, however, a disjunctive category *by itself* is merely an enumeration of its members. It cannot describe the initial generation of this category because these members share nothing but being contained in the same category.

Why should a new *disjunctive* category be generated at all? New referents could simply be integrated into already existing ones.

Why should a referent be contained in one category but not in another? According to Dreyfus, the fact that we experience ourselves always already in some situation provides us with the mental categories that we assume to be relevant for the current referents. This context-sensitivity enables the generation of disjunctive, but still discriminate, mental representations because *it favours one category over another.*

The content of abstract mental representations serves as such a *contextual frame.* The generation of a new representation is always preceded by the retrieval of the most appropriate available representation. A new representation is introduced *only* if even the most appropriate representation that is available is not appropriate enough.

In the selection of this representation, the system implements a particular expectation bias. Mitchell defines biases as "*any basis for choosing one generalization over another, other than strict consistency with the observed training instances*" (Mitchell 1990). `expectation` according to section 16.2.1 provides such a basis. It enables to prioritise one among many representations that are structurally consistent with the current referent.

Algorithm 8 shows that the recognition of referents as representations is primarily biased towards the last representation. If this fails, then it is biased towards the representation that has been previously experienced to follow immediately after the last representation according to the current context. If this fails as well, all representations have to be considered to find an appropriate one.

With a prioritisation like this, a disjunctive accumulation of features *can* be discriminate. A new representation will only be generated if none can be found that presents the referent well enough.

After retrieving an appropriate representation, the model is adapted to the referent in two ways. On the one hand, the model is adapted to better describe *the structure* of the referent (i.e. content). On the other hand, the model is adapted to better describe *the conditions* for its representation (i.e. the context).

This adaptation corresponds to Mountcastle's *equipotential* function at the foundation of cognition (see section 19.2).

Part VI.

Evaluating the Simulation

24. Introduction

The intentional content of mental representations cannot be described. What *can* be described, is a process that *generates* structures with intentional content. The description of a process is a procedure and a computational procedure is an algorithm.

Algorithm 8 in the previous part is designed to describe the generation of a mental model according to phenomenological theories of intentionality. These theories are based in empirical data from introspection.

The algorithm is considered a contribution *to the cognitive sciences* if its computational instances generate structures that correspond to this data. On top of that, the algorithm is considered a contribution *to machine learning* if it also enables to solve a particular problem as good or even better than comparable approaches.

The generated model can be evaluated in each of the three types of machine learning. From these three perspectives on the same procedure follows the need for three different kinds of evaluation.

In *supervised learning* and *reinforcement learning,* the algorithms can be compared quite easily to similar approaches. In supervised learning, the performance measure can be determined as the difference between output and target. In reinforcement learning, the performance is determined as the discounted cumulative reward received over time

The performance in *unsupervised learning,* however, depends on a particular purpose. In our case, this purpose is to simulate the generation of mental models according to phenomenological theories on the mind.

24.1. Unsupervised Simulation of Mental Modelling

First and foremost, the learning of a semiotic model is *unsupervised* because the system receives an unlabelled, temporally discrete sequence of qualitatively distinguishable elements, one at a time. The model identifies reoccurring parts of this sequence like mental models identify objects in their cognitive system's stream of consciousness. In this respect, the procedure resembles algorithms for *sequence clustering* (Ye 2004, pp. 277 sqq.).

The goal of an unsupervised machine learning system "is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc." (Ghahramani 2004, p. 73)

From this follows that the performance of an unsupervised learning system depends on *what the system is used for.* Algorithms can only be evaluated in a particular *application* (Saitta and Zucker 2013, p. 277). In the next section, prediction is presented as this application.

An appropriate application for the generation of a semiotic model in general is any task that can be solved faster or better with the model of a dynamic system which can be generated and adapted on-line.

At its core, however, algorithm 8 is meant as a simulation for the generation of mental models. This concerns evaluation in two points. Firstly, it puts a focus on the *temporal development* of the model—more specifically, the generation and adaptation of new representations over time. Secondly, it puts a focus on the *consistency* of the generated representations—more specifically the relation between internal representations and external referents. The following evaluation as an unsupervised learning procedure takes both of these points into account.

24.2. Supervised Sequence Prediction

One application for the learning of a semiotic model is *sequence prediction.* This process can be considered as *supervised* because, after each time step, the system has an input (i.e. the transition condition) and a target (i.e. the consequence shape) that it was supposed to predict.

A natural task for such systems is to predict the emissions of a *partially observable environment* (see section 16.1.1). Prediction with incomplete perception requires a stateful model that is able to differentiate emissions that *appear* identical but follow different causal dynamics.

The state of a model is supposed to represent parts of a trajectory where the probability of each emission is determined. The state of order-n Markov predictors is the history of previous emissions. The state of a semiotic model, in contrast, is a context that describes the *situation* (i.e. segments of the trajectory) where this probability is determined.

Think of a card game: the complete state of the game goes far beyond the observable cards in your hand. However, your current belief about this state is generated only from your actual observations throughout the game.

Another illustrative example is orientation in a partially observable grid world. Here, the unsupervised task of the agent is to generate a representation for the hidden state of the environment (i.e. the agent's absolute position). The supervised task is to use this representation to predict the next sensor emission that it will receive from the environment.

Due to the unobservability of phenomenal content, an observer cannot simply determine the match between representations which have been generated from this content and their referents in the agent's simulated environment. What *can* be determined, however, is whether these representations enables the agent to *recognise* their referents.

24.3. Reinforcement Learning with Hidden State Representations

Usually, reinforcement learning agents receive a 3-tuple after each time step t that consist of their last motor emission m_{t-1}, the current sensor emission s_t, and a real-valued reward for their prior actions r_t. This is enough to determine optimal policies in fully observable environments. Accordingly, various research improves the *efficiency* of algorithms that infer such an optimal policy (i.e. decrease the time it takes to find an optimal policy).

The common understanding of 'optimal' behaviour in reinforcement learning submits to perceptual aliasing. 'Optimality' effectively means 'the most effective behaviour without any way to distinguish apparently identical sensor data'.

In a partially observable environment, however, *identical* sensor emissions might require *different* motor activations to achieve a particular goal-state. This one of the symptoms of perceptual aliasing.

Consider a path that forks at two different locations. To reach its goal, the agent has to go left at the first fork but right at the second. Although both forks might appear identical, the agent needs to learn to differentiate them in order to reach its goal.

Only little research exists on on-line methods that increase *the eventual effectiveness* of policies in partially observable environments (i.e. exceeding the performance of *optimal* policies according to the above interpretation) beyond history-based approaches like the order-n Markov model.

Cognitive systems are able to resolve perceptual aliasing. Their mental models provide them with criteria that enable to differentiate identical sensor activations into separate representations. A good simulation for the generation of mental models, therefore, should enable an agent to learn a 'better-than-optimal' policy as well. (Crook 2007, pp. 101–102)

Models that enable to predict the emissions of a dynamic environment by representing its hidden state can also use this representation to better *navigate* the environment during goal-directed behaviour. Successful prediction is a sufficient indicator for effective recognition.

Lonnie Chrisman brings the benefit of model states under perceptual aliasing to the point.

> Interestingly, the ability to predict is not the characteristic that makes predictive models useful for overcoming perceptual aliasing. Instead, it is *the internal state* that is formed and utilized to make predictions which is valuable to the reinforcement learner. The central idea behind the current approach is that the information needed to maximize predictiveness is usually the same information missing from perceptually aliased inputs. (Chrisman 1992, p. 184, emphasis added)

The state of a semiotic model contains part of the information mentioned by Chrisman. To show this, the experience of a *reinforcement learning* agent is augmented with the current state of a semiotic model of the environment.

If the agent accumulates more reward over time *with* this state than it does *without,* then the state must represent some information about the environment that is actually hidden to the agent.

Because the evaluation setting is spatial, tasks are differentiated by referring to the unsupervised case as ‘localisation’, to the supervised case as ‘prediction’, and to the reinforcement learning case as ‘navigation’. However, these tasks cannot only be applied to spatial domains.

24.4. Proceeding

The evaluation concerns two types of data: an *uncoupled trajectory* (i.e. linear data) and a *coupled trajectory* (i.e. interactive data). The linear data is a *sequence of characters.* The interactive data is a *sequence of sensorimotor information* that a simulated agent exchanges with a discrete and partially observable environment (i.e. a grid world). In the reinforcement learning task, sensorimotor information is extended by a *reward* that enables to infer goal-directed behaviour.

The linear data is separated into an *artificial sequence* and a *natural sequence.* The artificial sequence is constructed with the intention to highlight the specificities and advantages of semiotic models. The natural sequence is chosen independent from the particular procedure and, therefore, more realistic.

Each evaluation consists of three general parts. The evaluation as *an unsupervised model generator* investigates the generated data structure as a model for the sequence

	unsupervised		comparative evaluation	
trajectory	development	representation	*supervised*	*reinforcement*
uncoupled artificial	Section 25.1.1	Section 25.1.2	Section 25.2	-
uncoupled natural	Section 26.1.1	Section 26.1.2	Section 26.2	-
coupled	Section 28.1.1	Section 28.1.2	Section 28.1	Section 28.3

Table 24.1.: The Different Evaluation Settings in their According Section.

and as a simulation for mental representations. The evaluation as *a supervised sequence predictor* provides an objective measure of performance in on-line sequence prediction. Lastly, the evaluation as *reinforcement learning under perceptual aliasing* simulates how the generation of semiotic models facilitates goal-directed behaviour in partially observable settings.

The unsupervised evaluation is exclusively qualitative. It presents *the temporal development* of the generated model and *the relation between representations and referents.* The supervised and reinforcement learning evaluations are comparative as well. They examine performance in comparison to a baseline approach. Table 24.1 provides an overview.

25. Modelling an Artificial Sequence

The evaluation with a linear dataset shows that semiotic models enable to predict the next emission in an *uncoupled* dynamic trajectory. Similar to order-n Markov predictors, semiotic models can also be adjusted to consider more than only one previous emission. This chapter only considers trajectories according to definition 2 on page 121 with a history of length of 1. The evaluation with a natural sequence in the next chapter considers longer histories as well.

The uncoupled dynamic trajectory is an infinite sequence of emissions from one of two alternating emission functions $f_a, f_b \in \mathcal{F}$, such that $\forall f \in \mathcal{F}.f : [0, 9] \rightarrow [0, 9]$.

One function emits ascending digits $f_a(x) = (x + 1) \texttt{ mod } 10$ and the other emits descending digits $f_b(x) = (x - 1) \texttt{ mod } 10$. After each time step t, a new digit is sampled from one of both functions and depending on the emission at $t - 1$.

After each time step, there is also a chance of $p = 0.2$ that the emission function from which elements are sampled *changes.* This random change is analogous to the randomly changing behaviour of the erratic centrifugal governor in chapter 17.

An example part of the trajectory is illustrated in figure 25.1. Each transition from one emission to the next either increments or decrements this element by exactly one. The source code used to generate this data can be found in appendix D on page 221.

25.1. Unsupervised Model Generation

The qualitative evaluation investigates the generated semiotic model for similarities with mental models. Among the relevant properties are 1) a number of new representations that is *asymptotic* over time, 2) the representation of *similar* parts of the environment with *the same* representation, and 3) a *hierarchical* organisation of these representations (i.e. in a partially ordered set).

Cognitive systems realise these properties reactively and in an *on-line* manner. This

```
...101234321098765456765432109876543210901232109878901234...
```

Figure 25.1.: Part of an Uncoupled Dynamic Trajectory.

translates into two particular requirements for an according machine learning algorithm.

The first requirement is that incoming data has to be processed in *a single pass.* This means that samples can only be used once to adjust the model. As soon as the system has adapted to the sample, *this sample is discarded.*

This premise does not only follow from how cognitive systems process information. It rather pays tribute to the fact that information from the environment cannot be assumed to be from a single source. Considering *past* samples in *current* adjustments to the model might change the model towards describing circumstances which are no longer the case.

The second requirement follows from this: data cannot be divided into periods of training and periods of test. Test data might be sampled from a radically different probability distribution than training data. An obvious way to deal with this fundamental uncertainty is to adapt the model continuously and on-the-fly (see section 21.1.3.

25.1.1. Structural Model Development

The table 25.1 shows an order-1 Markov predictor that has been generated according to these conditions. The random changes between f_a and f_b cause a strong uncertainty concerning potential successor emissions. This uncertainty expresses in a close-to-uniform distribution from each emission to its successors.

Once all relevant information has been captured, the structure of the model remains constant: the model has *converged* onto the current environment. The most frequent successors according to an order-1 Markov model of the example trajectory, however, alternate indefinitely.

Tables 25.2 to 25.4 show the normalised transition frequencies in the representations at level $l = 1$ of a semiotic model of the trajectory. Table 25.5 shows the normalised transition frequencies between *those* representations in the abstract representation at level $l = 2$.[1]

The model contains only point distributions which reflects a high degree of certainty concerning the transition of emissions. Accordingly, figure 25.2 shows that the semiotic model converges after 24 time steps. Level 0 contains all emissions, level 1 contains all representations at level $l = 1$, and level 2 contains a single representation.

Already during development, the model separates segments of the trajectory where two subsequent emissions are in a functional relation. However, three representations have been generated where two would have sufficed (i.e. one for ascending and one for descending digits). The reason for the additional representation is that representation update is *lazy.*

[1] Unless described differently, in the following, $\alpha = 0.0$ and $\sigma = 1.0$ for the generation of semiotic models (see algorithms 8 and 9 on page 166 and on page 168).

	consequence (probability in %)									
condition	‘0’	‘1’	‘2’	‘3’	‘4’	‘5’	‘6’	‘7’	‘8’	‘9’
‘0’	0	52	0	0	0	0	0	0	0	48
‘1’	53	0	57	0	0	0	0	0	0	0
‘2’	0	44	0	56	0	0	0	0	0	0
‘3’	0	0	53	0	47	0	0	0	0	0
‘4’	0	0	0	54	0	46	0	0	0	0
‘5’	0	0	0	0	48	0	52	0	0	0
‘6’	0	0	0	0	0	50	0	50	0	0
‘7’	0	0	0	0	0	0	48	0	52	0
‘8’	0	0	0	0	0	0	0	51	0	49
‘9’	52	0	0	0	0	0	0	0	48	0

Table 25.1.: Order-1 Markov Model of an Uncoupled Dynamic Trajectory.

Referents are integrated into the current representation as long as it represents them. Only if there is a significant difference (i.e. less than σ), the representation changes.

One one hand, this enables redundancy and high error tolerance. On the other hand, it can also lead to cases where *one* referent is contained in *all* representations. If a newly received referent contradicts this omnipresent referent, a new representation needs to be introduced. This is the case here.

25.1.2. Representation Associations

To predict the emissions in a dynamic trajectory requires to recognise emission functions for segments in this trajectory. Figure 25.3 shows part of the dynamic trajectory in the time interval $900 \leq t < 1000$. Segments in this part are coloured according to the representation at level $l = 1$ of the semiotic model that they have been recognised as.

The average length of these segments is $\bar{l} \approx 2.27$. This means that, on average, every 2.27-th time step, the model has to correct for an unexpected observation. As the generating function changes only every 0.2 time steps on average, these corrections appear relatively frequent.

However, it also means that, for 2.27 time steps in a row, the model is correct. As a consequence, the model's approximate loss $\widehat{L}$ can be inferred immediately from the average segment length: $\widehat{L} = \frac{1}{\bar{l}} \approx \frac{1}{2.27} = 0.44$.

This would be slightly better than a order-1 Markov predictor with as estimated loss of $\widehat{l_b} \approx 0.5$—in which each emission has two *different* successors with the *same* probability. In the next section, this estimate is tested empirically.

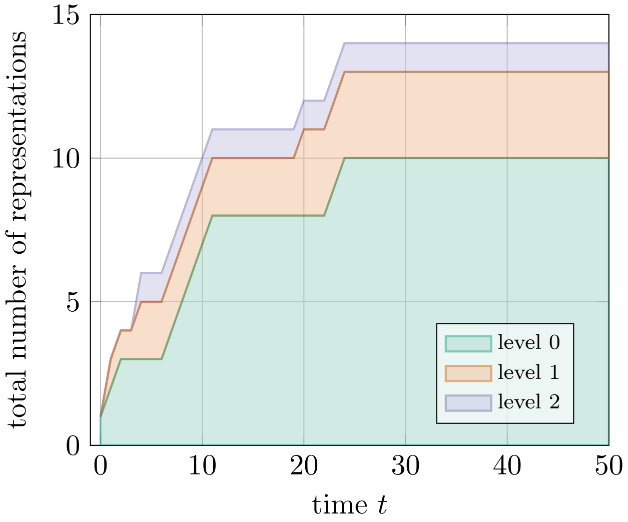


Figure 25.2.: Representation Generation in Semiotic Model.

