Modélisation logique du raisonnement et de l'apprentissage : une approche bio-inspirée.

Christel Grimaud

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ LILLE NORD DE FRANCE

École doctorale Sciences de l'Homme et de la Société

Spécialité Philosophie

Laboratoire Savoirs, Textes et Langage (UMR 8163)

Présentée par

Christel GRIMAUD

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ LILLE NORD DE FRANCE

Sujet de la thèse :

Logical modelling of reasoning and learning: a bio-inspired approach.

Soutenue le 31 Mars 2016

devant le jury composé de :

M. Shahid	RAHMAN	Directeur de thèse, Université de Lille III, France
Mme. Joke	MEHEUS	Présidente du Jury, Université de Gand, Belgique
M. Andreas	HERZIG	Examinateur, Université de Toulouse III, France
M. John	SYMONS	Examinateur, Université du Kansas, États-Unis d'Amérique
M. Tero	TULENHEIMO	Examinateur, Université de Lille III, France

Pré-rapporteurs :

M. Andreas	HERZIG,	Université de Toulouse III, France
M. Robert	JACKSON,	Université de Canterbury, Nouvelle-Zélande

Abstract

Logical modelling of reasoning and learning: a bio-inspired approach

In this dissertation, we take inspiration in cognitive sciences to address the issue of the logical modelling of reasoning and learning. Our main thrust is that to address these issues one should take inspiration in the way natural agents (*i.e.*, humans and animals) actually proceed when they draw inferences and learn. Considering that reasoning incorporates a wide range of cognitive abilities, and that it would thus be unreasonable to hope to model the whole of human's reasoning all at once, we focus here on a very basic kind of inferences that, we argue, can be considered as the primary core of reasoning in all brained animals. We identify a plausible underlying process for these inferences, first at the mental level of description and then at the neural level, and we develop a family of logical models that allow to simulate it. Then we tackle the issue of providing sets of rules to characterize the inference relations induced by these models. These rules are a by-product of the posited process, and should thus be seen as rules that, according to the model, result from the very functioning of brains. Finally we examine the learning processes attached to the considered inferences, and we show how to they can be modelled within our framework. To conclude we briefly discuss possible further developments of the framework, and in particular we give indications about how the modelling of some other cognitive abilities might be envisioned.

Résumé

Modélisation logique du raisonnement et de l'apprentissage : une approche bio-inspirée.

Dans ce mémoire on s'inspire des sciences cognitives pour aborder la question de la modélisation logique du raisonnement et de l'apprentissage. Notre principale conviction est qu'il faudrait, pour traiter ce problème, prendre modèle sur la manière dont les agents naturels (c'est à dire les humains et les animaux) procèdent lorsqu'ils raisonnent ou apprennent. Considérant que le raisonnement fait appel à un grand nombre de facultés cognitives distinctes, et qu'il ne serait donc pas raisonnable d'espérer modéliser d'un seul coup l'ensemble du raisonnement humain, on se concentre ici sur un type d'inférences très simples dont on soutient qu'elles constituent le coeur du raisonnement chez tous les animaux à cerveau. On identifie un processus sous-jacent plausible pour ces inférences, d'abord au niveau mental de description, puis au niveau neuronal, et on développe une famille de modèles logiques permettant de le simuler. On s'attache ensuite à produire un ensemble de règles d'inférence caractérisant les relations d'inférence induites par ces modèles. Ces règles résultent du processus suggéré, et doivent donc être vues comme des règles qui, d'après le modèle, émergent fonctionnement des cerveaux. Enfin, on analyse les processus d'apprentissage attachés aux inférences considérées, et on montre comment le formalisme proposé permet de les modéliser. Pour conclure on évoque brièvement les possibles développements futurs du modèle, et notamment on donne quelques indications quant à la manière dont la modélisation d'un certain nombre de facultés additionnelles pourrait être envisagée.

Table des matières

Version anglaise

Version française

Version anglaise

Introduction

Logical modelling of reasoning and learning is confronted with a handful of long-lasting problems. One of them is how the modelling of reasoning and that of learning should be articulated. Indeed, the acquisition of information by a cognitive agent triggers both inferential processes and learning, that is, a revision of the agent's dispositions to infer. An agent's dispositions to infer should therefore depend on the sequence of its previously acquired informations, that is, on its past experience. Closely related with the previous one, another issue is that a suitable model of learning should afford iteration, since learning is essentially a continuous process. Yet at this time there is no consensus about how the revision of an agent's dispositions to infer should be conducted [1].

The present dissertation introduces a logical framework for the modelling of inferences that aims at opening the way for a solution to these problems. Its main thrust is that to address these issues one should take inspiration in the way natural agents (*i.e.*, humans and animals) actually proceed when they draw inferences and learn. Indeed, learning and reasoning are natively articulated in these agents, and a logic that would model their very mental processes would be naturally immune to the problem. In a first attempt to provide such a logic, we focus here on a very basic kind of inferences that, we shall argue,

1 The most emblematic example of this is certainly the long-running debate on how to iterate Alchourrón, Gardenfors and Makinson's modelling of belief revision ([Alchourrón et al., 1985]). See [Konieczny and Pino Perez, 2002] for a review of different attempts in this direction, and their respective drawbacks. See also [Freund, 2004] for an ulterior proposal.

can be considered as the primary core of reasoning in natural agents. We suggest a plausible underlying process for these inferences, and develop a family of logical models that allow to simulate it. Then we tackle the issue of providing sets of rules to characterize the inference relations induced by the models. These rules are a by-product of the posited process, and should thus be taken as reasoning rules that arise from the functioning of a brain running such a process. In this view, the introduced framework gives an insight into how it is possible that intelligent reasoning and logic emerge from a brain at work. Finally we give an analysis of the learning processes attached to the considered inferences, and we show how the suggested logical structures allow to model these processes.

Chapter 1

Motivation and intended interpretation of the logical framework

Traditionally, the approach taken in logical modelling of reasoning is more or less tacitly a normative one: it considers how ideal agents should reason, not how real agents do. Moreover, it generally provides a purely external account, in the sense that it disregards the internal (mental/neural) processes by which reasoning is completed. By contrast, the approach that shall be taken here is descriptive and internal, which means that it shall seek to model what really happens in natural agents as they reason, and to match their reasoning processes. Our first task is thus to identify these processes, and for this we shall naturally turn to cognitive sciences. This shall be the first part of the present chapter. Then, once a plausible and tractable process is identified, in the second part we shall give a first overview of the envisioned logical modelling.

1.1 The modelled mental process

When looking at reasoning in natural agents, the first thing that one may notice is that it is not a single and homogeneous process, but that it rather incorporates a wide range of cognitive abilities. Some of these can presumably be found in human beings only, as for instance reflexivity and superior language abilities, while others are more widely spread and appear at varying extents in the so-called 'higher species' (as for instance the ability to handle relations)

and finally others are pervasive throughout animal kingdom. One can safely assume that each of these abilities is supported at the neural level by some corresponding neural process. It is also patent that the abilities and processes involved in human reasoning are manifold and intricate, and for many of them, still poorly understood. It would thus be unreasonable to hope to model the whole of human's reasoning all at once. A more reasonable plan is to search for some simple, elemental and plausible process to model.