1-0	consequence (%)							
condition	'0'	'1'	'2'	'3'	'4'	'5'	'8'	'9'
'0'	0	100	0	0	0	0	0	0
'1'	0	0	100	0	0	0	0	0
'2'	0	100	0	0	0	0	0	0
'3'	0	0	0	0	100	0	0	0
'4'	0	0	0	100	0	0	0	0
'5'	0	0	0	0	100	0	0	0
'6'	0	0	0	0	0	100	0	0
'7'	0	0	0	0	0	0	100	0
'8'	0	0	0	0	0	0	0	100
'9'	100	0	0	0	0	0	0	0

Table 25.2.: Representation 0 from Level 1.

4 3 210 121 01 2 34 3 2109 01 0 9 012 3 2 3456767 8 76 5 67 8 76 5 6767 890121 0 987 890
98 9 876 54 5 43 2 34 3 21 21 0 987654 567 89012 34567

Figure 25.3.: Represented Segments at Level 1.

1-1	consequence (%)							
condition	‘0’	‘1’	‘3’	‘4’	‘5’	‘6’	‘7’	‘8’
‘0’	0	100	0	0	0	0	0	0
‘1’	100	0	0	0	0	0	0	0
‘2’	0	0	100	0	0	0	0	0
‘3’	0	0	0	100	0	0	0	0
‘4’	0	0	0	0	100	0	0	0
‘5’	0	0	0	0	0	100	0	0
‘6’	0	0	0	0	0	0	100	0
‘7’	0	0	0	0	0	100	0	0
‘8’	0	0	0	0	0	0	100	0
‘9’	0	0	0	0	0	0	0	100

Table 25.3.: Representation 1 from Level 1.

1-2	consequence (%)									
condition	‘0’	‘1’	‘2’	‘3’	‘4’	‘5’	‘6’	‘7’	‘8’	‘9’
‘0’	0	0	0	0	0	0	0	0	0	100
‘1’	100	0	0	0	0	0	0	0	0	0
‘2’	0	100	0	0	0	0	0	0	0	0
‘3’	0	0	100	0	0	0	0	0	0	0
‘4’	0	0	0	100	0	0	0	0	0	0
‘5’	0	0	0	0	100	0	0	0	0	0
‘6’	0	0	0	0	0	100	0	0	0	0
‘7’	0	0	0	0	0	0	100	0	0	0
‘8’	0	0	0	0	0	0	0	100	0	0
‘9’	0	0	0	0	0	0	0	0	100	0

Table 25.4.: Representation 2 from Level 1.

	consequence (%)		
condition	1-0	1-1	1-2
1-0, (‘0’)	0	0	100
1-0, (‘1’)	0	100	0
1-0, (‘2’)	0	100	0
1-0, (‘3’)	0	0	100
1-0, (‘4’)	0	100	0
1-0, (‘5’)	0	100	0
1-0, (‘8’)	0	100	0
1-0, (‘9’)	0	100	0
1-1, (‘0’)	0	0	100
1-1, (‘1’)	100	0	0
1-1, (‘3’)	0	0	100
1-1, (‘4’)	100	0	0
1-1, (‘5’)	100	0	0
1-1, (‘6’)	100	0	0
1-1, (‘7’)	100	0	0
1-1, (‘8’)	100	0	0
1-2, (‘0’)	100	0	0
1-2, (‘1’)	100	0	0
1-2, (‘2’)	0	100	0
1-2, (‘3’)	100	0	0
1-2, (‘4’)	0	100	0
1-2, (‘5’)	0	100	0
1-2, (‘6’)	0	100	0
1-2, (‘7’)	100	0	0
1-2, (‘8’)	100	0	0
1-2, (‘9’)	100	0	0

Table 25.5.: Representation 0 from Level 2.

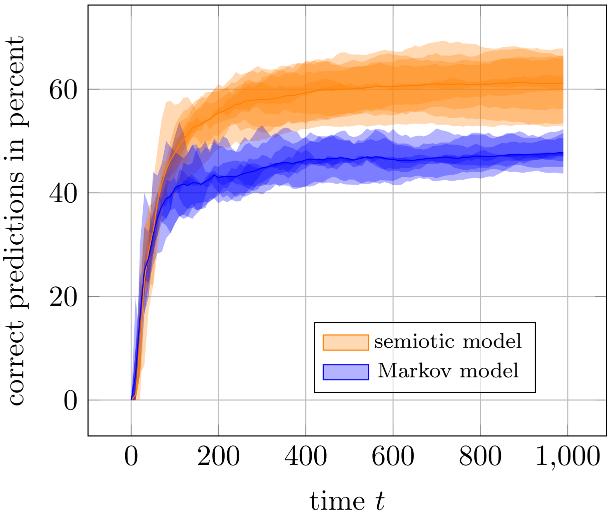


Figure 25.4.: Averaged Performances in Artificial Sequence Prediction.

25.2. Supervised Prediction

Models that count transition frequencies suggest themselves to an evaluation in prediction tasks. The previous section already provided performance estimates for semiotic models and the baseline approach. To verify these estimates, their actual performance over ten learning passes has been recorded.

Each pass is independent from the others and goes on for 1000 time steps. The emission received by the system after each of these time steps is determined by the dynamic uncoupled trajectory defined in the introduction and illustrated in figure 25.1.

In the previous section, the baseline performance is estimated to be $\widehat{L}_b \approx 0.5$ and the performance of the semiotic model to be $\widehat{L}_s \approx 0.56$. Figure 25.4 shows the average success $\bar{S}$ of both approaches, where $\bar{S} = 1 - \bar{L}$.

The estimate for the baseline approach is within the range of results. The estimate for the semiotic model is slightly above the estimated average success rate of $\bar{S} \approx 1 - 0.44 = 0.56$.

This might be due to the fact that the average segment length is only determined from a particular—potentially non-representative—part of the whole trajectory. Also, the first and last segment in figure 25.3 are cropped, effectively reducing the average segment length. Finally, the average learning curve has a variance that accommodates for the observed deviation.

26. Modelling a Natural Sequence

The evaluation with a natural dataset shows that semiotic models can also be generated to describe a trajectory with *unknown* properties. In the comparative part of this chapter, the influence that assumptions about these properties have on predictive success are explored experimentally.

The natural trajectory is written English language. Such texts consist of smaller segments that can be described as a functional transition from one character to another.

In contrast to the artificial trajectory in the previous chapter, the transitions in language occur at multiple levels: between the characters in a syllable, between the syllables in a word, between the words in a sentence, between the sentences in a paragraph, and so on.

Another difference is that linguistic transitions do not occur randomly. Instead, linguistic segments transition according to high-level functions which are influenced, for example, by syntactic rules and, eventually, the author's conception of a story.

This applies to all types of text (e.g. different languages) for as long as the text itself is orthographically and grammatically correct. Due to its length and easy availability through Project Gutenberg, *Pride and Prejudice* by Jane Austen provides a handy example for a natural trajectory. (Austen 1998)

The only pre-processing is the conversion of upper case into lower case letters and the removal of any character that is neither a Latin letter, an Arabic digit, or punctuation. Figure 26.1 shows part of the text.

26.1. Unsupervised Model Generation

Great parts of natural text are determined by syntax, language, or semantics that could—at least hypothetically—be inferred from the text alone. However, there is always the writer's *intention* which cannot be reasonably expected to be predictable even by a

```
...jane had sent caroline an early answer to her letter, and was counting
        the days till she might reasonably hope to hear again...
```

Figure 26.1.: Part of a Natural Trajectory.

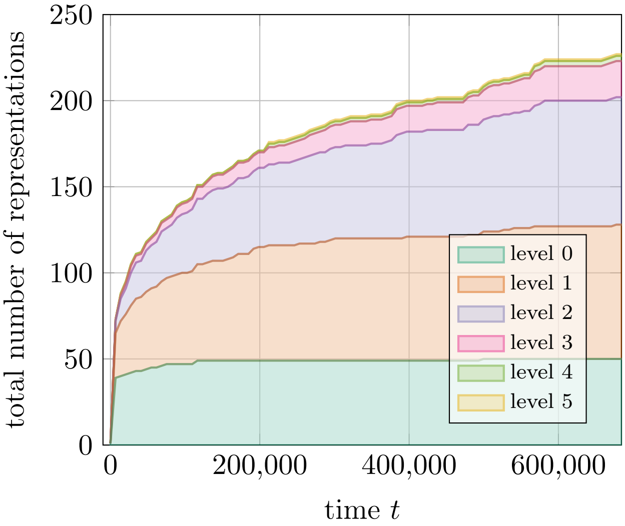


Figure 26.2.: Representation Generation in Semiotic Model.

perfect linguistic model.

During the generation of the model, a certain amount of unpredictability must therefore be tolerated. This fundamental uncertainty is considered by setting the parameters of the semiotic model to $\alpha = 1.0$ and $\sigma = 0.1$.

Decreasing σ enables to recognise facts with a certainty of at least 10 percent. Increasing α provides unknown referents with a 'head start' certainty (for details, see definition 4 on page 124).

The model does not converge despite the reduction in precision that follows from these changes. Considering the fact that language is eventually determined by the author's *intention,* however, it does not come as a surprise that Austen's intention cannot be determined by reading *Pride and Prejudice* once. At the end of the linear evaluation, the results from *several* reading passes are analysed in comparison.

26.1.1. Structural Model Development

After the roughly 680 000 characters in the text, the model features 50 representations at level 0 (i.e. characters), 78 representations at level 1, 74 representations at level 2, 21 representations at level 3, 3 representations level 4, and 1 representation at level 5. Figure 26.2 shows the development over time.

Of course, the representations at the various levels of the semiotic model do not correspond to the linguistic representations of natural language (e.g. syllables, words, or paragraphs). However, they *do* segment the text into reoccurring parts that follow one another according to a higher-level function.

In contrast to the artificial trajectory before, here, the length of the sequence does not allow to infer estimates on the predictive abilities of the model. Lowering σ makes the model more tolerant to wrong predictions. The loss would therefore be much higher than the inverse of the average length of segments.

26.1.2. Representation Associations

The semiotic model of the natural trajectory is much more complex than the model before. Therefore, the presentation of recognised segments is limited to only three of the total number of six levels. Like in the chapter before, the lowest level is omitted. Because only a small segment of the whole sequence can be presented, levels with segments that last longer than what can be printed on a single page are omitted as well. This concerns levels four and five.

Figure 26.3 shows the recognised segments at levels one to three that have been associated with the characters in the first part of the last paragraph of the text. Especially figure 26.3b shows interesting segmentations, where representations often cover one or several words without breaking them apart.

This can be attributed to the fact that words are separated by spaces. Representations at level one cannot reliably predict the next emission after a space. Therefore, representations at level one and above end considerably more often with spaces than with any other character.

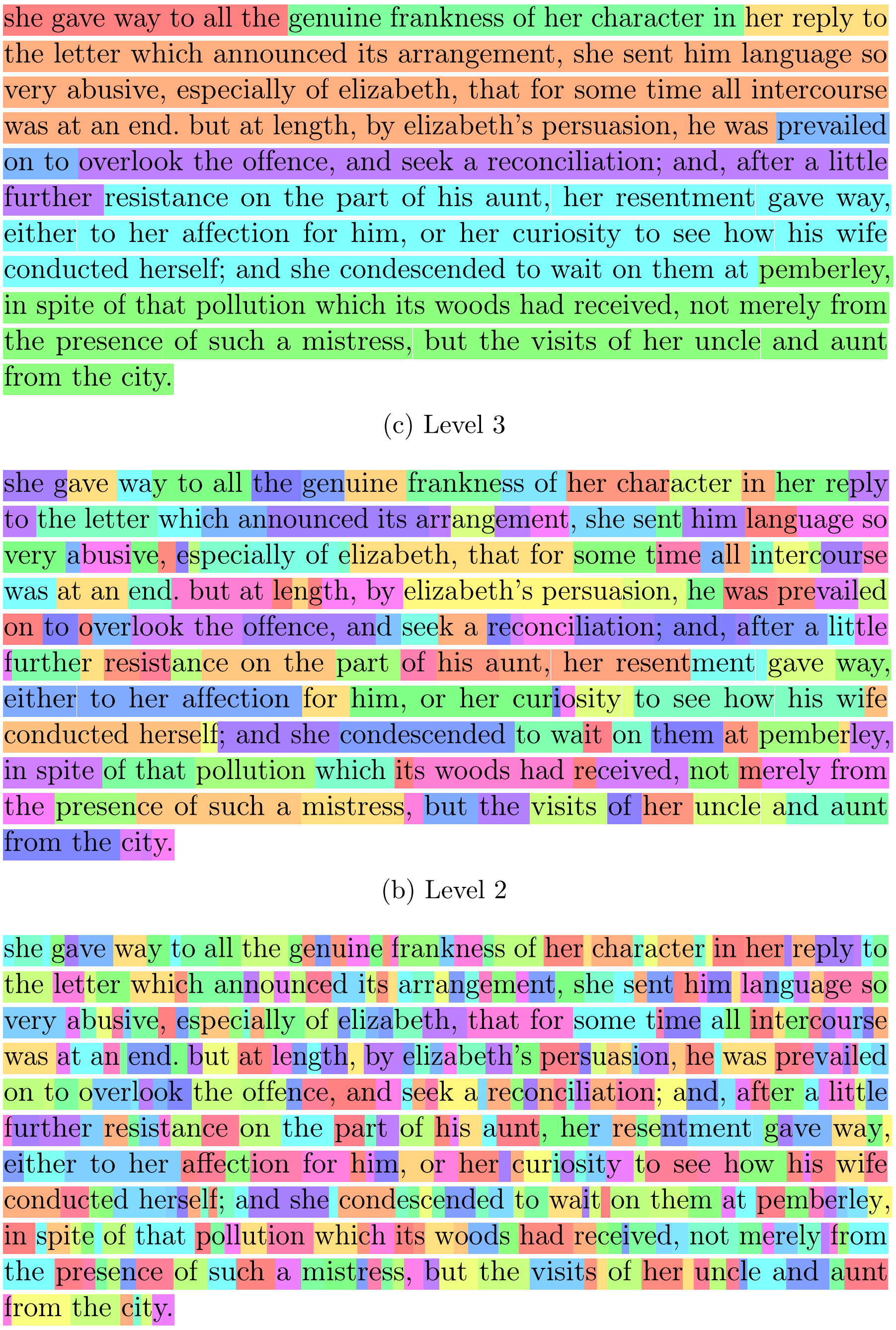
26.2. Supervised Prediction

If cognitive systems understand a particular text, then they are also able to predict its characters with a high rate of success. Successful character prediction indicates access to *the content* of the text. Unfortunately, no system that perceives *only text* can access the same semantic content like systems that feature other modes of perception as well.

However, *this is not due to the symbolic quality of these characters* but rather due to the fact that the a mode of perception that is based on linguistic characters follows essentially different *contingencies* than the basic modes of biological perception (see section 8.3).

Given the inaccessibility of content that results from essentially different modes of perception, predictive success at least enables to show that a system is able to uncover hidden structural regularities.

Figure 26.4 shows the predictive performance during the on-line learning of a semiotic model and a order-1 Markov predictor. The semiotic model reaches an eventual perform-



she gave way to all the genuine frankness of her character in her reply to the letter which announced its arrangement, she sent him language so very abusive, especially of elizabeth, that for some time all intercourse was at an end. but at length, by elizabeth’s persuasion, he was prevailed on to overlook the offence, and seek a reconciliation; and, after a little further resistance on the part of his aunt, her resentment gave way, either to her affection for him, or her curiosity to see how his wife conducted herself; and she condescended to wait on them at pemberley, in spite of that pollution which its woods had received, not merely from the presence of such a mistress, but the visits of her uncle and aunt from the city.

(c) Level 3

she gave way to all the genuine frankness of her character in her reply to the letter which announced its arrangement, she sent him language so very abusive, especially of elizabeth, that for some time all intercourse was at an end. but at length, by elizabeth’s persuasion, he was prevailed on to overlook the offence, and seek a reconciliation; and, after a little further resistance on the part of his aunt, her resentment gave way, either to her affection for him, or her curiosity to see how his wife conducted herself; and she condescended to wait on them at pemberley, in spite of that pollution which its woods had received, not merely from the presence of such a mistress, but the visits of her uncle and aunt from the city.

(b) Level 2

she gave way to all the genuine frankness of her character in her reply to the letter which announced its arrangement, she sent him language so very abusive, especially of elizabeth, that for some time all intercourse was at an end. but at length, by elizabeth’s persuasion, he was prevailed on to overlook the offence, and seek a reconciliation; and, after a little further resistance on the part of his aunt, her resentment gave way, either to her affection for him, or her curiosity to see how his wife conducted herself; and she condescended to wait on them at pemberley, in spite of that pollution which its woods had received, not merely from the presence of such a mistress, but the visits of her uncle and aunt from the city.

(a) Level 1

Figure 26.3.: Represented Segments.

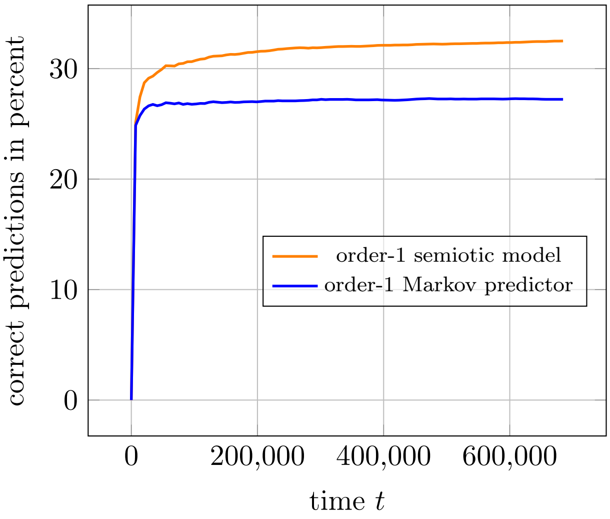


Figure 26.4.: Performance Comparison for the Prediction of a Natural Trajectory.

ance of 32.50 percent while the Markov predictor tops out at 27.22 percent. The semiotic model also maintains a stronger upward trend, indicating its ability to describe more of the trajectory, would it continue.

Due to the static trajectory of one particular text—in contrast to the probabilistic trajectory from before—, the text can only be extended by repetition. It is important to note that this contradicts the initial premise of a single pass over the data. However, it enables to illustrate whether previously obtained knowledge about sequential structure can be successfully applied again and in similar (i.e. identical) circumstances.

A small change in algorithm 8 has been implemented to consider *historical transition conditions* such that the state of a semiotic model maintains a history of n past representations at each level. After this modification, semiotic models can be compared to order-n Markov predictors, where $n \geq 2$.

As a consequence, the updates of *each individual level* in the state of a semiotic model in lines 4, 6, and 9 of algorithm 8 on page 166 now follow algorithm 5 on page 163 that describes the update of *the complete* state of an order-n Markov predictor.

Queries for the current representation at leach level in lines 19 and 23 of algorithm 8 have to be adjusted accordingly to return only the last element of these histories.

Figure 26.5 provides an overview for $n = 1$, $n = 2$, $n = 3$, $n = 4$, and $n = 5$ while learning a ten times repetition of the original text. The vertical red line indicates the end of the first pass. The semiotic models and Markov predictors with $n = 1$ left of this line are identical to those in figure 26.4.

Markov predictors reach their performance peak faster, but these peaks are consistently

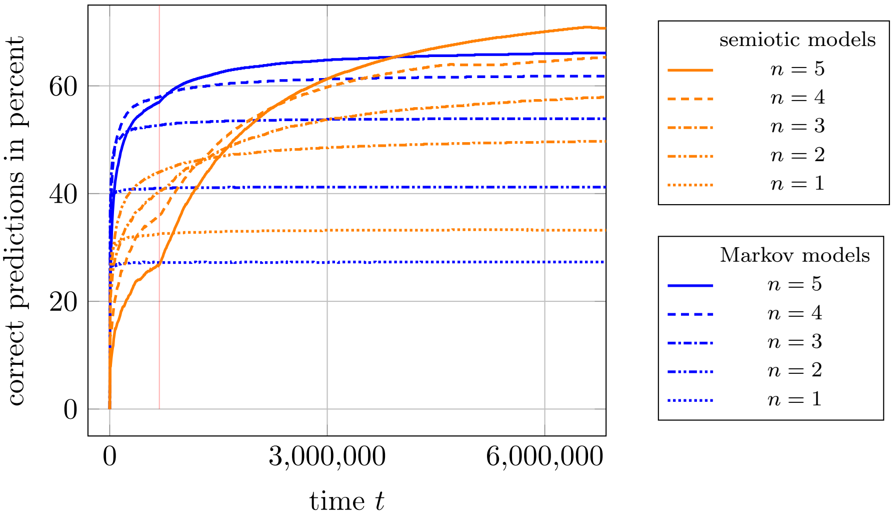


Figure 26.5.: Performance Comparison for the Prediction of a Repetitive Natural Trajectory with Varying History Lengths.

lower, than those of semiotic models. This result shows that, the more complex a model is, the longer it requires to reach peak performance. It also shows that semiotic models are able to capture considerably more structural dependencies between the emissions of the trajectory than Markov predictors.

Most of these dependencies can be assumed to be long-term, due to the fact that semiotic models outperform specifically Markov predictors, for which temporal dependencies beyond n are plain impossible to describe. Semiotic models cover such long-term dependencies in virtue of their hierarchical state which can maintain a particular context for *arbitrarily* extended periods of time.

This assumption is substantiated by the fact that semiotic models benefit considerably more from repetitions in the trajectory, as the steep incline at the start of the first repetition demonstrates. It can be inferred that dependencies between characters at the beginning of the text are maintained, whereas Markov predictors overwrite them with the most current information.

27. Simulating Cognitive Modelling with Reinforcement Learning

Reinforcement learning is a general framework to describe the goal-directed interaction between autonomous agents and another system. Often times, the other system is conceived of as the agent's environment. But it could just as well be another agent or any system that emits shapes for the agent to receive.

In this chapter, a reinforcement learning framework for evaluating simulations of cognitive systems is presented. The focus is on three abilities of cognitive systems: 1) generating internal representations for hidden states of the environment, 2) making predictions that depend on these representations, and 3) using these predictions for goal-directed behaviour.

The agent's environment in the evaluation with a coupled trajectory is a *grid world.* Therefore, 1) *the state of the model* is effectively a representation of the agent's absolute position in this grid world, 2) *the predicted emissions* are the agent's neighbouring cells in this grid world, and 3) *goal-directed behaviour* is the finding of a path from one position to another.

Localisation is the agent's ability to represent its position and orientation in space. A representation of position and orientation in a partially observable grid world is a representation of the hidden state of the environment.

The content of these representations is inaccessible to the observer. Localisation performance can only be determined indirectly: in virtue of the agent's performance in *predicting* the next emission of the environment.

Prediction performance can be determined like in the linear case before, independent from the agent's action policy. Transition conditions are composed of the last sensor activation *and* one of several possible motor activations (see section 16.3).

Navigation, eventually, requires a particular action policy. Successful navigation is a sequence of goal-directed motor activations. The agent's goal in the grid world is to reach a particular position. In reinforcement learning, agents identify goals in virtue of reward that they receive as soon as the environment enters a particular state.

An internal model of the environment can help the agent to consider the outcomes of its actions and to find a policy more *efficiently* (i.e. faster). An internal representation of states that are hidden to the agent, makes it possible to find more *effective* (i.e. better) policies.

27.1. Defining the Grid World

In each grid world is only *one* agent. At any point in time, this agent occupies exactly one open space position and is oriented towards one of four major directions.

Each cell in this grid is either occupied by a wall, which is represented to the agent with the character '`x`', or by an open space, represented to the agent with the character '`.`'.

The grid world is a labyrinth on a discrete, two-dimensional Cartesian coordinate system. It is generated such that each open space can be reached from any other open space by repeatedly transitioning a single cell towards any of the four major directions.

If the agent tries to move onto a cell which is occupied by a wall, it remains in its original position and orientation. Sensor and motor activations beyond the area defined by this grid are handled with the modulo operator, effectively making the grid world *toroidal.*

In the following, two sensor modes and two motor modes determine a total of four different modes of interaction between agent and environment. Also, parameters are introduced for making the agent's actions uncertain.

These different variations provide a general framework, in which a dynamically coupled trajectory between agent and environment can be modelled. These different modes enable to evaluate algorithms in *different* partially observable Markov decision processes that are based on the *same,* relatively easy to conceive, grid world.

The evaluation is constrained to one of these variants where the outcome of actions is fully determined by the state of the environment (i.e. no action uncertainty). The same grid world, however, can also be used to evaluate models that have been generated by 'clumsy' agents during other modes of interaction.

27.1.1. Modes of Interaction

Rotational movement is performed by emitting one of three possible motor activations. Two motor activations rotate the agent ninety degree clockwise or counter-clockwise. They are represented by '`l`' for counter-clockwise rotation and '`r`' for clockwise rotation. The third activation transitions the agent one cell into the direction of it's current orientation and is represented by '`f`'.

Transitional movement is performed by emitting one of four possible motor activations, each of which transports the agent to one of the four adjacent positions. When moving transitionally, the agent is always oriented to the north. The four transitional activations are represented by '`n`' for moving to the north, '`e`' for moving to the east, '`s`' for moving to the south, and '`w`' for moving to the west.

The agent can also receive two mutually exclusive types of sensor activations. Both of which can be combined with each motor mode. It receives either the *four surrounding cells* or the *eight surrounding cells.* In each case, the received sensor emission is a tuple starting with the cell that the agent is oriented towards, continuing clockwise.

For benchmarking purposes, there is also a sensor mode that provides the agent with its *absolute* coordinates within the grid world. In this mode, the grid world environment effectively becomes *fully observable* to the agent.

To introduce noise, the agent can be made 'clumsy'. This is controlled with a real value between 0 and 1. For each action, a uniformly distributed random real value from $[0, 1]$ is drawn. In case this value undercuts 'clumsiness', the agent performs a uniformly random action drawn from the actions available in the current motor mode.

27.1.2. Formalisation

Grid worlds can be formalised as *partially observable Markov decision processes.* A partially observable Markov decision processes is a 7-tuple $(\ S, A, T, \Omega, O, R, \gamma\ )$.

The first five elements describe a partially observable environment according to section 16.1.1, where $S = E$, $A = \mathsf{motor}$, and $\Omega = \mathsf{sensor}$. The last two are required to define a particular goal for the agent.

The element S is a set of world states, A is a set of motor emissions, $T : s \times A \times S \to [0, 1]$ are the transition probabilities from one state to another given a particular motor emission, Ω is a set of sensor emissions, and $O : s \times A \times \Omega \to [0, 1]$ are the probabilities for a particular sensor emission in a given state when emitting a particular motor activation.

In the grid world, the description of sensor probabilities can be simplified to $O : s \to \Omega$ because the emissions of the environment are determined only by the agent's position and not probabilistically conditional on the agent's emissions.

Also, in a grid world, S is usually the cross-product of all agent positions and orientations. Figure 27.1 shows, for example, that the coordinates $(3, 2)$ can be occupied by the agent and '`n`' is a valid orientation for the agent. Therefore, $\big\langle (3, 2), \text{'}\texttt{n}\text{'} \big\rangle \in S$.

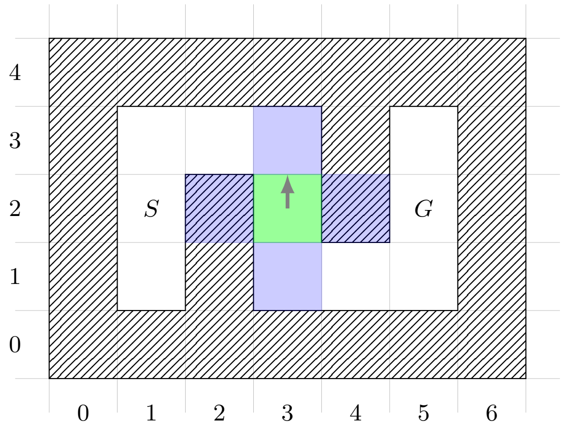


Figure 27.1.: An Example Grid World.

The set of possible motor emissions during transitional movement is $A = \{ \text{‘}\texttt{n}\text{’}, \text{‘}\texttt{e}\text{’}, \text{‘}\texttt{s}\text{’}, \text{‘}\texttt{w}\text{’} \}$. During rotational movement, $A = \{ \text{‘}\texttt{l}\text{’}, \text{‘}\texttt{r}\text{’}, \text{‘}\texttt{f}\text{’} \}$, respectively.

Take the following two states during transitional movement in the grid world from figure 27.1: $s_0 = \big\langle (3,2), \text{‘}\texttt{n}\text{’} \big\rangle$ and $s_1 = \big\langle (3,3), \text{‘}\texttt{n}\text{’} \big\rangle$. When the agent emits ‘`s`’ without being clumsy, then $T(s_1, \text{‘}\texttt{s}\text{’}, s_0) = 1$. When being maximally clumsy, on the other hand, $T(s_1, \text{‘}\texttt{s}\text{’}, s_0) = \frac{1}{4}$, because the motor emission is completely random.

Each element c in a sensor emission can have one of two different values $c \in C = \{ \text{‘}\texttt{x}\text{’}, \text{‘}\texttt{.}\text{’} \}$. Accordingly, for agents that receive four surrounding cells, the set of sensor activations is $\Omega \subseteq C \times C \times C \times C = C^4$.

The sensor activation in state $s = \big\langle (3,2), \text{‘}\texttt{n}\text{’} \big\rangle$, for example, is $O(s) = (\text{‘}\texttt{.}\text{’}, \text{‘}\texttt{x}\text{’}, \text{‘}\texttt{.}\text{’}, \text{‘}\texttt{x}\text{’})$. In state $s' = \big\langle (3,2), \text{‘}\texttt{e}\text{’} \big\rangle$, the activation is ‘rotated’ by 90 degree clockwise: $O(s') = (\text{‘}\texttt{x}\text{’}, \text{‘}\texttt{.}\text{’}, \text{‘}\texttt{x}\text{’}, \text{‘}\texttt{.}\text{’})$. For agents that receive eight cells, $\Omega \subseteq C^8$, accordingly.

The element $R : s \times A \to \mathbb{R}$ is the reward the agent receives when performing a particular action in a particular world state. In the grid world, the agent's goal is a position. Therefore, the reward function can be simplified to $R : s \to \mathbb{R}$.

In the grid world, the agent receives a continuous reward of -1, *except* it reached goal position. If this is the case, the agent immediately receives a reward of 10 and its position is reset.

For the agent's immediate decisions, however, reward that occurs one thousand time steps in the future might not be as relevant as reward that occurs sooner.

The relevancy of future reward can be controlled by $\gamma \in [0, 1]$. For $\gamma = 0$, the agent only cares about the immediate reward, whereas for $\gamma = 1$, it cares about all future rewards to the same extend.

Figure 27.1 illustrates an example grid world with start position ‘`S`’, goal position ‘`G`’, and an agent that is oriented to the north and that receives the four surrounding cells as ('.', 'x', '.', 'x').

27.2. Cognitive Systems as Reinforcement Learning Agents

Reinforcement learning is usually performed quite different from how cognitive systems learn to interact with external reality. In the following, two differences are described that are most important according to the experience gathered during evaluation.

Two according changes to the classical reinforcement learning paradigm are proposed. These changes maintain the possibility for traditional agents to compete with agents that generate a partially observable model of the environment.

Traditional reinforcement learning algorithms are not intended to describe the interaction between cognitive system and external reality. This shows in the fact that they perform comparatively weak under these changes.

27.2.1. Preserving the State of the Agent

Sarsa-learning according to Sutton and Barto (1998, chapter 6.4) enables goal-directed action selection. Sarsa is a conventional value iteration reinforcement learning algorithm. Action selection follows the maximum of an evaluation function which is determined as follows.[1]

$$Q(s_{t-1}, a_{t-1}) \leftarrow Q(s_{t-1}, a_{t-1}) + \alpha\big(r_t + \gamma Q(s_t, a_t) - Q(s_{t-1}, a_{t-1})\big) \qquad (27.1)$$

The element s_{t-1} is the environment's last emission, a_{t-1} is the agent's last emission, and r_t is the reward received after the following time step. The parameter α determines the learning rate and γ determines the discount factor (see the end of section 27.1.2).

In the comparative setting, s_t is the current state of the semiotic model or the state of the Markov predictor respectively. The action a_t is the system's motor emission $m_t \in \mathsf{motor}$. For each action selection, there is a $\epsilon = 0.1$ chance for uniformly random exploratory behaviour.

After each time step, reinforcement learning agents receive an experience $\langle s_t, a_t, r_t \rangle$, where r_t depends not on s_t and a_t but on s_{t-1} and a_{t-1} instead.[2] To relate action and reaction, therefore, the agent needs to *memorise* s_t and a_t, to access them after the next

[1] In comparison to Sutton and Barto, we shifted the temporal index by one into the past so as to avoid a reference to future rewards.