1.1.1 Automatic inferences

Among the cognitive abilities that are commonly displayed by natural agents, there is one that appears to be at the same time very basic, pervasive and essential to reasoning. It is the ability to draw non-monotonic inferences out of the available information.

In humans, the most elemental form in which it manifests itself is through those 'quick and dirty' inferences that we routinely draw based on our intuitive knowledge of things and situations, as for example when observing a bird we expect it to fly away if we attempt to come closer. An essential characteristic of these inferences is that they are not deliberate, but occur in an automatic manner. In fact, they do not even require consciousness for their processing, and they generally remain unconscious unless we bring them to consciousness by a reflexive move on ourselves. As far as introspection tells us, they do not seem to rely on verbal representation either, as is suggested by the fact that most of the time we simply 'feel' the premisses and the conclusion, without bothering to put them into words. And finally, they seem almost effortless, which makes it probable that they run on a very simple and 'cheap' process. They certainly do not constitute the whole of humans' inferential capacities — obviously, reflexivity and verbal representation support much more sophisticated reasoning — but they take part in most of the decisions and judgements we make in our everyday life. In particular, we often use their conclusions as premisses in more complex reasoning processes such as, precisely, those involving reflexivity

or language. Therefore, they appear as a simple, autonomous (*i.e.* not relying on any other ability) and fundamental (*i.e.* on which other abilities rely) piece in human reasoning.

But the ability to draw such inferences is far from being specific to humans. The drawing of inferences joins the making of decisions, all the more when it comes to automatic, unconscious and non-verbal inferences. Now all animals do make decisions, and the question is whether these are the result of a rigid reflex or of a more flexible process. If the decision process is flexible, that is, if the animal's decisions non-monotonically depend on the available information, then it can be seen as running a non-monotonic inferential process. It turns out that flexible decision making is the norm rather than the exception all across animal kingdom. Indeed, even minute-brained animals such as insects or spiders, the reasoning capacities of which are supposedly very limited, have been observed to make non-monotonic decisions depending on what they perceive from the situation at hand [1]. The ability to draw non verbally-supported unconscious non-monotonic inferences in an automatic manner therefore appears as one of the most widespread of all reasoning abilities.

1 A well studied case is that of salticid spiders' hunting strategies relative to their prey's characteristics. For instance, when stalking a spider from the genus *Scytodes*, the salticid species *Portia labiata* normally chooses to approach from the rear. As *Scytodes* species catch their preys by spitting venom-impregnated silk at them from a distance, this strategy minimizes the risk of *P. labiata* being captured and eaten by its intended prey. But *Scytodes* females usually carry their eggs in egg-bags they hold in their mouthparts, and spitting requires that they first release their egg-bags, which they are reluctant to do. Egg-carrying females *Scytodes* are therefore less prone to spit than non-egg-carrying *Scytodes*. When stalking an egg-carrying *Scytodes* female, *P. labiata* usually forgoes the detour and takes a frontal approach instead, which allows it to bite its prey in cephalic areas for a better efficiency of venom ([Jackson and Cross, 2011], pp. 130-131). Numerous examples of similar 'conditional strategies' are documented in the literature about salticids (see for instance [Bartos, 2008]). Another striking example of non-monotonic decision making in salticids (yet involving certainly more than simple non-monotonic decision making) is trial-and-error behaviours. Typically, given a situation and a goal to achieve (*e.g.*, escaping from an unsuitable position), the spider opts for some strategy A. If this succeeds, it will repeat the same strategy if placed again in a similar position. But if this fails, it will renounce the strategy A and switch to some alternative strategy B ([Jackson et al., 2001]).

I shall call this kind of inferences *automatic inferences*. Due to the profound similarities that, despite evident superficial differences, prevail across species in the general functioning of brains, there is no reason to suppose that the drawing of these inferences is realized in a very different way from one species to another. One may thus imagine that the ability to draw automatic inferences relies on an autonomous process, simple enough to fit in smallest brains, and that it forms the primary core of reasoning in natural agents. By contrast, the other cognitive abilities would appear as additional modules, that plug into it to enhance reasoning in higher species.

1.1.2 A plausible process for automatic inferences

To identify the process that underlies the drawing of automatic inferences in natural agents, one may rely on the intuition one has of its own reasoning processes (*i.e.*, appeal to introspection), or look at the neural machinery that supports them. These two methods are not exclusive, and even they usefully supplement each other.

Introspection

Introspection for its part suggests that to draw automatic inferences, we rely on the mental representations we have of the objects and situations that we know. For example, if we have to decide whether a bird that we are observing will fly away if we attempt to come closer, we appeal to our experience of birds. This means that we search our memories for similar situations, and check if in these situations the bird did fly away or not. If it did fly away in all the similar situations we can recall, then we infer that it will do the same in the present case.

It should be noted that these memories need not be individual memories for each and any situation we faced in our life. On the contrary, in most cases the memorization process sums up similar experiences into a unique memory, leaving aside insignificant details. For example, we know about sparrows without

remembering each and any situation in which we faced one. So the memories we have of things, and therefore those on which we rely to draw automatic inferences, are in fact mental representations of some archetypal cases, that is, they are some kind of concepts.

But they are not any kind of concepts. As is well known, concepts can be more or less general, with more general concepts subsuming more specific ones, thus forming clustering upward chains of more and more general concepts. A key point here is that to draw automatic inferences, we rely only on the most specific of our concepts. The reason is that these are the most detailed and precise representations we have of things, and thus the most suited to base our inferences upon. We search these most precise representations, looking for some that match what we perceive of the present situation, that is, that satisfy all the features we are able to grasp of it, and we check whether the ones we have found also satisfy the feature 'flies away'. If all of them do, then we infer that the bird will fly away.

But in doing so we do not check each and any of our most specific concepts that match what we perceive of the present situation, which would be far too long and mentally demanding. Rather, we simply check the first ones that come to our mind, that is, the ones that are the most vivid in our memory. If all of these satisfy the feature 'flies away', then we conclude that the bird will fly away. It is this partial recollection of memories that causes automatic inferences to be non-monotonic, as a more complete information about the current situation would probably reactivate a different set of memories and so bring us to different conclusions. However this incomplete retrieval does not significatively affect the relevance of our conclusions, because our most vivid memories are also likely to be the most significant and important for us, and thus the best able to provide us with useful conclusions.

It may also happen that we fail to find a memory of similar object(s)/situation(s), and that therefore the present situation appears new to us. In such cases the above process cannot be completed, and this in turn triggers learning

processes. These essentially consist in the creation of a new memory, or the supplementation of existent ones.