[2] This is because reward is considered as part of the environment's *reaction* to the actions of the agent (see, for example, Sutton and Barto 1998, Summary of Notation, under 'r_t').

time step as s_{t-1} and a_{t-1}. This enables to incorporate the current reward into the Q-evaluation of the *last* experience as shown in equation (27.1).

Navigation tasks in reinforcement learning are commonly assumed to be *episodic.* Once the agent has reached its goal, the current episode ends. At the end of each episode, the agent's position is reset.

Without resetting the agent's position, it would merely linger at its goal state. Such behaviour can hardly be called 'navigation'—although it is cognitively quite realistic.

By resetting the position of the agent after reaching a goal state, however, the reward in this new position 'clouds' the agent's evaluation of the immediately preceding goal state. To avoid this, the agent's memory is usually reset after each episode as well.

Unfortunately, resetting the agent's memory after reaching a goal state is an infusion of external knowledge. Externally resetting the agent implies that the ultimate goal state has been reached *beyond* what the agent can infer from the reward it receives. *Cognitive* agents have no access to this kind of knowledge.

27.2.2. Circular Goals

A 'teleportation' at the end of each episode is also hardly justifiable with the actual interaction between real cognitive systems and external reality. It appears to be necessary, however, to avoid the agent lingering at the same goal state where it cannot learn anything new.

Only if the agent's memory is left *intact,* the displacement of the agent is an actual problem. The position reset is initiated externally with no way for the agent to integrate it into its model. It 'breaks' the agent's stream of consciousness with little chance for anticipation. As a consequence, the agent's memory is simply wiped such that this discrepancy is not integrated into the agent's model in the first place.

Circular goals provide a cognitively more justifiable approach. Circular goals are defined as a list of goal states in the same environment, each of which becomes active only once its immediate predecessor has been achieved.

This alternation of goals is supposed to be a simplified simulation for *the agent's needs.* These needs tend to emerge in an alternating manner as well. This is a more natural incentive for agent activity than malignant displacement. Also, it allows to test the transfer of knowledge from one task to another.

Usually, knowledge transfer is investigated either by placing an agent in a somehow self-similar or repetitive environment (e.g. in hierarchical reinforcement learning) or by placing an agent who has learned in a *particular* environment into *another* environment where it is supposed to apply previously acquired knowledge.

The former serves to increase the speed (i.e. the efficiency) of learning, but *not* its eventual effectiveness. Although humans exploit similarities in their environment as well, they recognise *dissimilar* segments of the environment with *the same* representations: they recognise 'creatively'. This aspect can hardly be observed in objectively repetitive environments. In the various different segments of one environment with circular goals, however, it can.

The latter presents the same problem like resetting the agent's position after reaching a certain goal state. An unpredictable change in the state of the environment necessarily introduces uncertainty into the agent's model of the environment. In the interaction between cognitive systems and external reality, there is usually no such unpredictable transition.[3]

[3] In the case of a *breakdown* (see section 12.6), the system adapts its mental model such as to avoid it in the future.

28. Modelling a Dynamic System

The evaluation with another dynamic system shows that semiotic models can be generated to describe dynamically *coupled* trajectories. This is demonstrated within the reinforcement learning framework from the previous chapter. This framework serves as a basis for more faithful simulations of the interaction between cognitive system and reality—especially cognitive modelling.

The qualitative evaluation in the grid world setting is analogue to the qualitative evaluation in the linear case. It concerns the temporal development of representations in the model and the segments of the environment that are associated with these representations.

The quantitative evaluation is separated into a prediction task and an interaction task. The prediction task investigates the agent's ability to *localise* itself within the grid world. The interaction task investigates its ability to use localisation to *navigate* successfully.

28.1. Unsupervised Model Generation

The structure of the grid world is inspired by Sutton (1990). It is illustrated in figure 28.1. During model generation in the localisation task, the agent moves randomly across all open cells. Its mode of action is *transitional,* its mode of perception covers *eight neighbouring cells,* and 'clumsiness' is set to zero.

There is no goal and, therefore, no reward. Each simulation run consists of five million iterations. During each iteration, the state of the agent's semiotic model is recorded so that it can later be associated with the true state of the environment (i.e. the agent's position).

28.1.1. Structural Model Development

Figure 28.2 shows the development of the model during one simulation run. At the end of the simulation, the model contains 99 representations at seven levels. The lower levels converge relatively early on. The latest of which is level 2 with consistent 18 representations starting from around time step 2 680 000.

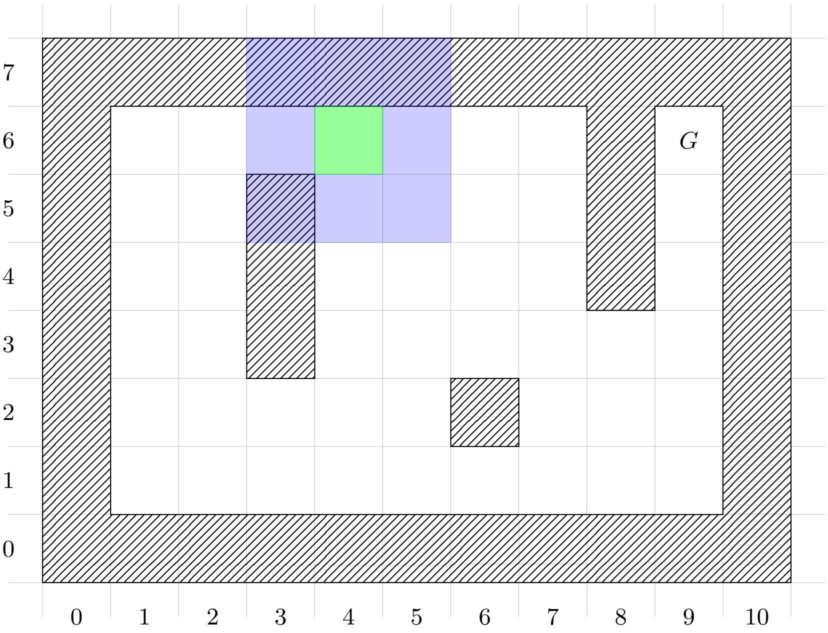


Figure 28.1.: Sutton's Grid World.

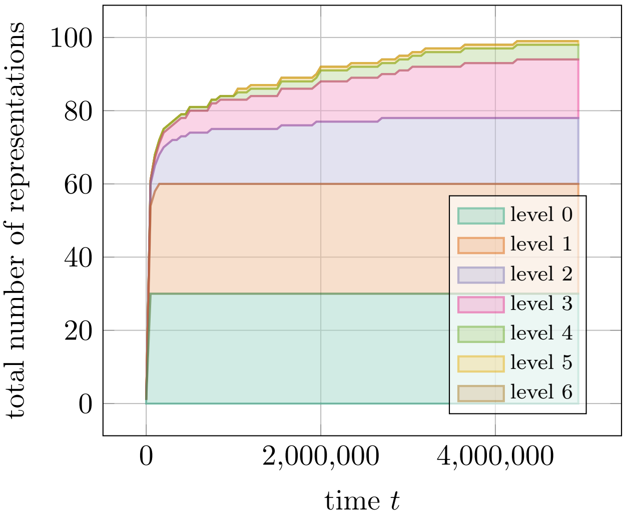


Figure 28.2.: Representation Generation during Localisation.

Levels 3 and above appear as though more representations are generated in future time steps. The total number of new representations in each time step, however, approaches zero, as does the number of new levels.

Model convergence can be influenced by adjusting α and σ. The selection of $\alpha = 1$ and $\sigma = 0.1$ was not optimised in this respect but carried over from the linear evaluation in the previous chapter.

The total number of representations could be considerably reduced by making the system 'forget' representations that have not been used for some time. However, this also introduces the need for another parameter to determine when *exactly* representations are to be removed from the model.

Unfortunately, the definition of any additional parameter implies the danger of specialising the system to the task at hand. Therefore, the ability to forget representations will not be introduced until more research is available.

28.1.2. Representation Associations

It is not trivial to illustrate segments of the grid world that are consistently associated with the same representation. The reason is, on the one hand, that the same segment can be represented differently depending on the current context and, on the other hand, that the same representation can refer to various different segments, depending on their perceived similarity.

The presentation of *all* representations at *all* positions in the grid world goes far beyond the limits of this work. Therefore, the four most frequent representations at level 1 are described and it is illustrated where they are employed most frequently by the agent.

Figures 28.3 to 28.6 show the spatial distribution of these representations. Lighter areas are less frequently associated with the given representation than darker areas.

The four representations do not represent objectively coherent parts of the grid world. They do, however, represent characteristically *different* parts. None of the distributions is uniform over the complete grid world and, although there is some overlap, the core regions of each representation are rather exclusive. This indicates a good coverage of individual and separate segments of the environment.

28.2. Supervised Localisation

The predictions of the system are used to estimate its ability to localise itself within the grid world. Figure 28.7 shows the system's average prediction performance over 10 runs

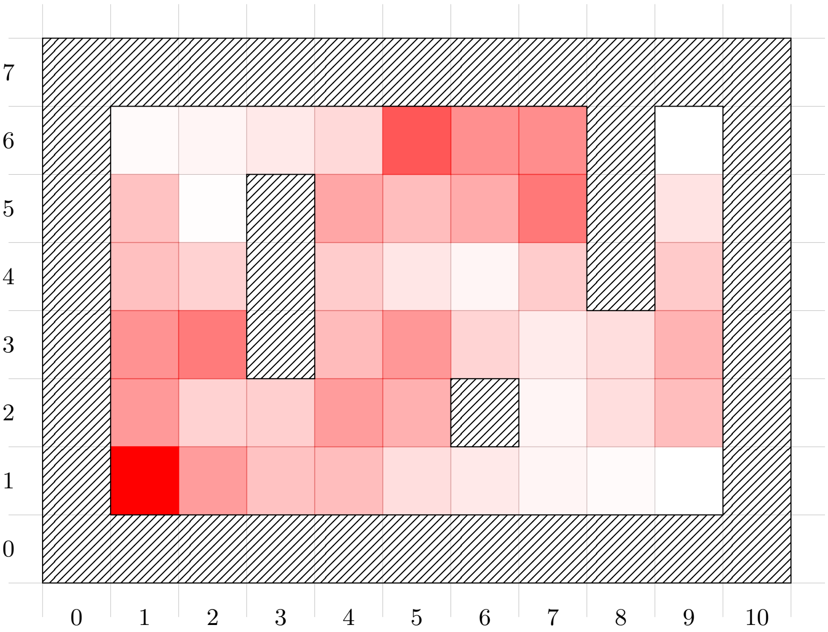


Figure 28.3.: Representation *a* for Localisation.

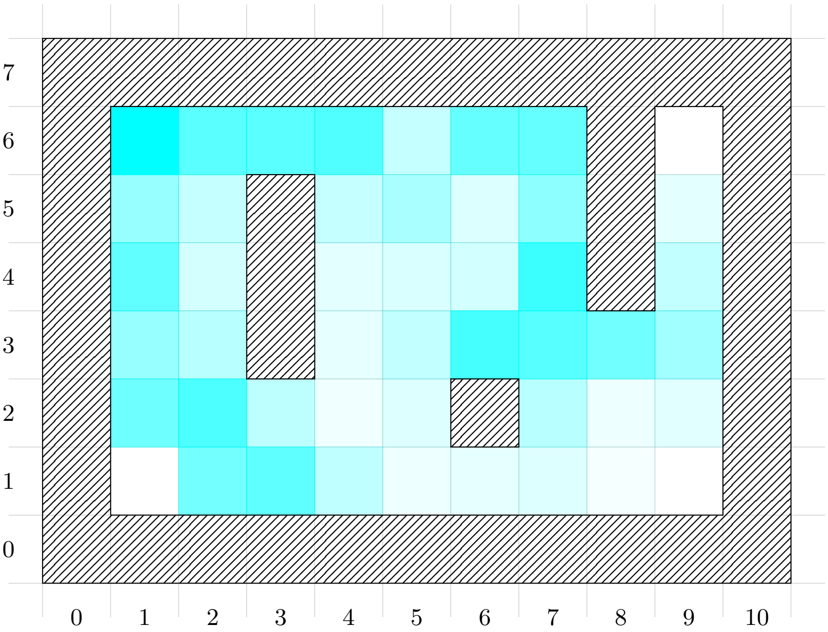


Figure 28.4.: Representation *b* for Localisation.

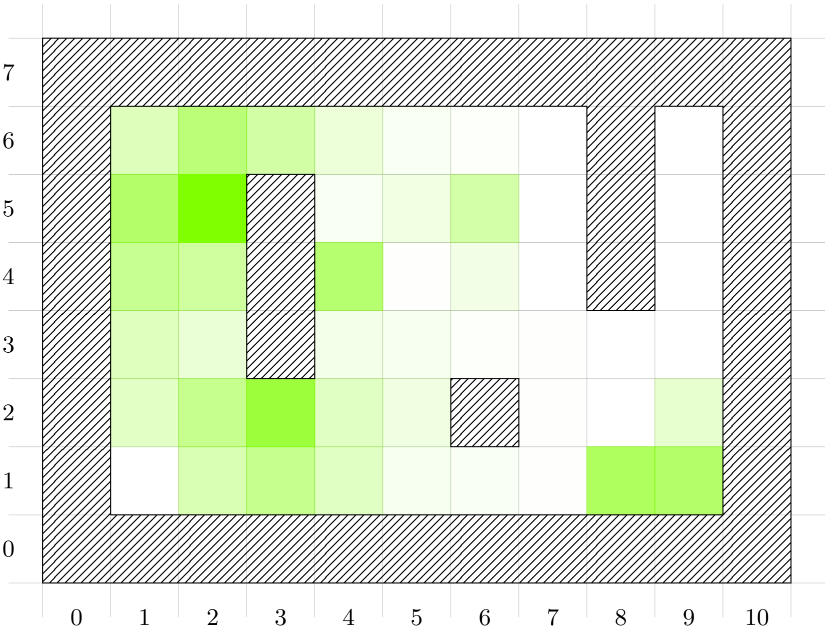


Figure 28.5.: Representation c for Localisation.

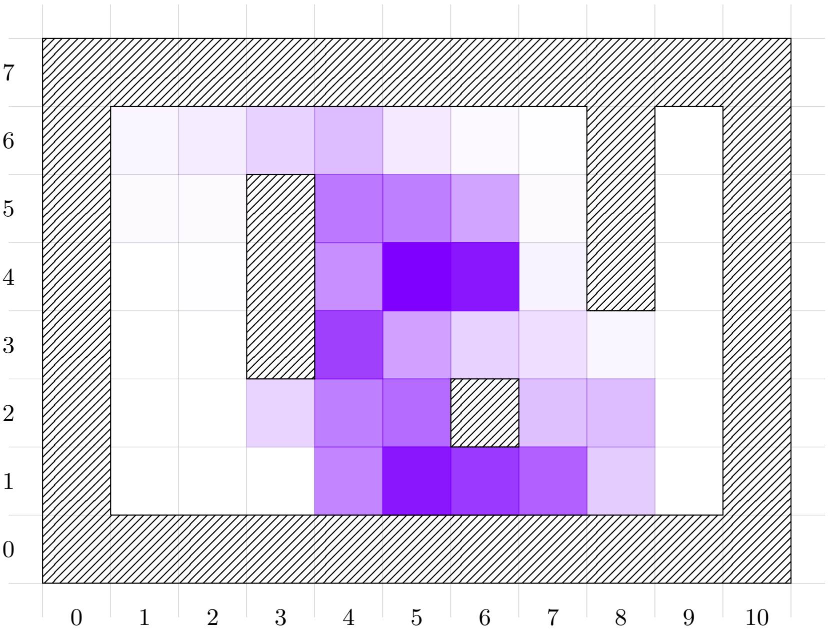


Figure 28.6.: Representation d for Localisation.

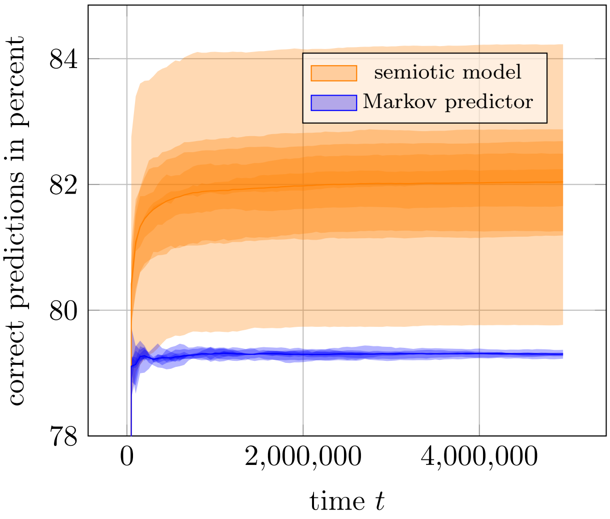


Figure 28.7.: Averaged Performances in Localisation Task.

of 5 000 000 iterations each. The performance of the semiotic model is compared to a Markov predictor with the same history length of $n = 1$.