One may object that such a process cannot be hypothesized in animals, since it critically relies on mental representations and (one may argue) animals do not have mental representations. It is true that the idea of animals having mental representations has long been dismissed, as animals were mostly regarded as mere stimulus-response machines, notably by philosophers. Yet over the past decades this view has been seriously challenged by animal cognition studies, and nowadays it is widely acknowledged by researchers in the field that most animals, including small brained ones such as arthropods, do have mental representations [2]. The reluctance that the non-specialist may feel to regard animals, and particularly tiny brained ones, as having mental representations, presumably comes from the fact that he/she tends to associate to the meaning of these words the intuition he/she has of his/her own human subjective experience, which he/she needs not do. From the cognitive standpoint — and as for instance Jackson & Cross [3] put it — a mental representation is much more simply *'something more like an internal state that carries information and is then put to use during decision making. A key idea is that representations are used for processing that happens several steps removed from simple stimulus-response chains'*. It is in such a dispassionate sense that we take the term.

Brain sciences

Brain sciences for their part provide details on how the above described inferential process could plausibly be realized in brains. At first glance, brains show noticeable disparities across animal kingdom, notably in size or anatomical plan. But beyond these obvious differences they all share the same general functional schema.

2 See for instance [Srinivasan, 2006], [Chittka and Niven, 2009] and [Jackson and Cross, 2011] on mental representations and concepts in arthropods.

3 [Jackson and Cross, 2011] p. 120.

In particular, in all brained species perception is achieved through a series of parallel sensory channels, one for each sensory modality (vision, olfaction, etc.) occurring in the considered species. Each of these channels collects cues from the outside world by the means of a set of specific neural receptors (light receptors in the retina for the visual channel, chemico-receptors in the olfactory channel, and so on) and further processes them to extract relevant features. For instance, edges, colours or directional motion are features typically extracted in visual channels, while particular odours are features extracted in olfactory channels. In all cases, the extraction of a given feature is achieved by a set of dedicated neurons that specifically respond to this feature [4]. As one progresses along a sensory pathway, features may combine to form higher-level features, such as a particular angular arrangement between edges or a particular texture [5]. Which features are extracted by a given channel depends on its particular neural organization, so the type, number and complexity of the features that are eventually perceived varies across species, but the general rule for all brained species from arthropods to humans is that sensory information is analysed into features.

And in all brained species too, sensory channels finally output to some central brain areas. These are diversely organized — obviously an insect's brain does not look like that of a mammal — but they all have two main characteristics in common. First, they contain neurons that receive inputs

4 The seminal work on the subject is that of Hubel and Wiesel, who around the sixties identified edge detectors in cats' and monkeys' primary visual cortices ([Hubel and Wiesel, 1959, 1962, 1968]). Ever since, numerous studies have been carried out in a great number of species, demonstrating the existence of feature detectors in both vertebrates and invertebrates, and this for all sensory modalities. About visual features in invertebrates, one may consult for example [O'Carroll, 1993] (edge detectors in dragonflies), [Barnett et al., 2007] (motion detectors in flies) and [Paulk et al., 2008, 2009] (colour and motion detectors in bumblebees). For a brief comparison between vertebrate and invertebrate visual and olfactory channels, see [Chittka and Niven, 2009] p. 996-997; for a detailed survey of olfactory channels in vertebrates and invertebrates, see [Wilson and Mainen, 2006]; for a comprehensive analysis of the homologies between the first layers of insect and vertebrate visual systems, see [Sanes and Zipursky, 2010].

5 For a short overview of complex features extraction in different sensory pathways, see [Taylor et al., 2006] p. 8239. For examples of complex visual features in macaques, see [Tsunoda et al., 2001] and [Tanaka, 2003].

from distinct sensory channels, so that they respond to the co-occurrence of features from different sensory modalities; and second, they are involved in learning and memories [6].

In insects for example, the lateral and medial protocerebrum and the mushroom bodies, three highly interconnected neuropils from the central brain, receive direct or indirect inputs from all sensory channels [7]. Neurons from the mushroom bodies have been shown to respond to stimuli from multiple sensory modalities, and even more interestingly, to respond to particular combinations of stimuli from multiple sensory modalities [8]. On the other hand, the mushroom bodies have long been thought to be involved in insect learning and memory [9], although at this time it is unclear whether these occur in the mushroom bodies themselves or whether the mushroom bodies are but a relay of information towards higher brain centers located in the lateral and medial protocerebrum [10].

Similarly in mammals, sensory information converges to the medial temporal lobe (MTL), a brain region that has long been known to be crucially involved in learning and memories. Yet here we need to go more into details,

6 The reader unfamiliar with animal cognition studies may be reluctant to imagine small-brained animals as being able to learn, as the contrary has long been a popular assumption. Yet during the last decades, a large amount of evidence showing that these creatures are capable of learning has been gathered. In fact, even the minuscule nematode worm *C. elegans*, the nervous system of which is composed of exactly 302 neurons, has been shown to be able to learn to avoid stimuli associated with noxious environments ([Zhang et al., 2005]). The idea that is gaining ground today is rather that *'learning ability may be an emergent property of nervous systems and, thus, all animals with nervous systems should be able to learn'* ([Hollis and Guillette, 2011] p. 24). This, however, is not to say that any learning should necessarily rely on memories, and one could valuably argue that *C. elegans*' nervous system is too scanty to be called a brain, and also to leave room for mental representations. Rather, it is likely that selective pressure for learning has given rise to memories and concepts, because these are a very efficient support for learning.

7 A general schema of the neural interconnections within the insect brain can be found for example in [Strausfeld et al., 1998] p. 30.

8 See for examples [Schildberger, 1984] in cricket, and [Strausfeld et al., 1998] pp. 27–31 in cockroach. For a general account on multimodal sensory integration in insects, see [Wessnitzer and Webb, 2006].

9 For a brief historical overview of the concept of mushroom bodies as learning and memory centers, see [Strausfeld et al., 1998] p. 14.

10 See [Strausfeld et al., 1998] pp. 31–32.

as cognitive psychology usually distinguishes between several kinds of memories, not all of which are believed to depend on the MTL, and not all of which appear to play a role in automatic inferences. Indeed, although the exact extent and definition of each of them is still under debate, researchers in the field generally agree to distinguish at least four kinds of memories.

A first one is called *working memory*. It is defined as a short-time memory that temporarily stores the information required to carry out complex cognitive tasks. It is the memory that for example allows us to keep in mind the different parts of a problem while solving it, or to remember having already completed a given sub-task as part of a more complex one. A second kind of memory is known as *procedural memory*. It is thought to be responsible for the learning of motor skills, such as, for instance, how to ride a bike. Contrary to working memory, procedural memory is a long-term memory. It also has the salient feature that it generally operates outside of the subject's awareness. A third kind of memory is called *episodic memory*. It is defined as the memory of personally experienced events, and as such it is believed to store the salient episodes of one's life. It is a long time memory, just as procedural memory. Finally, the last kind of usually distinguished memory is called *semantic memory*. It is defined as a long term memory that stores one's knowledge of general facts — that is, one's knowledge about things. Semantic and episodic memories are strongly related, since most of one's knowledge about things is abstracted from one's personal experience. Jointly, they are referred to as *declarative memory*. Contrary to procedural memory, declarative memory requires awareness, in the sense that one can only recall facts and events one was first aware of[11].