The semiotic model outperforms the Markov predictor with an average of 82.0 percent compared to an average of 79.3 percent. The graph also shows that the variance of the semiotic model's performance is much greater than the variance of the Markov predictor.

This is probably due to the fact that a lot more transition samples are available for the single function approximation in a Markov predictor whereas a semiotic model has to distribute these samples across several different function approximations (i.e. content). Depending on the agent's random actions, this distribution can be more or less appropriate.

28.3. Reinforced Navigation

In the navigation task, all reward is uniform, except in the goal state. Therefore, all discount factors $0 < \gamma$ are effectively equivalent. In our experiments, we set $\gamma = 1$ arbitrarily.

All free cells are designated as starting positions and one as goal position. This follows from the original presentation of the grid world in Sutton (1990) and does not allow for circular goals as they are introduced in section 27.2.2.

To avoid lingering, navigation tasks are implemented with only one goal and with a random displacement of the agent, once this goal is reached. Also, a simplified navigation task is presented with circular goals and without this displacement.

Due to the dilution of the goal state evaluation when randomly displacing the agent while

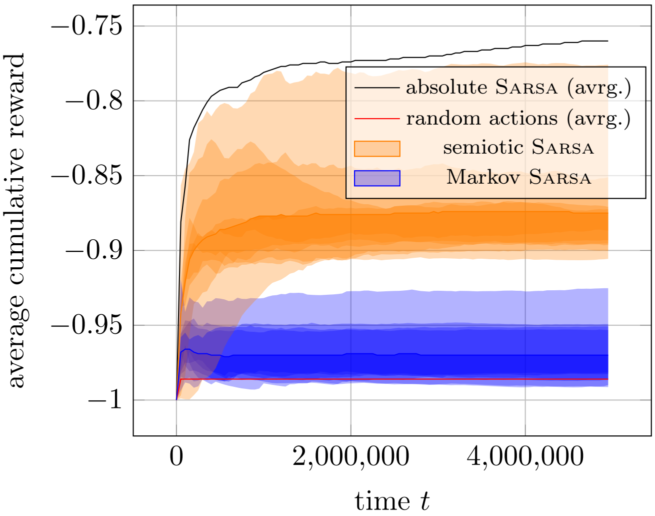


Figure 28.8.: Averaged Performances in Navigation Task.

maintaining its memory, the performance of agents in this environment is considerably lower than usual (again, see section 27.2.2).

Figure 28.8 shows a performance comparison between actions inferred from the states of Markov predictors and semiotic models over 10 runs of 5 000 000 iterations. It also shows the average navigation performance of a system with full perception of the environment's hidden states in black (i.e. the agent position and orientation) and the average performance of a system that acts completely random in red.

The results show that random actions generate on average the least amount of cumulative reward. A Markov-based SARSA is only slightly better on average but sometimes even worse than a random policy.

Similar to the orientation task before, the approach with a semiotic model shows the most variance. After the 1 000 000-th time step, however, even the worst run of semiotic SARSA is better than the best run with Markov SARSA.

The best run with a semiotic model almost reaches the average performance of a SARSA agent that receives its absolute position in the grid world, effectively making the environment *fully observable*.

28.4. Combined Evaluation of Localisation and Navigation

During evaluation, several peculiarities could be observed that afford directions for research. In the remaining section, the most interesting are presented.

Figure 28.9 illustrates the average predictive performance during a single run in the

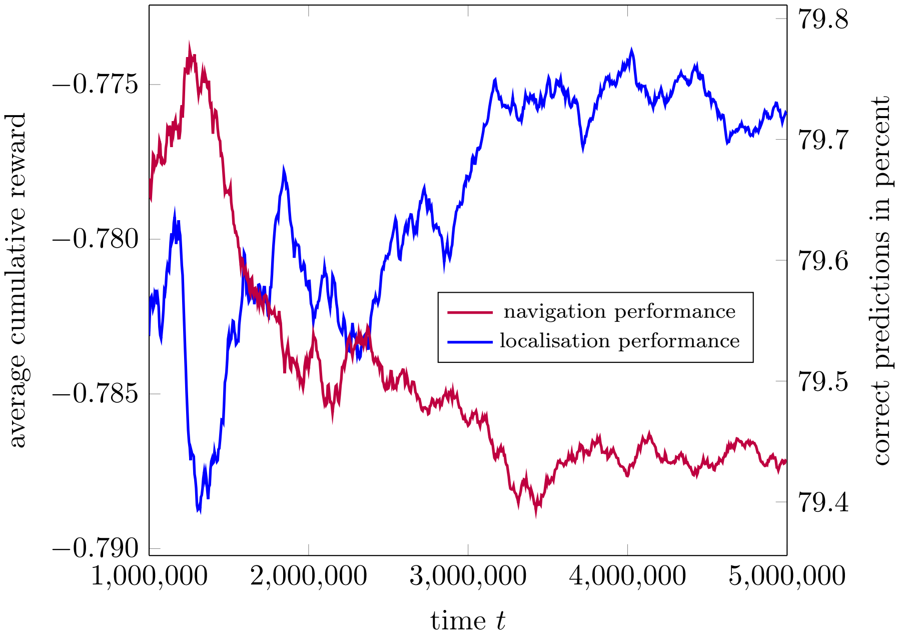


Figure 28.9.: Combined Performance during Navigation and Localisation.

grid world in comparison to the average cumulative reward. Starting from approximately time step 2 000 000, both values develop approximately *inverse.*

If there is in fact an inverse relation between prediction and interaction performance, this would contradict one basic premise onto which semiotic models are built: representations that enable successful *prediction* enable successful *interaction* as well.

Although the evaluation has confirmed this premise in general, it appears as if there is a particular 'incompatibility' between the use of the same representations for prediction *and* interaction.

The same effect also shows in the temporal development of the semiotic model during navigation in figure 28.10. During non-random, goal-directed interaction, *more* representations are necessary for the system to represent its environment than they are during a random action policy according to figure 28.2 on page 198.

It remains to find out whether this is the case in humans as well. If it is not, this might be an indicator that value iteration reinforcement learning like Sarsa is a successful, but possibly not a genuinely *cognitive,* approach to learning goal-directed interaction.

There are alternatives to reinforcement learning. However, only few qualify as potential simulations for the cognitive processes during problem solving. With the exception of goal-agnosticism, this is mainly due to the general conditions for the simulation of mental models in section 21.1.2.

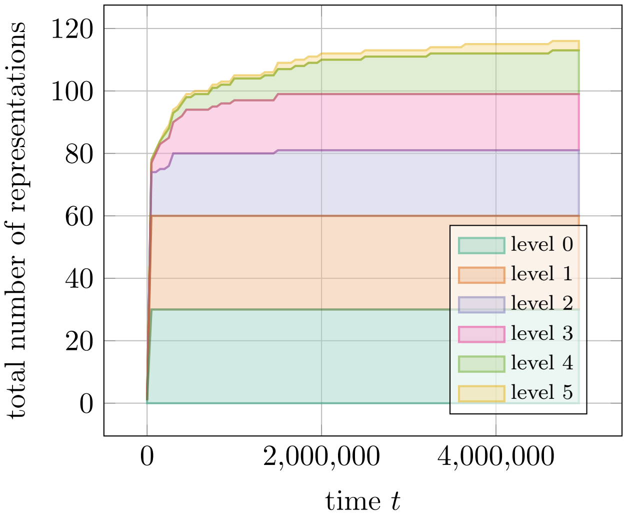


Figure 28.10.: Representation Generation during Navigation.

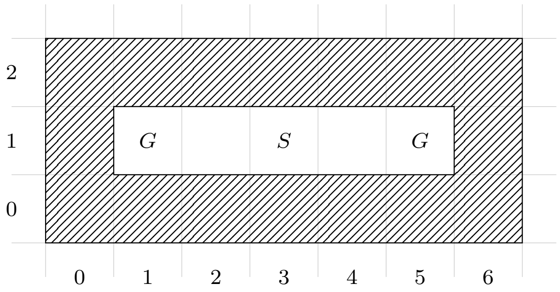


Figure 28.11.: Grid World with Alternating Goals.

Sarsa- and Q-learning can be interpreted in such a way that they do not contradict these conditions. However, neither was implemented with the express intention to simulate human information processing. If they differ substantially from what humans do, alternatives that are *not* incompatible with the corresponding cognitive processes are yet to be found.

28.5. Knowledge Transfer with Changing Goals

The minimal grid world in figure 28.11 enables to investigate the agent's ability for knowledge transfer across different tasks. In more complex environments, the effect is still visible but far less distinct. This indicates the need for improvement or revision of some aspects of our approach. What these aspects are, remains to be determined.

Figure 28.12 illustrates the average navigation performance for ten simulation runs over

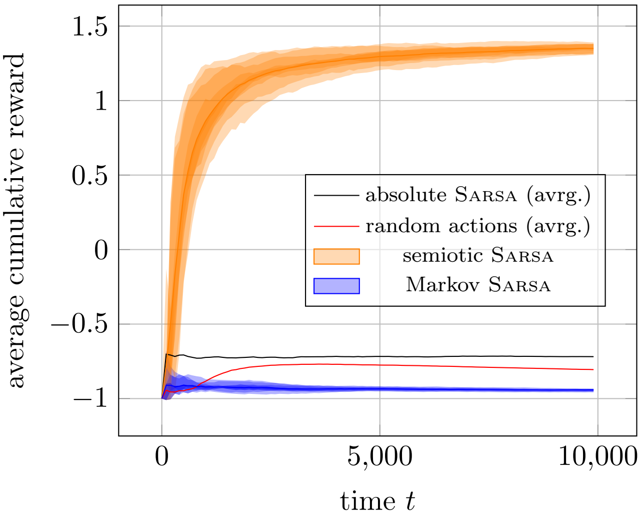


Figure 28.12.: Averaged Performances in Knowledge Transfer.

10 000 iterations. It shows that even the immediate perception of the agent's absolute position enables only slightly better-than-random behaviour if goals change circularly.

SARSA-learning with a Markov model is even worse than random because it gets stuck in local evaluation optima that do not correlate with actual reward. SARSA-learning with a semiotic model, in contrast, achieves a segmentation of the environment that corresponds to the two alternating goal states. It accumulated the most average reward over time.

An optimal policy in this case generates an average cumulative reward per time step of $\bar{r}^* = {}^7\!/_4 = 1.75$. It requires $\Delta t = 4$ to cross from one goal to the other and, during the crossing, a reward of $\Delta r^* = 7$ is accumulated.

With an average accumulated reward of $\bar{r} \approx 1.35$, semiotic SARSA comes very close to $\bar{r}^*$. This is a very promising result, especially considering the fact that $\bar{r}^*$ is not achievable in practice due to ten percent exploratory behaviour.

This shows that representations that describe segments of the environment 'redundantly' enable to provide the information necessary to solve different tasks in the same environment. The redundant representations generated to solve one task can be 'misused' to distinguish different *states of need.*

28.6. Shortcomings of the Baseline Approach

Order-n Markov predictors are not fit to simulate the mental model in a cognitive system for two reasons. The first reason is that they are *phenomenlogically different* from mental models. The second reason is that they are fundamentally limited in their predictive power *due to this difference.*

A simple history is just not a good representation for the current state of reality. It cannot describe nested objects and it does not feature the hierarchical structure of contexts as Dreyfus describes it in section 2.2.1.

Each order-n Markov predictor has up to $n^{|S|}$ states, where S is the set of emissions This state changes almost every time step—except when the complete history consists of one and the same emission. The state of a semiotic model in contrast is updated lazily and it therefore applies to more extended periods of time. Markov predictors change their state *by default* whereas semiotic models state changes occur only *when necessary.*

With each increment in n, the number of states in a Markov predictor multiplies by the total number of emissions $|S|$. With each newly observed emission, it can also multiply by n. Also, emissions that depend only on a history of $m < n$ previous emissions are not only considered *once* but in fact can be stored up to $|S|^{n-m}$ times.

The number of states in a semiotic model, in contrast, can be up to $\Pi_{l=0}^{L}|S_l|$. Which one is more space efficient depends therefore on the rate in which new representations are generated at each level. In a semiotic model, this rate can be changes by modifying α and σ.

The most important drawback of order-n Markov predictors is that there are *always* long-term dependencies that cannot be covered *in principle.* No exclusively history-based approach with a finite n can cover dependencies over more than n time steps.

This difference between order-n Markov predictors and semiotic models is equivalent to the difference between a finite state automaton and a stack automaton. Without *some* form of long-term memory, each condition must be ‘forgotten’ at some point in the future.

In an interactive settings, where the state is used by an agent to interact with the environment, Markov predictors can fail catastrophically. The performance of exclusively history-based approaches can even become *worse than random.* The state of a semiotic model, on the other hand, can be maintained for arbitrarily extended periods of time.

29. Conclusion

The state of external reality is fundamentally unknown to cognitive systems. From first-person-perspective, the simple solution is to assign *meaning* (i.e. mental content) to the phenomenal shapes that are emitted by external reality. Symbolic mental representations are a combination of this shape and content. The symbol grounding problem asks for the origin of the content.

The generation of mental content is not observable from third-person-perspective. Only the generation of structures that *contain* this meaning can be observed. To the observer, these structures are representations for a state of the environment that is *hidden* from the observed system. From third-person-perspective, a solution to the symbol grounding problem is therefore a cognitively plausible generation of representations for hidden states.

The present project followed this general strategy. This conclusion provides an overview over the line of argument behind this strategy, its general premises, the obtained results, shortcomings, alternative routes, and possible extensions.

29.1. Summary of Argument

The present project investigates the border between real and simulated cognition. It suggests a thin overlap under certain premises. The plug of the project is the famous symbol grounding problem. An interpretation of the symbol grounding problem as a *linguistic* problem is rejected and arguments are brought forward against its author's intention to ask merely for the origins of language. Instead, the problem concerns *the very origin of understanding itself*—prior to any form of communication.

It is argued that basic perception is essentially a subjective symbolic representation of external reality because the phenomenal shape of basic perception is neither in a similarity relation (i.e. iconic) nor in a causally determined relation (i.e. indexical) with its referent in external reality. Instead, the relation between an immediately imperceivable external referent and the phenomenal shape as which it appears to a cognitive system is determined only by, and only to, the system itself. It is a *symbol* that can only be interpreted by a single system.

The symbol grounding problem therefore asks the fundamental question for the ultimate semantic foundation of basic perception. The field of embodied cognition suggests various answers. Due to the fact that many of the theories in this field express criticism towards symbolic mental representation, special attention is paid to exposing the conception of mental representation that is concerned by this critique and its difference to symbolic mental representation in the above sense.

It can be shown that embodied cognition rejects only a very particular conception of mental representation. This conception is common in the cognitive sciences. The philosophy of mind provides *another* conception, that is *not* incompatible with embodied cognition. It can be shown that this subjective conception is essentially different from the one criticised by embodied cognition. As a consequence, subjective mental representations with intentional content enable theories of embodied cognition to provide solutions to the symbol grounding problem.

The relation between the philosophy of mind (i.e. phenomenology) and embodied cognition is also supported by a semiotic conception of mental representation. According to semiotics, mental representations are signs and each sign is a ternary relation between shape, content, and referent. The content of symbolic signs is structural, and so is the content of basic perceptions. Theories from embodied cognition describe exactly this pre-conceptual structure in the most basic subjective experience.

Uncovering this similarity enables to apply the same processes that embodied cognition describes among pre-conceptual elements onto more complex representations that the cognitive system is aware of. This implements an *equipotential structure:* different functions are realised by the same basic processes at various levels of abstraction.

An explanation for the generation of basic perception explains only the most basic mental content. The mental models that cognitive systems use to conceive of their world consists of much more complex structures that span over time and multiple levels of abstraction. To infer the generation of these complex structures from basic perception, Searle's theory on intentional mental content is presented. From the structure of intentional content, requirements for its component representations are derived. These requirements accord to Peircean semiotics and include the pre-conceptual elements postulated by embodied cognition.

A semiotic formalisation of intentional mental models is developed that is based in semiotic mental representations. Examples are provided that show the difference between semiotic models of fully observable, and semiotic models of partially observable, systems. It is argued for an intricate coupling between cognitive systems and external reality. Semiotic models establish such a coupling and are therefore in line with theories from embodied cognition.