11 It should be stressed that this does not mean that it requires self-awareness (*i.e.* self-consciousness). Indeed awareness of something should not be mistaken with self-awareness. The cat stalking a bird is aware of a bird being there, but this does not entail that it is aware of its own awareness of a bird being there. Self-awareness is but a special case of awareness, that can only occur in animals that possess appropriate pathways able to collect information about their own neural states and to integrate it with that from other pathways for further processing. Most animals are not self-aware, although they are quite aware of the events and objects they react to.

Among these various kinds of memory, semantic memory naturally appears as the very plausible support for automatic inferences, and so it is the one in which we shall be principally interested, as well as in its relations with episodic memory. Contrary to procedural and working memories, episodic and semantic memories critically depend on the MTL. This is evidenced by the fact that brain lesions to the MTL give rise to severe impairments in both episodic and semantic memories, but not in procedural and working memories. More precisely, lesions to the MTL entail a profound *anterograde* amnesia (inability to form new episodic and semantic memories) along with a *retrograde* amnesia (loss of episodic and semantic memories that were acquired before the occurrence of the lesions). This retrograde amnesia is typically temporally graded, which means that more recent memories are lost while more remote ones are spared [12]. In humans for example, the retrograde amnesia that follows MTL lesions can extend across a period varying from a few years to several decades before the lesion, depending on the extent of the damage [13]. This temporal gradation has led to the prominent view that the MTL is critical for the formation of episodic and semantic memories, but that these are then progressively 'backed up' in other brain areas, so that over the years they become independent from the MTL [14].

To get an insight into how the MTL may form episodic and semantic memories, a rapid look at its anatomical organization can be useful. The MTL is a composed and hierarchically organized structure, that comprises the hippocampal formation and the entorhinal, perirhinal, and parahippocampal cortices. At the lower end of the hierarchy, the perirhinal and parahippocam-

12 For a general review of the role of the MTL in declarative memory, see [Preston and Wagner, 2007]; for a detailed survey of its role in the acquisition of semantic memories, see [Bayley and Squire, 2005] and [Levy et al., 2004]; for a study of anterograde and retrograde amnesia for semantic knowledge, see [Manns et al., 2003] and [Bayley et al., 2006].

13 [Squire and Bayley, 2007] pp. 185–186.

14 See p. 24 bellow.

pal cortices receive direct or indirect inputs from different sensory pathways, and in particular from the visual, somatosensory, olfactory, and auditory pathways [15]. Perirhinal and parahippocampal cortices then provide the major inputs to the entorhinal cortex, which in turn provides the major inputs to the hippocampus, at the higher level of the hierarchy [16]. All these connections are reciprocal, suggesting that information may flow in both directions. In addition, all components of the MTL have substantial reciprocal connections with the amygdala, an adjacent area known to be involved in the processing of emotions [17].

At the top end of the hierarchy, neurons from the entorhinal cortex and hippocampus have been shown to respond specifically to the presentation of various stimuli corresponding to a same concept, regardless of the stimuli's other qualities. Crucially, such results were obtained not only in humans, but also in non-human mammals. For instance, neurons responding specifically to particular nests-situations have been found in mice hippocampus. More precisely, three different types of nest-responding neurons were identified. A first type of neurons responded specifically to 'inside-nest' situations, firing robustly and persistently whenever the mice were in a nest, while remaining nearly silent otherwise. A second type of neurons did quite the contrary, and responded to 'outside-nest' situations (that is, they fired whenever the mice were not in a nest, and ceased to do so as soon as they entered one). A last kind of neurons responded to nest encounters, firing transiently but robustly each time the mice were reaching a nest, and only then. All three kinds of neurons appeared unresponsive to any of the many other objects that were presented to the mice, and their responses were insensitive to the nests' particular qualities such as geometrical shape, location, odour, or the mater-

15 See for example [Suzuki and Eichenbaum, 2000] p. 176.

16 For a synthetic schema of neural interconnections between the structures that compose the MTL, see [Quiroga, 2012] p. 592; for a more comprehensive description, see [Suzuki and Eichenbaum, 2000] pp. 176–178 and [Preston and Wagner, 2007] pp. 306–308.

17 See for example [Phelps and LeDoux, 2005].

ial they were made of. In fact, the only discernible criteria for the neurons' firing was whether or not the considered object could be used as a nest by the mice [18].

Similarly in humans, neurons responding specifically to particular persons or to particular landmarks have been found in the hippocampus and the entorhinal cortex. Again, the identified neurons fired to any stimulus corresponding to their target concept, regardless of its particular qualities and regardless even of its modality, and they did not fire to any of the many other presented stimulus. In one study for instance, a neuron was recorded that fired indifferently to various photos of a famous actress, and also to her name, either written or pronounced. Other recorded neurons however were less selective, and responded to several stimuli from the stimulus set. But in these cases the stimuli they responded to were obviously related, and could be seen as instances of a more general concept. For example, one neuron fired to both the Eiffel Tower and the Tower of Pisa, while another fired to different actors featuring in a television series, and another to pictures of animals [19].

It should be specified here that the fact that only one single neuron responding to a given concept was identified in each case does not mean in any way that this precise neuron was the only one in the subject's brain to do so. Being living cells, neurons are intrinsically subject to noise, damage and death, and the robustness of brain processes largely relies on redundancy [20]. It is therefore very likely that for each concept represented in the brain, there is a whole assembly of neurons that specifically respond to it. The reason why only one was found during the experiments is because only a few neurons were recorded in each session, due to technical and/or medical constraints.

18 [Lin et al., 2007].

19 [Quiroga, 2012].

20 There are however exceptions to this general rule. A well documented one can be found in insects, were a single neuron has been showed to be the unique reward signalling neuron in relation with gustatory stimuli (see for example [Menzel and Giurfa, 2001] p. 63). Overall, redundancy appears to be much lower in arthropods, which is presumably linked to the fact that the risk of death of neurons is less critical in these short living animals than in long living ones.

At the same time, another group of studies evidenced that entorhinal and hippocampal neurons do not fire only when the subject perceives their target stimuli, but also when he imagines or recalls these same stimuli [21]. In sum, they fire whenever the subject is 'thinking about' their target stimuli.

Taken together, the hierarchical structure of the MTL and the selective firing of entorhinal and hippocampal neurons strongly suggest that MTL neurons integrate co-occurrent features from all modalities into multimodal representations of the encountered objects and situations. A common view is that each MTL neuron at its own hierarchical level receives convergent inputs from a definite set of neurons from the next lower level, and thus encodes the conjunction of the (conjunctions of the) features encoded by its input neurons [22]. For example, perirhinal and parahippocampal cortices are generally thought to encode conjunctions of features from different sensory modalities. To account for their processing in parallel, it has been proposed that the perirhinal cortex might be more particularly involved the representation of objects [23], and the parahippocampal cortex in the representation of contexts [24]. These separate representations would then be assembled together within the entorhinal cortex and the hippocampus, along with emotional information coming from the amygdala. In this manner, entorhinal and hippocampal neurons would encode multimodal, contextualized and emotionally valued mental representations of objects and situations. It is however difficult to better separate the respective roles of entorhinal and hippocampal neurons at this time, as these are still under intense debate.

Another salient property of MTL neurons is their ability to change their connection pattern almost in real time. This versatility is believed to rely on *long term potentiation* (LTP), a neural mechanism through which excitative

21 [Gelbard-Sagiv et al., 2008].

22 See for example [Shimamura, 2010] pp. 11206-1209.

23 [Taylor et al., 2006]; see also [Holdstock et al., 2009].

24 [Aminoff et al., 2013].

connections between co-active neurons can be rapidly and durably strengthened [25]. In practice, a neuron that arborises in the vicinity of some higher level neuron but has very few connections to it, might, if it happens to be activated at the same time as it, have its connections to this neuron reinforced and become in this manner one of its input neurons. In this manner, the (conjunction of) feature(s) encoded by the lower level neuron could be quickly integrated into the conjunction of features encoded by the higher level one [26].

Furthermore, the reciprocal connections within the MTL suggest that the firing of MTL neurons might (re)activate feature-responsive neurons in higher modality-specific areas, thus endowing MTL neurons with subjective semantical content [27]. For example, the simple thinking about the concept of red is enough to elicit in ourselves a reminiscence of the sensation of redness. This might be due to the fact that the activation of the MTL neurons that respond to the concept of red triggers the activation of those that encode the feature 'red' in higher visual areas.

The above considerations have led to the prevailing view that one of the main functions of the MTL neurons might be to form mental representations (*i.e.*, concepts) of experienced objects and situations by rapidly binding together simultaneously perceived features [28]. However, the fact that some of these neurons were found to fire specifically to individual persons or objects and others to various related stimuli indicates that they might be able to encode specific concepts as well as more general ones. How the encoding of more general representations articulates with that of more specific ones and especially with that of particular memories is still an open question, but obviously the semantical content of a more general representation — that is, the set of

25 For a general account on long term potentiation, see for example [Izquierdo, 1993].

26 Such a proposition can be found for example [Shimamura, 2010] p. 1206.

27 [Martin and Chao, 2001], [Eichenbaum, 2000] pp. 47-48, [Squire et al., 2004] pp. 282–283.

28 [Preston and Wagner, 2007] pp. 312–313.

features that compose it — should depend on that of more specific ones. A possibility would be that more general representations are drawn from more specific ones as the set of their common salient features. Indeed, it seems that (for example) our general concept of bird is abstracted from the more specific concepts we may have of birds, such as for instance concepts of particular bird species, mental representations of particular individual birds or of particular bird-situations. These more specific concepts in turn might be drawn from even more specific ones, and so on to individual memories of particular bird-situations [29]. As time passes, the individual memories at the source of a given concept would generally get lost, but the concept itself would remain as the trace of their repeatedly encoded common salient features.

At the neural level, this process might result from the fact that neurons supporting similar representations (*i.e.*, that share a lot of features in common) respond to similar inputs, hence due to neural noise it may happen that a neuron wired so as to respond to one input accidentally responds to another, similar enough one. In such a case, the combined action of long term potentiation and long term depression (LTD, a neural mechanism complementary to LTP that rapidly and durably decreases the efficacy of synapses [30]) might change the neuron's connections in such a way that it will now respond to the conjunction of the features that are common to the inputs. In this manner, a fraction of the neurons initially responding specifically to the one or the other of a given class of similar inputs might progressively come to respond to the general concept that subsumes them all. In some cases, the original representations might end up by losing all neuronal support, accounting for the loss of the individual memories at the origin of the concept.

29 Here we disregard the particular and strictly human case of concepts that may be acquired through verbal definitions.

30 For a general and accessible introduction to long term depression, one may consult [Bear and Abraham, 1996]. Here we are more particularly interested in the so-called *heterosynaptic* long term depression ([Bear and Abraham, 1996], pp. 437-442), which selectively decreases the efficacy of the inactive input synapses of a currently active neuron.

Whether or not this suggested process is correct in its details, it remains highly probable that MTL neurons are able to extract relevant — which means, in particular, invariant — features from the continuous and fickle flow of immediate memory, and to bind them into more general and more lasting representations, that is, into concepts. The degree of generality of these can be very variable, with some of them being very general and others much more specific. In some cases, the individual memories at the origin of a concept might be preserved for some time, but generally they will eventually get lost through the abstraction process.

As previously mentioned, firing MTL neurons probably send back excitative inputs to higher sensory areas. According to the prevailing view, the neurons from different modality-specific areas that are regularly co-(re)activated by a given assembly of MTL neurons might progressively grow direct interconnections, so that over the time they would form an autonomous network capable of supporting the corresponding representation on its own [31]. As a result, mental representations encoded by entorhinal and hippocampal neurons would be progressively 'backed-up' in a distributed form over higher sensory areas, accounting for the temporally graded character of the retrograde amnesia that follows MLT lesions [32].

Yet this does not mean that the entorhinal and hippocampal neurons that encode a given representation should cease to do so after it is backed up in modality-specific areas. Rather, it is likely that they keep on supporting a given representation as long as it is relevant for the subject [33]. They thus appear to be the first and principal support of concepts in mammals' brains. For

31 [Squire and Alvarez, 1995] pp. 171–174.

32 See for example [Shimamura, 2002], [Squire et al., 2004] p. 296 and [Preston and Wagner, 2007] pp. 312–315.

33 Such a claim can be found for example in [Quiroga, 2012] pp. 594–595, and [Shimamura, 2002].

this reason we shall call them *concept neurons* [34].

Due to their character of primary support of concepts, it is likely that concept neurons subserve a number of mental functions involving concepts and memory. In particular, it has been proposed that transitions between concepts in a thinking mind might reflect transitions in the activation of the assemblies of concept neurons that support them [35]. Another appealing suggestion (which is not incompatible with the previous one) has been that the memories of episodes constituting episodic memory might be encoded in the hippocampus as ordered sequences of discrete mental representations encoded by assemblies of concept neurons [36].

Taking up the first of these proposals, we suggest that concept neurons might in particular support a very special kind of transition between concepts, namely the drawing of automatic inferences. This is quite natural an hypothesis if one considers that from the evolutionary standpoint the adaptive character of memory essentially lies in the fact that it allows to store information from past experience, and to use in it decision making so to adapt decisions to the agent's environment [37].