After a description of semiotic models as a formal counterpart to mental models, an algorithmic model for the generation of a semiotic model is presented. If semiotic models are considered as a feasible formalisation for mental models, then an algorithm that generates a semiotic model can be evaluated *as cognitive model:* regarding its capacity to simulate the process of generating a mental model. The procedural generation has to adhere to particular conditions in the generation of the model. These conditions serve *cognitive justifiability.*

The developed algorithm is compared to a baseline approach that satisfies most of these conditions as well. The evaluation concerns the procedurally generated semiotic model in three parts. 1) Does it share structural and functional similarities with a real mental model? 2) Does it enable better predictions than the baseline approach? 3) Does it enable better goal-directed interaction than the baseline approach?

29.2. Summary of Contribution

A set of theoretical contributions is necessary for developing an algorithmic model for the generation of mental representations. The first is resolving the implicit ambiguity in research concerning the symbol grounding problem. This ambiguity is caused by two individually appropriate but mutually exclusive conceptions of mental representation. These conceptions are made explicit and their influence on the symbol grounding problem is described.

This theoretical contribution enables to develop an algorithm that *imitates* the generation of mental representations. The evaluation of this algorithm according to the three questions in the previous section facilitates further contributions. In general, the results of the evaluation confirm *semiotic models* as a formalisation for mental models as well as *the algorithmic model* that describes the generation of mental models.

More specifically, it can be shown that the relations between representations in a semiotic model imitate the relations between mental representations. This can be exemplified with the following points. 1) The same referent is interpreted differently depending on context and the same representation can refer to different referents depending on expectation. 2) The rate of representation generation reduces over time as older representations are reused and adapted to new circumstances. 3) The overlap of representations is minimal, functionally segmenting the environment. 4) The content of abstract representations creates a context in which already established basic representations can form new relationships.

The generated representations enable to predict behaviour that depends on hidden states. In an on-line reactive scenario, this is usually achieved with order-n Markov predictors. The developed type of semiotic models outperforms this baseline approach while remaining on-line feasible.

The results are presented in a grid world test bed specifically designed for simulating mental modelling (see chapter 27). This variant of a reinforcement learning setting is *cognitively more plausible* than the traditional setting, *easily conceivable* by an outside observer, *variable* so as to increase or decrease complexity, *partially hidden* to test for the construction of hidden state representations, and applicable in *qualitative analyses,* exclusively *predictive tasks* as well as *goal-directed interaction.*

Within this test bed, the representations in a semiotic model can be shown to improve goal-directed behaviour. The same structures that enable prediction based on hidden states also enables to increase performance in reinforcement learning. Both improvements suggest that the state of a semiotic model can successfully represent the hidden state of another system.

29.2.1. Room for Improvement

Despite the promising empiric results, there are some points that can be improved given more time/resources. Room for improvement shows in the two major parts of this work: the presentation of the theoretical foundation and the practical part that follows after it.

First, and mentioned several times throughout the work, there is rather little overlap between phenomenological and artificial intelligence research. This, in turn, makes the presentation and explanation of way more theoretical groundwork from both areas necessary. In this thesis, this methodological problem is overcome in virtue of embodied cognition: it serves as a 'glue' between both fields. Eventually, however, it is desirable to connect phenomenology and artificial intelligence on a more fundamental level and without a middle man.

The dependency on fundamental basics brings with it another problem. Often times, concepts and theories are introduced without a lot of context. Although their relevance should become clear in the chapters that follow after their introduction, it is hard to justify their initial presentation straight from the start. To do this would either require the reader to be familiar with the scientific background from which they emerged or to lay out this background in detail. Both cannot be easily achieved within the frame of a thesis from only one field. Consequently, there are parts that depend on the good will of the reader that their content will later prove to be relevant.

The practical part can be improved as well. Given more time and resources, the presented algorithm should be analysed far more extensively concerning its runtime and memory characteristics. In the present work, this has been circumvented by introducing 'on-line feasibility' as a concept that combines both (see section 20.4.2). This concept enables to restrict the class of comparable algorithms considerably and it therefore serves its purpose. The relation between the type of the modelled system, the spatial and temporal requirements of the algorithm, and the resulting model itself, however, are far too strong not to be studied in more detail.

Also, the reinforcement learning framework for simulating the generation of mental models allows a far more versatile evaluation than what has been presented. The purpose of the simulation is to offer a highly variable but *standardised* system that is in a permanent coupling with an agent that tries to model it. The evaluated results in this work concern only one variant but could be gathered from numerous other combinations of properties such as to explore the limits of semiotic models and the processes that generate them in more detail. The extend of such an exploratory analysis and the development of an appropriate heuristics to choose from the multitude of parameters for an environment, however, go beyond the scope of this thesis.

The philosophical conclusion concerning genuine mental modelling eventually requires to abandon simulated environment for good. According to the final conclusion, the next step is to evaluate the algorithm with a *physical robot.* However, the time and material resources required for experiments with physical robots far exceed what can be achieved in a simulated environment. They require at least one more project of similar scale.

29.2.2. Shortcomings of this Approach

To simulate the modelling of *some thing* requires to simulate *this thing* as well. In the case of *mental* models, this is problematic: mental models describe external reality but reality cannot be conceived of independent from *our own* mental model.

In a *simulated* reality like the grid world, the system's 'external reality' is *the designer's conception* of a grid world. The system's representations can only describe *this conception* and only with regard *to this conception,* their appropriateness can be determined and evaluated.

The mind of the designer filters different aspects of external reality according to what they consider to be relevant. Aspects that are relevant to *another* system with a *different* body (i.e. the simulated agent) are simply unknown. Only reality itself could provide the agent with *all* the information that might become relevant at some point.

The filtering of aspects shows, for example, in the fact that a programmer is incapable to generate a truly erratic environment. Random events cannot be described without implying one particular probability distribution while ignoring another.[1]

This critique is more fundamental than simply remarking that the grid world is a 'toy example'. To criticise the simplicity of a simulation implies that another, a more complex, simulation might uncover shortcomings that do not occur if the task is too easy.

Strictly speaking, the generation of a mental model with intentional content *cannot be simulated.* Here, Searle's distinction between original and derived intentionality shows its practical implications. Every conception leaves out aspects of its real referent because they are dispensable only to the particular system that has this conception.

The practical problem with a simulation is that a designer can always only simulate *some* aspects of reality. However, what aspects of reality are relevant to solve a particular task should be determined by the system that needs to solve this task in the first place.

29.3. Summary of Findings

During the development of this work some unexpected insights have been made. Most important and crucial to the methodological proceeding was the fact that contemporary embodied cognition proposes solutions to the symbol grounding problem. It can even approach the hard part of the symbol grounding problem because it is not only based in natural theories on cognition but also in phenomenological perspectives on the mind and, therefore, compatible with first-person-perspective.

From the inaccessibility of external reality follows that *it cannot be conceived of* independent from a particular mental model. In a simulation, this condition cannot be satisfied because simulations are always conceived of by their designer. However, the condition can be put in formal terms. The concept of dynamic trajectories from definition 2 allows to describe systems with essentially unknown hidden state transitions. This takes into account the difficulty of the task to model an essentially inaccessible external reality. The insight finally gives a practical face to the decade old philosophical warning that cognitive simulations somehow *differ* from real cognitive systems.

Due to the resources necessary for replacing simulations with physical robots in real-world environments, this insight could not be taken into account in the remainder of this project. What can be taken account in future research, however, is that only system that interact with the real world can be considered to generate something similar to what we experience as our own mental model.

[1] Distributions can be designed to *change* (i.e. be non-stationary). According to which probability distribution the properties of the distribution change, however, must be determined at some point or it must change itself according to some determined distribution.

Beyond the theoretical findings, the empiric evaluation produced various insights as well. Among the most interesting ones is that predictive and interactive performance evolve inversely in a mature semiotic model. This finding contrasts one premise in the design of the empiric evaluation: that those structures that enable successful prediction also enable successful interaction. Although this premise could be confirmed, it appears not to be the case *in general.* Predictive and interactive performance are both comparatively high, however, slight variations in both develop inversely. This indicates a certain incompatibility between performing these two functions with the same underlying structures. It suggests that at least one of both functions or the structure itself might need to be revised.

Eventually, the failure of traditional agents in the more cognitively plausible non-episodic reinforcement learning setting (i.e. without world and model state reset and with circular goals) suggests that they did not develop with the intend to simulate cognitive processes. In contrast, the success of the presented algorithmic model for the generation of mental models in the same setting shows that it might perform processes that resemble those in real cognitive systems. The reaching of circular goals turns out to be a quite natural formalisation for knowledge transfer from one task to another under the same causal dynamics.

29.4. Real-world Implications

The advantage of a computational model for the generation of mental models is that it can be instantiated in any computer system—for example in a physical robot that interacts with external reality, like real cognitive systems. The immediate and unfiltered access that a physical robot has to external reality has an important epistemological caveat.

The robot's internal representations now represent the same *immediately inconceivable, external* reality that real cognitive systems represent in *their own* mental models. They cannot be treated like the representations in a simulation because every comparison between the robots representation, and our conception, of reality would effectively only be a comparison of two *different* representations of reality.

This provides a necessary condition for symbol grounding in artificial systems. Systems that interact with something which has already been conceived of (e.g. simulated environments), can only have symbolic representations with content that is *derived* from original intentional content. Only symbolic representations that are *grounded in potentially undiscovered external reality* can have original intentional content.

A system that is able to reactively generate partially observable models from real world systems also has important practical implications. Any control system can be extended with a module that generates a semiotic model of the controlled system on-the-fly. Empiric results suggest that existing systems that are based on order-n Markov predictors can be considerably improved by switching over to semiotic models. This also applies to the non-interactive case. Situations that require real-time predictions and model adaptations can benefit from semiotic models as well. Possible applications are traffic analyses, the prediction of user interaction, unsupervised sequence clustering, or the implementation of artificial opponents in computer games that adapt to player behaviour.

29.5. Avenues of Further Research

The insights provided by this project allow several routes of further research. One point is indicated by the application of semiotic models as a tool for *discrete normative* predictions. The base representations can be easily modified, however, to perform *continuous regressive* prediction. This could be implemented with various methods, one that suggests itself due to ease of implementation and low complexity requirements is linear regression. Applied to sequences of rational numbers, semiotic models could enable to identify repeating segments of linear development, repeating segments of *those* segments, and so on. This could prove to be useful in the technical analysis of financial exchange charts.

The number of the model's parameters was kept minimal on purpose. It would be interesting, however, to explore modifications, where these parameters change similar to an ϵ-decreasing strategy in reinforcement learning. At first, the model is very tolerant towards deviations between referents and representations. But with each step in time the parameters change towards a more precise representation of the environment. Modifications in the model parameters can also be made dependent on the current level. An intuitive approach would be to increase representation tolerance with each additional level.

Another loose thread is the implementation of different modalities as O'Regan and Noë describe it in section 8.3. According to them, different types of perceptual modes emerge not from different sensorimotor interfaces with a physical environment but from different contingencies that show in the information that passes this interface. An according adaptation of the present algorithm would be to decompose the sensor emissions that the agent receives and treat each of it as an individual sensor emission. Each of these component emissions can be integrated into, and predicted from, the same semiotic model. From different sensors, different causal relations should emerge between each type of component emission. These different contingencies should reflect in *modal regions* of the semiotic model, where there are representations that are specific to each mode of perception.

For the particular case of a comprehensive cognitive model for the generation of mental representations, various extensions and improvements are necessary as well. The most pressing one is the implementation of *forgetting.* The removal of unused representations enables a more adaptive model because it avoids the need to meticulously adapt representations to new and radically different situations. It avoids the need to search through a plethora of representations that were generated in the past. It also introduces a stronger generalisation bias towards representations that can be applied to various different referents that appear under similar circumstances.

Appendices

A. The Centrifugal Governor

```python
#!/usr/bin/env python3

f_friction = .05
v_friction = .2

state = (50., 80.)

print("time, flywheel, valve")
for time_index in range(20):
    print("{}, {:.2f}, {:.2f}".format(time_index + 1, *state))
    flywheel, valve = state
    next_flywheel = flywheel + (1. - f_friction) * (90 - valve -
    ↪  flywheel)
    next_valve = valve + (1. - v_friction) * (90 - flywheel - valve)
    state = next_flywheel, next_valve
```

Appendix A has been referenced in footnote 1 on page 129.

B. Simulating the Centrifugal Governor

```python
#!/usr/bin/env python3

flywheel_model = lambda x: -.89 * x + 83.25
valve_model = lambda x: -.63 * x + 79.58

state = (50., 80.)

print("time, flywheel, valve")
for time_index in range(1, 21):
    state = flywheel_model(state[1]), valve_model(state[0])
    print("{}, {:.2f}, {:.2f}".format(time_index + 1, *state))
```

Appendix B has been referenced in footnote 4 on page 131.

C. Emissions from a Simulated Centrifugal Governor

Time t	Flywheel f_t	Valve v_t
2	12.05	48.08
3	40.46	71.99
4	19.18	54.09
5	35.11	67.50
6	23.18	57.46
7	32.11	64.98
8	25.42	59.35
9	30.43	63.57
10	26.68	60.41
11	29.48	62.77
12	27.38	61.00
13	28.96	62.33
14	27.78	61.34
15	28.66	62.08
16	28.00	61.52
17	28.49	61.94
18	28.12	61.63
19	28.40	61.86
20	28.19	61.69
21	28.35	61.82

Table C.1.: Emissions from a Simulated Centrifugal Governor.

Appendix C has been referenced in footnote 3 on page 131.

D. Generating an Erratic Linear Sequence

```python
#!/usr/bin/env python3

import random

def sequence_generator(switch_prob=.2, iterate=-1):
    i = 0

    last_element = 0
    state = 1

    while True:
        yield None, last_element
        i += 1
        if 0 < iterate <= i:
            raise StopIteration()
        elif random.random() < switch_prob:
            state *= -1
        last_element = (last_element + state) % 10
```

Appendix D has been referenced in chapter 25 on page 177.

Abbreviations

CP Charles Peirce (1931-1935). *Collected Papers of Charles Peirce.* Ed. by Charles Hartshorne. Ed. by Paul Weiss. 1-6 vols. Cambridge, Mass.: Harvard University Press.

CP Charles Peirce (1958). *Collected Papers of Charles Peirce.* Ed. by Arthur W. Burks. 7-8 vols. Cambridge, Mass.: Harvard University Press.

CPR Immanuel Kant (2010-2013). *Critique of Pure Reason.* An electronic classics series publication. Ed. by Jim Manis. The Cambridge Edition of the Works of Immanuel Kant. Translated by John Meiklejohn. New York, NY: The Pennsylvania State University.

KRV Immanuel Kant (1998). *Kritik der reinen Vernunft.* Ed. by Jim Manis. Philosophische Bibliothek Band 505. Felix Meiner Verlag Hamburg.

Bibliography

Agre, Philip and David Chapman (1987). ‘Pengi: An Implementation of a Theory of Activity’. In: *AAAI*. Ed. by Kenneth Forbus and Howard Shrobe. Vol. 87. 4, pp. 286–272 (cit. on p. 58).

Ahmed, Ayesha and Ted Ruffman (1998). ‘Why do infants make A not B errors in a search task, yet show memory for the location of hidden objects in a nonsearch task?’ In: *Developmental Psychology* 34.3, p. 441 (cit. on p. 52).

Austen, Jane (1998). *Pride and Prejudice*. Vol. 1342. URL: ftp://uiarchive.cso.uiuc.edu/pub/etext/gutenberg/etext98/pandp10.zip (cit. on pp. 184 sq.).

Baars, Bernard (1993). *A cognitive theory of consciousness*. Cambridge University Press (cit. on p. 14).

Baillargeon, Renée and Julie DeVos (1991). ‘Object permanence in young infants: Further evidence’. In: *Child development* 62.6, pp. 1227–1246 (cit. on p. 52).

Barsalou, Lawrence (1999). ‘Perceptual Symbol Systems’. In: *Behavioral and Brain Sciences* 22.04, pp. 577–660 (cit. on pp. 64–67).

Bense, Max (1975). *Semiotische Prozesse und Systeme*. Baden-Baden (cit. on p. 101).

Bielecka, Krystyna (2015). ‘Why Taddeo and Floridi did not solve the symbol grounding problem’. In: *Journal of Experimental & Theoretical Artificial Intelligence* 27.1, pp. 79–93 (cit. on p. 42).

Block, Ned (1996). ‘Mental paint and mental latex’. In: *Philosophical issues*, pp. 19–49 (cit. on p. 51).

Boden, Margaret (1988). *Computer Models of Mind: Computational Approaches in Theoretical Psychology*. Problems in the Behavioural Sciences. Cambridge University Press (cit. on p. 78).