On the neural side, such a transition process might rely on a very simple mechanism. Since a mental representation is evoked as soon as the neurons

34 We borrow the term from [Quiroga, 2012], although contrary to him we do not take it to encompass all the MTL neurons that exhibit selective firing, but only those from the entorhinal and hippocampal cortices. Indeed neurons from lower MTL areas rather appear to convey partial representations to entorhinal and hippocampal neurons, the latter being (as it seems) responsible of their further integration into fully fledged concepts.

35 See for example [Quiroga, 2012].

36 [Eichenbaum, 2004].

37 Similar ideas can be found in a number of relatively recent papers ([Buckner, 2010] provides an introductory review of the subject). Yet these papers generally consider forms of inference that are more complex than the one we are interested in here, and in particular, that rely in a decisive manner on episodic memory. But it is generally acknowledged that most non-human mammals have, at best, very limited episodic memory capacities, although they undoubtedly do make inferences, in the sense defined page 13 above. It follows that the forms of inference considered in these papers cannot be the whole of the inferences based on memory functions, and also that automatic inferences do not depend on episodic memory.

that support it fire, its vividness in the mind (that is, the ease with which it can be recalled) probably depends on the ability of the considered neurons to respond to an appropriate input. Differences in numerical size of neural assemblies, coupled with mutual excitative connections between neighbouring neurons supporting a same concept, might yield differences in response-time of neural assemblies supporting different concepts. In addition, mutual inhibitory connections between neighbouring neurons supporting different concepts might hamper the activation of concept-supporting neural assemblies with longer response-times. This would bring neural assemblies whose neurons receive a common input to compete for activation, with the result that only those with the shortest response-times would finally get activated, which means that only the mental representations supported by these assemblies would finally be recalled.

This of course is a mere hypothesis, that would need to be checked against observational data. In particular, it would be necessary to verify whether the connections patterns within the MTL and the existing synaptic mechanisms are compatible with such a process. But the current knowledge on the MTL is too fragmented and uncertain to allow such a straightforward assessment, and all we can do for now is to make plausibility arguments. In this respect, it should be emphasized that mutual activation and inhibition between neurons and competition between neural assemblies are very common mechanisms in brains, which makes the suggested process all the more plausible.

It would however remain to explain how it is that only the neural assemblies supporting most specific representations are involved in this process [38], and also how the conclusions of the inferences are effectively retrieved from the finally recalled memories. Various hypotheses relative to both of these issues can be made, but in the absence of more detailed anatomical and functional data they would be purely speculative and so of little interest, therefore we

38 Recall that, according to our analysis, automatic inferences rely only on the subject's most precise mental representations (*c.f.* p. 15 above).

shall leave these questions unsettled.

Leaving the special case of mammals and turning back to brained species in general, we may recall that in all these species sensory information is analysed into features that then converge to central brain areas where they are integrated, and that in all these species too, these same central brain areas are involved in learning. These commonalities make it very likely that not only in mammals, but more broadly in all brained species, general knowledge supporting decision making is stored in mental representations encoded as sets of co-occurrent features. But this is perhaps not surprising, if one considers that features and sets of features are probably the most economical way to store representations of objects and situations in a brain [39].

Furthermore, the encoding of sets of co-occurrent features by dedicated neurons that sum up co-occurrent inputs, as do MTL concepts neurons in mammals, is probably also the most simple and economical one possible. Therefore, it would not be surprising either that a similar organization might be conserved across species, and that neurons analogous to MTL concept neurons might exist in all brained animals, although they have not yet been identified in non-mammalian species.

Finally, the neural mechanisms involved in the above suggested process are also very simple and common ones, and can certainly be found in the simplest brains. Accordingly, there is no reason to suppose that such a process should only occur in mammals. On the contrary, its very basic character suggests that it might take place in a broadly similar manner in all brained species, accounting for the pervasiveness of automatic inferences across animal kingdom.

39 An enlightening argumentation in this sense can be found in [Srinivasan, 2006]. One may also consult Chittka and Niven's comment on Srinivasan ([Chittka and Niven, 2009] p. 1000).

1.2 Informal presentation of the framework

In the previous section, we argued that automatic inferences form the primary core of reasoning in natural agents, and we described a simple process that plausibly supports them, first at the mental level of description and then at the neural level. In light of this, it appears that the most convenient approach for the modelling of reasoning in natural agents is to proceed by steps and to model first this core process, and leave the modelling of additional cognitive abilities to further work. Accordingly, the present dissertation shall be dedicated to the modeling of automatic inferences and their associated learning processes.

In practice, we shall consider an unspecified cognitive agent (hereafter 'the agent') which shall be assumed to draw automatic inferences according to the above described process, and we shall set up a logical framework that intends to match this process. Therefore, it should be kept in mind that the framework to be introduced is meant to account for automatic inferences alone, to the exclusion of any other cognitive ability the agent may enjoy. In particular, it shall not intend to model more sophisticated kinds of inferences, notably those that rely in a decisive manner on consciousness, reflexivity or verbal representation. But in cases where automatic inferences occur as a subprocess of these, it shall of course account for this precise subprocess.

We shall call *inferential system* the part of the agent's brain that, according to the above account, processes automatic inferences. It is this inferential system, or more exactly, its operations, that we shall intend to model.

To represent the information that is processed by the agent's inferential system, we shall use a propositional language the variables of which shall stand for the features the agent is physiologically able to perceive — that is, the features that its perceptual system is able to extract from sensory informa-

tion. Consequently, features shall not be regarded as properties of external world's objects, but as information (in the computational sense of the term) flowing within the agent's brain. However we shall keep in mind that, from the agent's point of view (*i.e.*, as it experiences them), features indeed appear as properties of external world objects. This is why we shall sometimes write such things as *'in the agent's view, the object satisfies the feature f'*. It should be noted that the finite number of neurons in the agent's brain only allows it to discriminate, and thus to perceive, a finite number of features, hence the logical language shall be finite. For simplicity, we shall assume that the set of the features that the agent is physiologically able to perceive does not change over time, so the logical language shall be supposed to be fixed. As is the custom, propositional variables shall be denoted by lower case italic latin letters p, q, r, *etc.*, while formulas shall be denoted by lower case italic greek letters α, β, γ, *etc.*.

One may notice that the neural organization of sensory channels as sketched above only allows agents to perceive 'positive' features. For example, we may perceive something as being red, but not as not being black. Yet 'red' and 'black' are mutually exclusive features, in the sense where recognizing something as red automatically brings us to reject the idea that it is black. In other words, it brings us to conceive the considered object as not-black. At the neural plane, this is most likely achieved through mutual inhibition between neurons, but from a computational standpoint, it comes down to having negative information such as 'not-black' being processed by the agent's inferential system. For this reason we shall also consider negative features, and quite naturally represent them by negative literals of the language. Features themselves shall be denoted by slanted lower case latin letters f, *etc.* We shall sometimes write '*not-f*' to denote the negative counterpart of the feature *f*. Note that it is impossible for an agent to conceive an object/situation as satisfying at the same time a feature f and its negation *not-f*.

It should also be remarked that this same neural organization of sensory channels entails that natural agents can never perceive anything but features and conjunctions of features. In particular, they can never perceive disjunctions of features. This is true even when the available information is ambiguous. For instance, an agent will never perceive anything as being 'dark blue or black': it will perceive it as being dark blue, or as being black, or simply as being dark. In some extreme cases, its perception may vary over time between these possibilities — this has been dubbed *'multistable perception'* in cognitive psychology [40]. If the agent is capable of reflexivity, then it may, by a reflexive move on itself, perceive its own hesitation, and from it infer that the object it considers is 'dark blue or black'. But this is a posterior reconstruction, and in any cases, it will never perceive the object as 'dark blue or black'.

The mental representation of the elements of a given class of objects or situations [41] in the agent's mind shall be construed as a the set of positive and negative features that, in its view, are satisfied by the corresponding objects/situations. Here we shall make the simplifying assumption that all the

40 As Sterzer et al. characterize it, *'Multistable perception occurs when sensory information is ambiguous and consistent with two or more mutually exclusive interpretations. When no additional cues are available that allow perceptual synthesis to converge on one unique interpretation, perception alternates spontaneously every few seconds between two ('bistable') or more ('multistable') interpretations of the same sensory input'* ([Sterzer et al., 2009], p. 310). Multistable perception has been recognized for a long time (see [Schwartz et al., 2012] for a brief historical review), and occurs within all sensory modalities. Although its precise neural bases are still under intensive debates (which go far beyond the scope of the present dissertation), there is a general agreement that *'only a single perceptual solution can exist at once'*([Leopold and Logothesis, 1999] p. 260).

41 As we saw, memorization process generally sums up multiple experiences into a unique memory, so that a mental representation is in fact the common representation of the elements of a class of objects or situations; in cases where a mental representation is built up from a single experience, this class will simply contain one single element. Furthermore, since mental representations are nothing but sets of features, they do not distinguish between objects and the situations in which these appear. For example, a representation of birds may contain some information about the context in which the birds in question were observed. Hence a mental representation can be seen at the same time as the representation of some object(s), and as the representation of some situations involving these/this object(s). The ability to separate objects and situations is most likely realized at a much more higher cognitive level — provided, of course, that such a level is implemented in the agent's brain.

features in this set equally contribute to the mental representation, in other words that this set is *not* structured. Under this assumption, a mental representation can be simply seen as the logical conjunction of all the features that compose it. We shall moreover assume that the agent's mental representations are consistent, which means that a feature and its negation cannot take part at the same time in a same representation. This will allow us to figure mental representations by partial worlds. More specifically, if r is some mental representation in the agent's mind, we shall represent r by the partial world w such that for any literal λ in the language, w satisfies λ if and only if λ stands for a feature that pertains to r. Mental representations shall be denoted by slanted lower case sans-serif latin letters a, b, c, r, etc. If r is a mental representation in the agent's mind, we shall sometimes denote w_{r} the partial world that stands for r in the model.

We shall say that a mental representation r *satisfies* a feature f if and only if f is in the set of features that compose r, that is, if and only if in the agent's opinion, the corresponding object(s)/situation(s) satisfy/ies f. For example, the agent's mental representation of sparrows (supposing it has one) will satisfy the feature 'has a beak' if and only if in the agent's opinion, sparrows have a beak [42]. The fact that a partial world standing for a mental representation satisfies a literal λ shall thus represent the fact that, in the agent's opinion, the corresponding object(s)/situation(s) satisfy/ies the feature λ stands for. Similarly, we shall say that a mental representation r satisfies the content of information represented by the formula α if and only if the partial world that represents r satisfies α.

42 For the ease of reading, in examples we shall regard such things as 'has a beak', 'flies away' and so on, as features. This is quite a simplification, as actual features are in fact more likely to be of the kind of those identified in [Tsunoda et al., 2001] and [Tanaka, 2003], while 'has a beak', 'flies away' and so on, are more likely to be conjunctions of a great number of such features. However this simplification is harmless, since we regard mental representations as conjunctions of features and conjunction is associative.

To know something is nothing but to have a certain mental representation of that thing. Therefore, for any class of objects/situations that the agent knows, there is a corresponding mental representation in its mind. As explained above, from a computational standpoint only the most specific (precise) of them play an effective role in the drawing of automatic inferences, so we can restrict our interest to these. The set of all these most precise mental representations can straightforwardly be represented by the set $\mathcal{U}$ of partial worlds w such that w represents one of the agent's most precise mental representations in the above sense. Furthermore, since mental representations are supported in brains by assemblies of dedicated neurons, $\mathcal{U}$ can also be seen as representing a set of such assemblies: namely, the set of the assemblies of concept neurons in the agent's brain that support its most specific mental representations.

Our usual speaking of memories as 'more' or 'less' vivid one relative to another suggests that the difference in vividness of mental representations is somewhat quantitative. Ideally, we should check this intuition by looking at the neural mechanisms that govern the interactions between concept neurons. But for now these are poorly understood, and all we can do is to rely on introspection. Therefore, we shall assume that the vividness of mental representations is indeed quantifiable, which means that to any mental representation in the agent's mind corresponds a real number that is the measure of its vividness. On this basis, we shall account for the difference in vividness of mental representations by endowing $\mathcal{U}$ with a binary relation $<$. More specifically, if w and w' are partial worlds in $\mathcal{U}$, the fact that w $<$ w' will represent the fact that the mental representation figured by w is *more* vivid in the agent's mind than the one figured by w'.

We shall call the structure $\mathcal{M} = (\mathcal{U}, <)$ a *partial worlds model*. If one looks at $\mathcal{U}$ as the set of the agent's most precise mental representations, then $\mathcal{M}$ can be seen as representing the agent's worldview — its *Umwelt*, as Jackson and

Cross call it [43]. But if one looks at $\mathcal{U}$ as the set of assemblies of concept neurons that support the agent's most specific mental representations, then $\mathcal{M}$ can be seen as representing the agent's inferential system.

We shall say that the agent is *disposed to infer* β *from* α if and only if it is in such a condition that the following holds: if it were to consider the content of information α (where α represents the totality of the information that the agent considers at this time), then it would come, after a few instants, to consider, among others, the content of information β [44]. According to the inferential process suggested in section 1.1.2 above, the agent is disposed to infer β from α if and only if all its most vivid mental representations among those that satisfy α also satisfy β. At the neural level, a disposition to infer can be seen as a particular arrangement of the neural connections in the agent's inferential system such that, if a suitable excitative input was delivered within its inferential system specifically to its neurons that support concepts that satisfy α, then after a few instants only neurons supporting concepts that also satisfy β would be active [45]. In cases where the input originates from the agent's perceptual system, α will always be a literal or a conjunction of literals, since natural agents can never perceive anything but features and conjunctions of features. But in cases where automatic inferences occur as a subprocess in a more complex process the input may come from other brain areas, and then α may in principle be any formula. For this reason we shall work on the full language, and allow any formula to be a premiss or a conclusion of automatic inferences.