Borodin, Allan and Ran El-Yaniv (2005). *Online computation and competitive analysis*. cambridge university press (cit. on p. 150).

Bower, Thomas (1982). *Development In Infancy.* A Series Of Books In Psychology. San Francisco: W.H. Freeman (cit. on p. 52).

Brentano, Franz (1874). *Psychologie vom empirischen Standpunkte.* Psychologie vom empirischen Standpunkte Bd. 1. Duncker & Humblot (cit. on pp. 79 sqq., 83, 86, 90 sq.).

— (2012). *Psychology from an Empirical Standpoint.* International Library of Philosophy. Taylor & Francis (cit. on pp. 5, 79 sqq., 83 sq., 86, 95).

Brooks, Rodney (1990). 'Elephants don't play chess'. In: *Robotics and autonomous systems* 6.1, pp. 3–15 (cit. on pp. 58 sqq.).

— (1991). 'New approaches to robotics'. In: *Science* 253.5025, pp. 1227–1232 (cit. on pp. 57 sq., 60, 114).

— (1999). *Cambrian intelligence: the early history of the new AI.* Mit Press (cit. on pp. 59 sq., 62, 67, 111).

Buchbinder, Niv, Shahar Chen, Joshep Naor, and Ohad Shamir (2012). 'Unified algorithms for online learning and competitive analysis'. In: *Conference on Learning Theory*, pp. 5–1 (cit. on p. 150).

Calvo, Paco and Toni Gomila (2008). *Handbook of cognitive science: An embodied approach.* Elsevier (cit. on p. 50).

Cangelosi, Angelo, Alberto Greco, and Stevan Harnad (2002). 'Symbol grounding and the symbolic theft hypothesis'. In: *Simulating the evolution of language.* Springer, pp. 191–210 (cit. on p. 94).

Čertickỳ, Michal (2013). 'Action models and their induction'. In: *Organon F* 20.2, pp. 206–215 (cit. on p. 108).

Cesa-Bianchi, Nicolo and Gábor Lugosi (2006). *Prediction, learning, and games.* Cambridge university press (cit. on pp. 119 sq.).

Cesa-Bianchi, Nicolo, Gábor Lugosi, et al. (1999). 'On prediction of individual sequences'. In: *The Annals of Statistics* 27.6, pp. 1865–1895 (cit. on p. 119).

Chalmers, David (1992). 'Subsymbolic computation and the Chinese room'. In: *The symbolic and connectionist paradigms: Closing the gap*, pp. 25–48 (cit. on p. 41).

Cheeseman, Peter (2001). 'Probability, Foundations of'. In: *The MIT Encyclopedia of the Cognitive Sciences.* Ed. by Robert Wilson and Frank Keil. MIT press (cit. on p. 124).

Chen, Stanley and Joshua Goodman (1996). 'An empirical study of smoothing techniques for language modeling'. In: *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 310–318 (cit. on p. 125).

Chrisman, Lonnie (1992). 'Reinforcement learning with perceptual aliasing: The perceptual distinctions approach'. In: *AAAI*, pp. 183–188 (cit. on p. 175).

Clowes, Robert (2007). 'Semiotic symbols and the missing theory of thinking'. In: *Interaction Studies* 8.1, pp. 105–124 (cit. on p. 95).

Cole, David (2010). 'Against Derived Intentionality' (cit. on p. 78).

Crane, Tim (1995). *The mechanical mind: a philosophical introduction to minds, machines and mental representation*. Penguin philosophy. Penguin (cit. on pp. 84 sq.).

— (1998). 'Intentionality as the mark of the mental'. In: *Royal Institute of Philosophy Supplement* 43, pp. 229–251 (cit. on pp. 78, 81).

— (2000). 'The Origins of Qualia'. In: *The History of the Mind-Body Problem*. Routledge (cit. on p. 96).

— (2001). 'Intentional objects'. In: *Ratio* 14.4, pp. 336–349 (cit. on pp. 27 sq.).

— (2006). 'Brentano's concept of intentional inexistence'. In: *The Austrian contribution to analytic philosophy* 1, p. 20 (cit. on pp. 80, 84).

— (2013). 'The Given'. In: *Mind, Reason and Being-in-the-World: the McDowell-Dreyfus Debate*. Ed. by Joseph Schear. Routledge, pp. 229–249 (cit. on pp. 34 sq.).

— (2014). *Aspects of Psychologism*. Harvard University Press (cit. on p. 96).

Crook, Paul (2007). 'Learning in a state of confusion: Employing active perception and reinforcement learning in partially observable worlds'. In: (cit. on p. 174).

Crook, Paul and Gillian Hayes (2003). 'Learning in a state of confusion: Perceptual aliasing in grid world navigation'. In: *Towards Intelligent Mobile Robots* 4 (cit. on p. 159).

Dennett, Daniel (1981). *Brainstorms: Philosophical essays on mind and psychology*. 8. Mit press (cit. on p. 13).

— (1989). *The intentional stance*. Bradford Books. MIT Press (cit. on p. 78).

Dretske, Fred (1997). *Naturalizing the mind*. mit Press (cit. on p. 51).

Dreyfus, Hubert (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA, USA: MIT Press (cit. on pp. 12, 18, 95, 121, 169, 207).

Dreyfus, Hubert (2007). 'Why Heideggerian AI failed and how fixing it would require making it more Heideggerian'. In: *Philosophical psychology* 20.2, pp. 247–268 (cit. on p. 59).

Drummond, John (2012). 'Intentionality without representationalism'. In: *The Oxford Handbook of Contemporary Phenomenology, Oxford: Oxford University Press* 115, p. 133 (cit. on p. 78).

Eban, Elad, Aharon Birnbaum, Shai Shalev-Shwartz, and Amir Globerson (2012). 'Learning the experts for online sequence prediction'. In: *arXiv preprint arXiv:1206.4604* (cit. on p. 120).

Eells, Ellery (1999). 'Probability'. In: *The Cambridge dictionary of philosophy.* Ed. by Robert Audi. Second Edition. Cambridge University Press (cit. on p. 124).

Ernst, Damien, Pierre Geurts, and Louis Wehenkel (2005). 'Tree-based batch mode reinforcement learning'. In: *Journal of Machine Learning Research* 6.Apr, pp. 503–556 (cit. on p. 150).

Feder, Meir, Neri Merhav, and Michael Gutman (1992). 'Universal prediction of individual sequences'. In: *IEEE transactions on Information Theory* 38.4, pp. 1258–1270 (cit. on p. 119).

Feldman, Jerome and Dana Ballard (1982). 'Connectionist models and their properties'. In: *Cognitive science* 6.3, pp. 205–254 (cit. on p. 13).

Fish, Stanley (1982). 'With the Compliments of the Author: Reflections on Austin and Derrida'. In: *Critical Inquiry* 8.4, pp. 693–721 (cit. on p. 84).

Fodor, Jerry (1980). 'Methodological solipsism considered as a research strategy in cognitive psychology'. In: *Behavioral and brain sciences* 3.01, pp. 63–73 (cit. on pp. 26, 33).

— (1983). *The modularity of mind: An essay on faculty psychology.* MIT press (cit. on p. 78).

Forrester, Jay (1971). 'Counterintuitive behavior of social systems'. In: *Theory and Decision* 2.2, pp. 109–140 (cit. on pp. 76 sq.).

Frege, Gottlob (1892). 'Über Sinn und Bedeutung'. In: *Zeitschrift für Philosophie und philosophische Kritik* 100, pp. 25–50 (cit. on pp. 19, 83, 85, 90).

Gelfond, Michael and Vladimir Lifschitz (1998). 'Action languages'. In: (cit. on p. 118).

Ghahramani, Zoubin (2004). ‘Unsupervised Learning’. In: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures.* Ed. by Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch. Springer Berlin Heidelberg, pp. 72–112 (cit. on p. 173).

Gibbs, Raymond (2006). *Embodiment and cognitive science.* Cambridge University Press (cit. on p. 50).

Gibson, James (1986). *The ecological approach to visual perception.* Psychology Press (cit. on pp. 55 sqq., 62, 67, 70, 101).

Glenberg, Arthur (1997). ‘What memory is for: Creating meaning in the service of action’. In: *Behavioral and brain sciences* 20.01, pp. 41–50 (cit. on pp. 64, 67).

Glenberg, Arthur and Michael Kaschak (2002). ‘Grounding language in action’. In: *Psychonomic bulletin & review* 9.3, pp. 558–565 (cit. on pp. 67 sq.).

Gomes, Antônio, Ricardo Gudwin, and João Queiroz (2005). ‘Meaningful agents: a semiotic approach’. In: *Proceedings of the International Conference on Integration of Knowledge Intensive Multi-Agent Systems–Modeling, Evolution, and Engineering, Waltham, MA*, pp. 399–404 (cit. on p. 94).

Goodman, Nelson (1976). *Languages of art: An approach to a theory of symbols.* Hackett publishing (cit. on p. 51).

Harnad, Stevan (1989). ‘Minds, machines and Searle’. In: *Journal of Experimental & Theoretical Artificial Intelligence* 1.1, pp. 5–25 (cit. on pp. 16, 18, 22 sq., 40).

— (1990). ‘The symbol grounding problem’. In: *Physica D: Nonlinear Phenomena* 42.1, pp. 335–346 (cit. on pp. 7, 15 sq., 18–23, 27, 36–40, 42, 49 sq., 62, 73, 78, 100 sq., 169).

— (2001). ‘Mind, Machines and Searle II: What’s Wrong and Right About Searle’s Chinese Room Argument?’ In: (cit. on p. 23).

Haugeland, John (1987). ‘An Overview of the Frame Problem’. In: *The Robot’s Dilemma* (cit. on p. 14).

— (1993). ‘Mind Embodied and Embedded’. In: *Mind and Cognition: 1993 International Symposium.* Ed. by John Haugeland. Academica Sinica, pp. 233–267 (cit. on p. 72).

Hausknecht, Matthew and Peter Stone (2015). ‘Deep recurrent q-learning for partially observable mdps’. In: *CoRR, abs/1507.06527* (cit. on p. 159).

Hazan, Elad (2016). 'Introduction to Online Convex Optimization'. In: *Foundations and Trends in Optimization* 2.3-4, pp. 157–325. ISSN: 2167-3888. DOI: 10.1561/2400000013. URL: http://dx.doi.org/10.1561/2400000013 (cit. on p. 120).

Hengst, Bernhard (2010). 'Hierarchical Reinforcement Learning'. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey Webb. Boston, MA: Springer US, pp. 495–502. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_363. URL: https://doi.org/10.1007/978-0-387-30164-8_363 (cit. on p. 158).

Honeychurch, Sarah (2013). *Duck-rabbit*. URL: http://nomadwarmachine.wordpress.com/2013/10/26/duck-rabbit/ (visited on 11/07/2014) (cit. on p. 17).

Hurley, Susan (2002). *Consciousness in Action*. Harvard University Press (cit. on p. 114).

Hyslop, Alec (2013). *Other minds*. Vol. 246. Springer Science & Business Media (cit. on p. 28).

Hyslop, Alec and Frank Jackson (1972). 'The analogical inference to other minds'. In: *American Philosophical Quarterly* 9.2, pp. 168–176 (cit. on p. 28).

James, William (1892). 'The Stream of Consciousness'. In: *Psychology*. Ed. by William James (cit. on p. 97).

Johnson-Laird, Philip (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cognitive science series. Harvard University Press (cit. on p. 77).

— (1991). *Mental Models*. Ed. by Michael Posner. A Bradford Book. MIT Press (cit. on p. 77).

Kant, Immanuel (2010-2013). *Critique of Pure Reason*. An electronic classics series publication. Ed. by Jim Manis. The Cambridge Edition of the Works of Immanuel Kant. Translated by John Meiklejohn. New York, NY: The Pennsylvania State University (cit. on pp. 24, 28, 100).

— (1998). *Kritik der reinen Vernunft*. Ed. by Jim Manis. Philosophische Bibliothek Band 505. Felix Meiner Verlag Hamburg (cit. on pp. 6, 10, 24 sq., 28).

Karp, Richard (1992). 'On-line algorithms versus off-line algorithms: How much is it worth to know the future?' In: *IFIP Congress (1)*. Vol. 12, pp. 416–429 (cit. on p. 149).

Kenaan, Hagi (2002). 'Language, philosophy and the risk of failure: rereading the debate between Searle and Derrida'. In: *Continental Philosophy Review* 35.2, pp. 117–133 (cit. on p. 84).

Kobayashi, Yasushi and Tadashi Isa (2002). 'Sensory-motor gating and cognitive control by the brainstem cholinergic system'. In: *Neural Networks* 15.4, pp. 731–741 (cit. on p. 55).

Audi, Robert, ed. (1999). *The Cambridge dictionary of philosophy.* Cambridge University Press. Chap. phenomenology (cit. on p. 3).

Kosslyn, Stephen (1980). *Image and mind.* Harvard University Press (cit. on p. 51).

Kosslyn, Stephen and James Pomerantz (1977). 'Imagery, propositions, and the form of internal representations'. In: *Cognitive Psychology* 9.1, pp. 52–76 (cit. on p. 51).

Kriegel, Uriah (2013). 'Two notions of mental representation'. In: *Current Controversies in Philosophy of Mind*, p. 161 (cit. on p. 27).

Kuvayev, Leonid and Richard Sutton (1997). *Model-based reinforcement learning.* Tech. rep. Technical report university of massachusetts, Department of computer science (cit. on p. 155).

Lagoudakis, Michail and Ronald Parr (2003). 'Least-squares policy iteration'. In: *Journal of machine learning research* 4.Dec, pp. 1107–1149 (cit. on p. 150).

Lakoff, George and Mark Johnson (1980). *Metaphors we Live by.* Chicago: University of Chicago Press (cit. on pp. 63 sq.).

— (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* Collection of Jamie and Michael Kassler. Basic Books (cit. on pp. 62 sqq.).

Lewis, David (1971). 'Analog and digital'. In: *Noûs*, pp. 321–327 (cit. on p. 51).

Lin, Long-Ji (1992). 'Self-improving reactive agents based on reinforcement learning, planning and teaching'. In: *Machine learning* 8.3/4, pp. 69–97 (cit. on pp. 150, 159).

Lin, Long-Ji and Tom Mitchell (1993). 'Reinforcement learning with hidden states'. In: *From animals to animats* 2, pp. 271–280 (cit. on p. 159).

Littlestone, Nick and Manfred Warmuth (1994). 'The weighted majority algorithm'. In: *Information and computation* 108.2, pp. 212–261 (cit. on p. 119).

Lyre, Holger (2013). 'Verkörperlichung und situative Einbettung'. In: *Handbuch Kognitionswissenschaft.* Ed. by Achim Stephan and Sven Walter. Metzler, Stuttgart. Chap. III.7, pp. 186–193 (cit. on p. 50).

MacDorman, Karl (1997). 'How to ground symbols adaptively'. In: *Two Sciences of Mind: Readings in Cognitive Science and Consciousness* 9, pp. 135–178 (cit. on pp. 16 sqq., 66, 169).

Maturana, Humberto (1980). *Autopoiesis and cognition: The realization of the living.* 42. Springer (cit. on pp. 122, 134).

May, Mark (1996). 'Mentales Modell'. In: *Wörterbuch der Kognitionswissenschaft.* Ed. by Gerhard Strube. Klett-Cotta (cit. on p. 76).

McCallum, Andrew (1993). 'Overcoming incomplete perception with utile distinction memory'. In: *Proceedings of the Tenth International Conference on Machine Learning*, pp. 190–196 (cit. on p. 159).

— (1995). 'Instance-based state identification for reinforcement learning'. In: *Advances in Neural Information Processing Systems*, pp. 377–384 (cit. on p. 159).

— (1996a). 'Hidden state and reinforcement learning with instance-based state identification'. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 26.3, pp. 464–473 (cit. on p. 159).

— (1996b). 'Reinforcement learning with selective perception and hidden state'. PhD thesis. University of Rochester. Dept. of Computer Science (cit. on pp. 159 sq.).

McCarthy, John and Patrick Hayes (1969). 'Some Philosophical Problems from the Standpoint of Artificial Intelligence'. In: *Machine Intelligence.* Ed. by Matthew Ginsberg. Vol. 4, pp. 463–502 (cit. on p. 13).

McCorduck, Pamela (2004). *Machines who think: a personal inquiry into the history and prospects of artificial intelligence.* Ak Peters Series. A.K. Peters (cit. on p. 12).

McGinn, Colin (2004). *Consciousness and its objects.* Oxford University Press (cit. on p. 29).

Melnyk, Andrew (1994). 'Inference to the best explanation and other minds'. In: (cit. on p. 28).

Millikan, Ruth (2004). *Varieties of Meaning: The 2002 Jean Nicod Lectures.* Bradford books. MIT Press. ISBN: 9780262134446 (cit. on p. 57).

— (2006). 'Useless Content'. In: *Teleosemantics.* Ed. by Grahan Macdonald and David Papineau (cit. on p. 57).

Mitchell, Tom (1990). 'The need for biases in learning generalizations'. In: *Readings in Machine Learning*, pp. 184–191 (cit. on pp. 145, 170).

— (1997). *Machine Learning.* McGraw-Hill International Editions. McGraw-Hill (cit. on p. 142).

Mountcastle, Vernon (1978). 'An organizing principle for cerebral function: the unit model and the distributed system'. In: *The Mindful Brain.* Ed. by Gerald Edelman and Vernon Mountcastle. Cambridge, Mass.: MIT Press (cit. on pp. 139, 170).

Nes, Anders (2008). 'Are only mental phenomena intentional?' In: *Analysis* 68.3, pp. 205–215 (cit. on p. 81).

Newell, Allan (1988). 'Putting it all together'. In: Klahr, David and Kenneth Kotovsky. *Complex Information Processing. The Impact of Herbert A. Simon.* Ed. by David Klahr and Kenneth Kotovsky. Chap. 15 (cit. on pp. 31 sq.).

Newell, Allen (1980). 'Physical Symbol Systems'. In: *Cognitive science* 4.2, pp. 135–183 (cit. on pp. 32 sq.).

Newell, Allen and Herbert Simon (1976). 'Computer Science As Empirical Inquiry: Symbols and Search'. In: *Communications of the ACM* 19.3, pp. 113–126 (cit. on p. 31).

O'Regan, Kevin and Alva Noë (2001). 'A sensorimotor account of vision and visual consciousness'. In: *Behavioral and brain sciences* 24.05, pp. 939–973 (cit. on pp. 68–71, 74, 101, 215).

Pacherie, Élisabeth (2000). 'The content of intentions'. In: *Mind & Language* 15.4, pp. 400–432 (cit. on p. 90).

Peirce, Charles (1931-1935). *Collected Papers of Charles Peirce.* Ed. by Charles Hartshorne. Ed. by Paul Weiss. 1-6 vols. Cambridge, Mass.: Harvard University Press (cit. on pp. 5, 94–97, 99 sqq., 109).

— (1958). *Collected Papers of Charles Peirce.* Ed. by Arthur W. Burks. 7-8 vols. Cambridge, Mass.: Harvard University Press (cit. on p. 95).

— (1984). In: *Writings of Charles S. Peirce.* Ed. by Edward Moore. Ed. by Max Fisch. Ed. by Christian Kloesel. Ed. by Don Roberts. Vol. 2. 1-6 vols. Bloomington and Indianapolis: Indiana University Press (cit. on pp. 6, 95 sq., 100 sq., 103, 209).

Piaget, Jean (2013). *The construction of reality in the child.* Vol. 82. Routledge (cit. on p. 52).

Pitt, David (2013). 'Mental Representation'. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward Zalta. Fall 2013 (cit. on p. 30).

Pope, Kenneth (2013). *The stream of consciousness: Scientific investigations into the flow of human experience.* Springer Science & Business Media (cit. on p. 97).

Pope, Kenneth and Jerome Singer (1978). 'Regulation of the stream of consciousness: Toward a theory of ongoing thought'. In: *Consciousness and self-regulation.* Springer, pp. 101–137 (cit. on p. 97).

Potter, Mary, Brad Wyble, Carl Hagmann, and Emily McCourt (2014). 'Detecting meaning in RSVP at 13 ms per picture'. In: *Attention, Perception, & Psychophysics* 76.2, pp. 270–279 (cit. on p. 97).

Putnam, Hilary (1975). 'The Meaning of 'Meaning''. In: *Minnesota Studies in the Philosophy of Science* 7, pp. 131–193 (cit. on p. 33).

Pylyshyn, Zenon (1977). 'What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery'. In: *Images, Perception, and Knowledge.* Ed. by John Nicholas. Vol. 8. The University of Western Ontario Series in Philosophy of Science. Springer Netherlands, pp. 1–36 (cit. on p. 51).

— (1979). 'The rate of "mental rotation" of images: A test of a holistic analogue hypothesis'. In: *Memory & Cognition* 7.1, pp. 19–28 (cit. on p. 51).

Rabiner, Lawrence (1989). 'A tutorial on hidden Markov models and selected applications in speech recognition'. In: *Proceedings of the IEEE* 77.2, pp. 257–286 (cit. on p. 119).

Raffel, Stanley (2011). 'Understanding Each Other: The Case of the Derrida-Searle Debate'. In: *Human Studies* 34.3, pp. 277–292 (cit. on p. 84).

Rapaport, William (2012). *Intensionality vs. Intentionality.* Ed. by William Rapaport. URL: http://www.cse.buffalo.edu/~rapaport/intensional.html (visited on 30/01/2014) (cit. on pp. 84 sq.).

Raymond, Jane, Kimron Shapiro, and Karen Arnell (1992). 'Temporary suppression of visual processing in an RSVP task: An attentional blink?' In: *Journal of experimental psychology: Human perception and performance* 18.3, p. 849 (cit. on p. 97).

Robinson, Rex (2012). *An introduction to dynamical systems: continuous and discrete.* Vol. 19. American Mathematical Soc. (cit. on p. 120).

Rodriguez, Dairon, Jorge Hermosillo, and Bruno Lara (2012). 'Meaning in Artificial Agents: The Symbol Grounding Problem Revisited'. In: *Minds and Machines* 22.1, pp. 25–34 (cit. on pp. 38 sq.).

Rosenschein, Stanley and Leslie Kaelbling (1986). 'The synthesis of digital machines with provable epistemic properties'. In: *Proceedings of the 1986 Conference on Theoretical aspects of reasoning about knowledge.* Ed. by Joseph Halpern. Morgan Kaufmann Publishers Inc., pp. 83–98 (cit. on p. 58).

Routledge, Robert (1900). *Discoveries and Inventions of the Nineteenth Century.* 13th ed. Studio Edns. (cit. on p. 128).

Rumelhart, David (1989). 'The Architecture of Mind: A Connectionist Approach'. In: *Foundations of Cognitive Science.* Ed. by Michael Posner. Cambridge, MA, USA: MIT Press, pp. 133–159 (cit. on p. 13).

Rummery, Gavin and Mahesan Niranjan (1994). *On-line Q-learning using connectionist systems.* Vol. 37. University of Cambridge, Department of Engineering (cit. on p. 149).

Saitta, Lorenza and Jean-Daniel Zucker (2013). 'Abstraction in Machine Learning'. In: *Abstraction in Artificial Intelligence and Complex Systems.* New York, NY: Springer New York, pp. 273–327. ISBN: 978-1-4614-7052-6. DOI: 10.1007/978-1-4614-7052-6_9. URL: http://dx.doi.org/10.1007/978-1-4614-7052-6_9 (cit. on pp. 143, 169, 173).

Schlicht, Tobias (2008). 'Ein Stufenmodell der Intentionalität'. In: *P. Spät (Hg.): Zur Zukunft der Philosophie des Geistes*, pp. 59–91 (cit. on p. 81).

Schmid, Ute and Martin Kindsmüller (1996). *Kognitive Modellierung: Eine Einführung in logische und algorithmische Grundlagen.* Spektrum Akademischer Verlag (cit. on p. 138).

Schultheis, Holger (2013). 'Kognitive Modellierung'. In: *Handbuch Kognitionswissenschaft.* Ed. by Achim Stephan and Sven Walter. Metzler (cit. on p. 138).

Searle, John (1969). *Speech acts: An essay in the philosophy of language.* Vol. 626. Cambridge university press (cit. on p. 89).

— (1979). 'What is an intentional state?' In: *Mind* 88, pp. 74–92 (cit. on pp. 84–87, 89 sq.).

— (1980a). 'Intrinsic intentionality'. In: *Behavioral and Brain Sciences* 3.03, pp. 450–457 (cit. on pp. 95, 109, 213).

— (1980b). 'Minds, brains, and programs'. In: *Behavioral and Brain Sciences* 3, pp. 417–424 (cit. on pp. 5, 20 sq., 23, 28, 36–42, 44, 77 sq., 87 sq., 90, 94 sq., 102 sq.).

— (1983). *Intentionality: An Essay in the Philosophy of Mind.* Cambridge paperback library. Cambridge University Press (cit. on pp. 10, 21, 26 sq., 34, 36 sq., 40 sq., 78 sq., 81, 84, 88–93, 103, 140, 209).

— (1987). 'Minds and brains without programs'. In: *Mindwaves*, pp. 209–233 (cit. on p. 41).

— (1990a). 'Is the Brain a Digital Computer?' In: *Proceedings and Addresses of the American Philosophical Association* 64.3, pp. 21–37 (cit. on p. 42).

Searle, John (1990b). ‘Is the brain’s mind a computer program?’ In: *Scientific American* 262.1, pp. 26–31 (cit. on p. 41).

— (1991). ‘Consciousness, unconsciousness and intentionality’. In: *Philosophical Issues*, pp. 45–66 (cit. on p. 91).

— (1992). *The Rediscovery of the Mind.* A Bradford book. BRADFORD BOOK (cit. on p. 91).

— (1993). ‘The failures of computationalism’. In: *Think* 2, pp. 68–71 (cit. on pp. 7, 23, 38).

Shalev-Shwartz, Shai (2012). ‘Online learning and online convex optimization’. In: *Foundations and Trends in Machine Learning* 4.2, pp. 107–194 (cit. on p. 120).

Shani, Guy (2007). *Learning and solving partially observable markov decision processes.* Ben Gurion University (cit. on p. 149).

Shapiro, Lawrence (2011). *Embodied cognition.* Routledge London (cit. on pp. 36, 39, 50, 53).

Short, Thomas (1981). ‘Semeiosis and intentionality’. In: *Transactions of the Charles S. Peirce Society*, pp. 197–223 (cit. on p. 95).

— (2007). *Peirce’s theory of signs.* Cambridge University Press (cit. on p. 95).

Smith, Brian (1986). ‘Is computation formal’. unpublished (cit. on p. 78).

Smith, Linda and Esther Thelen (2003). ‘Development as a dynamic system’. In: *Trends in cognitive sciences* 7.8, pp. 343–348 (cit. on p. 52).

Sörensen, Bent, Torkild Thellefsen, and Sören Brier (2012). ‘Mind, Matter, and Evolution: An Outline of C. S. Peirce’s Evolutionary Cosmogony.’ In: *Cybernetics and Human Knowing* 19.1-2, pp. 95–120 (cit. on p. 95).

Stachenfeld, Kimberly, Matthew Botvinick, and Samuel Gershman (2017). ‘The hippocampus as a predictive map’. In: *bioRxiv*, p. 097170 (cit. on p. 71).

Steels, Luc (1999). *The Talking Heads Experiment. Volume 1. Words and Meanings.* Antwerpen: Laboratorium (cit. on pp. 27, 37, 40, 50, 62).

— (2008). ‘The Symbol Grounding Problem Has Been Solved. So What’s Next?’ In: *Symbols and Embodiment: Debates on Meaning and Cognition.* Ed. by Manuel de Vega. Oxford: Oxford University Press. Chap. 12 (cit. on pp. 7, 37, 39, 95).

Strasser, Anna (2011). ‘Mental Representation’. In: *Encyclopedia of the Sciences of Learning.* Ed. by Norbert Seel. Encyclopedia of the Sciences of Learning. Springer US (cit. on p. 29).

Strube, Gerhard (1996a). ‘Kognitionswissenschaft’. In: *Wörterbuch der Kognitionswissenschaft.* Ed. by Gerhard Strube. Klett-Cotta (cit. on p. 138).

— (1996b). ‘Kognitive Modellierung’. In: *Wörterbuch der Kognitionswissenschaft.* Ed. by Gerhard Strube. Klett-Cotta (cit. on pp. 76, 138).

— (1996c). ‘PSSH’. In: *Wörterbuch der Kognitionswissenschaft.* Ed. by Gerhard Strube. Klett-Cotta (cit. on p. 138).

Sun, Ron (2000). ‘Symbol grounding: a new look at an old idea’. In: *Philosophical Psychology* 13.2, pp. 149–172 (cit. on pp. 7, 39, 42, 94).

Sun, Ron and Chad Sessions (2000). ‘Self-segmentation of sequences: automatic formation of hierarchies of sequential behaviors’. In: *Systems, Man, and Cybernetics, Part B: Cybernetics* 30.3, pp. 403–418 (cit. on p. 159).

Sutton, Richard (1990). ‘Integrated architectures for learning, planning, and reacting based on approximating dynamic programming’. In: *Proceedings of the seventh international conference on machine learning*, pp. 216–224 (cit. on pp. 197, 202).

Sutton, Richard and Andrew Barto (1998). *Reinforcement learning: An introduction.* Vol. 1. 1. MIT press Cambridge (cit. on pp. 149, 194).

Szabó, Zoltán (2017). ‘Compositionality’. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward Zalta. Summer 2017. Metaphysics Research Lab, Stanford University (cit. on p. 63).

Taddeo, Mariarosaria and Luciano Floridi (2005). ‘Solving the symbol grounding problem: a critical review of fifteen years of research’. In: *Journal of Experimental & Theoretical Artificial Intelligence* 17.4, pp. 419–445 (cit. on p. 39).

Takahashi, Tatsuji, Kuratomo Oyo, and Shuji Shinohara (2009). ‘A loosely symmetric model of cognition’. In: *European Conference on Artificial Life.* Springer, pp. 238–245 (cit. on p. 147).

Tegtmeier, Erwin (2005). ‘Intentionality is not Representation’. In: *Metaphysica* 6, pp. 77–84 (cit. on p. 78).

Thelen, Esther, Gregor Schöner, Christian Scheier, and Linda Smith (2001). ‘The dynamics of embodiment: A field theory of infant perseverative reaching’. In: *Behavioral and brain sciences* 24.01, pp. 1–34 (cit. on pp. 52 sqq., 57, 62).

Thelen, Esther and Linda Smith (1996). *A dynamic systems approach to the development of cognition and action.* MIT press (cit. on p. 51).

Tool (2001). *Lateralus*. Ed. by Zoo Entertainment. Compact Disc.

Turing, Alan (1950). 'Computing machinery and intelligence'. In: *Mind* 59.236, pp. 433–460 (cit. on p. 21).

Van Gelder, Tim (1989). 'Distributed Representation'. PhD thesis. University of Pittsburgh (cit. on p. 98).

— (1995). 'What might cognition be, if not computation?' In: *The Journal of Philosophy*, pp. 345–381 (cit. on pp. 110, 129).

— (1999). 'Distributed vs. Local Representation'. In: *The MIT Encyclopedia of the Cognitive Sciences*. Ed. by Robert Wilson and Frank Keil. Mit Press (cit. on pp. 38, 40, 134).

Van Orman Quine, Willard (1960). *Word and Object*. MIT Press paperback series. Technology Press of the Massachusetts Inst. of Technology (cit. on pp. 83 sq.).

— (1980). *From a Logical Point of View: 9 Logico-philosophical Essays*. Logico-philosophical essays. Harvard University Press (cit. on p. 83).

Vogt, Paul (2001). 'Symbol grounding in communicative mobile robots'. In: *Proceedings of the AAAI Fall Symposium on Anchoring Symbols to Sensor Data* (cit. on p. 94).

— (2002). 'The physical symbol grounding problem'. In: *Cognitive Systems Research* 3.3, pp. 429–457 (cit. on p. 94).

Vogt, Paul and Federico Divina (2007). 'Social symbol grounding and language evolution.' In: *Interaction Studies* 8.1 (cit. on p. 95).

Von Eckardt, Barbara (1999). 'Mental Representation'. In: *The MIT Encyclopedia of the Cognitive Sciences*. Ed. by Robert Wilson and Frank Keil. Mit Press (cit. on p. 30).

Watkins, Christopher (1989). 'Learning from delayed rewards'. PhD thesis. King's College, Cambridge (cit. on p. 149).

Wilson, Margaret (2002). 'Six views of embodied cognition'. In: *Psychonomic bulletin & review* 9.4, pp. 625–636 (cit. on pp. 50, 110, 134).

Wittgenstein, Ludwig (2010). *Philosophical Investigations*. Wiley (cit. on pp. 16 sq., 169).

Wright, Edmond (1982). 'Derrida, Searle, Contexts, Games, Riddles'. In: *New Literary History* 13.3, pp. 463–477 (cit. on p. 84).

Ye, Nong (2004). *The Handbook of Data Mining*. Human Factors and Ergonomics. Lawrence Erlbaum Associates, Incorporated (cit. on p. 172).