The agent's disposition to infer β from α is naturally represented in the framework by the fact that all the $<$-minimal elements in $\mathcal{U}$ among those that satisfy α also satisfy β. The reader acquainted with preferential logics will

43 [Jackson and Cross, 2011], pp. 154–155.

44 This definition is freely inspired from Leitgeb's book [Leitgeb, 2004].

45 See p. 27 above.

have recognized here a familiar pattern: $\mathcal{M}$ induces an inference relation on the language exactly in the same way as Kraus, Lehmann and Magidor's cumulative models do [46]. In our framework, the inference relation induced by $\mathcal{M} = (\mathcal{U}, <)$ shall stand for the set of all the agent's dispositions to infer, that is, for its general knowledge about things. On the logical front, our main task shall be to characterize the inference relations induced by partial worlds models, that is, to identify the logical rules that, according to our account, structure the agent's general knowledge.

It may be useful to clarify the status of the logical language in the framework. As mentioned above, the logical language is the language in which we describe the information that flows within the agent's inferential system. This is not the same as describing beliefs, since beliefs result from the processing of information in the agent's brain, and so describing beliefs requires to take an external viewpoint on the agent — this is what standard modal logic does. By contrast, what we intend to do is to model the computational process itself, which does not operate on beliefs but on information alone.

Nor should the logical language be taken as the agent's language, even in the sense of some fodorian 'language of thought'. The fact is that the posited computation does not occur at some linguistic level and does not involve any language, even a computational one. Even though assemblies of concept neurons can be seen as symbolic representatives of the concepts they support, there are no symbolic representatives of the logical operators: logical operations are directly performed on the assemblies of neurons, as simultaneous activation or lack of activation. It follows that the computation doesn't involve any language in the agent, for which the logical language could be taken to stand. Rather, the logical language is the language that we use to describe the processing of information in the agent's brain, quite in the same way that a physicist may use some suitable mathematical language to describe some physical phenomenon.

46 [Kraus et al., 1990], hereafter KLM.

Therefore, the logical rules that shall be shown to characterize the induced inference relations shall not be interpreted as reasoning rules that the agent applies (which would require some inner language), but as rules it obeys, just as physical phenomena obey physical rules without knowing anything of them.

Chapter 2

The logical setup

We now turn to the formal introduction of the above drafted logical framework. For generality's sake and easier comparison with other author's works, throughout this chapter and the following chapter 3, we shall *not* ask the logical language to be necessarily finite. We shall focus down on finite languages from chapter 4 on.

The plan of this chapter is as follows: in section 2.1, we define a notion of partial worlds, along with some useful notions of monotonic consequence and equivalence; in section 2.2 we provide a brief reminder of KLM's framework, and in section 2.3 we take inspiration from their work to define partial worlds models. Then we show that inference relations induced by *smooth* partial worlds models form a strict subclass of KLM's cumulative relations. We identify some supplementary valid rules, but these are not sufficient to reach completeness.

2.1 Partial worlds, $\mathcal{U}$-consequence and $\mathcal{U}$-equivalence

Let $\mathcal{L}$ be a (not necessarily finite) propositional language, and $Var(\mathcal{L})$ the set of propositional variables of $\mathcal{L}$. A *partial truth assignment* on $Var(\mathcal{L})$ is a partial function $\boldsymbol{w} : Var(\mathcal{L}) \longrightarrow \{0, 1\}$. A *complete truth assignment* on $Var(\mathcal{L})$ is a total function $\boldsymbol{w}' : Var(\mathcal{L}) \longrightarrow \{0, 1\}$. In line with usual definitions, we regard total functions as partial functions that happen to be defined everywhere, and

thus complete truth assignments as partial truth assignments that happen to be defined for all propositional variables.

Each partial truth assignment $\boldsymbol{w}$ generates a *partial world* w, and each total truth assignment $\boldsymbol{w'}$ generates a *complete world* w' (except for the empty truth assignment, for which we state that it doesn't generate any world). We denote $\mathcal{W}_{\mathcal{L}}$ the set of all complete worlds for $\mathcal{L}$, and $\mathcal{W}_{\mathcal{L}}^{\text{P}}$ the set of all partial worlds for $\mathcal{L}$ ($\mathcal{W}_{\mathcal{L}} \subseteq \mathcal{W}_{\mathcal{L}}^{\text{P}}$). For each w $\in \mathcal{W}_{\mathcal{L}}^{\text{P}}$, we define:

. Var^{+}(w) is the set of propositional variables p from $\mathcal{L}$ such that $\boldsymbol{w}(p) = 1$;

. Var^{-}(w) is the set of propositional variables p from $\mathcal{L}$ such that $\boldsymbol{w}(p) = 0$;

. Lit(w) $= Var^{+}$(w) $\cup \{\neg p_i / p_i \in Var^{-}(\text{w})\}$ is the set of literals from $\mathcal{L}$.

In the particular case where $\mathcal{L}$ is finite we use the pair (Var^{+}(w), Var^{-}(w)) to denote w. For example, if $Var(\mathcal{L}) = \{p,\ q,\ r\}$, the pair ($\{p\}$, $\{q\}$) denotes the partial world w such that $\boldsymbol{w}(p) = 1$, $\boldsymbol{w}(q) = 0$ and $\boldsymbol{w}(r)$ is undefined. Moreover (and still in the case where $\mathcal{L}$ is finite), for each w in $\mathcal{W}_{\mathcal{L}}^{\text{P}}$, we define the $\mathcal{L}$-formula δ(w):

$$\delta(\text{w}) = \bigwedge \lambda / \lambda \in Lit(\text{w})\ ^{1}$$

We call δ(w) the *description* of the partial world w.

Satisfaction of $\mathcal{L}$-formulas by a complete world w is denoted $\models$, and defined as usual:

. if α is a propositional variable, then w $\models \alpha$ iff $\boldsymbol{w}(\alpha) = 1$;

. if $\alpha = \neg\beta$, then w $\models \alpha$ iff w $\not\models \beta$;

. if $\alpha = \beta \wedge \gamma$, then w $\models \alpha$ iff w $\models \beta$ and w $\models \gamma$,

other connectives being defined as usual in terms of negation and conjunction. We denote $\vdash$ the *classical consequence* relation on $\mathcal{L}$, and $\equiv$ the *classical equivalence*. If T is a set of $\mathcal{L}$-formulas, we say that w satisfies T and write w $\models T$

1 This formulation is freely inspired from [Dubois, 2008] p. 9.

if and only if for any formula $\alpha \in T$, w $\models \alpha$.

Satisfaction of $\mathcal{L}$-formulas by a partial world w is denoted $\Vdash$, and defined with the help of supervaluations:

For any $\mathcal{L}$-formula α and any partial world w, w $\Vdash \alpha$ iff for every complete world w' such that w' $\models Lit$(w), w' $\models \alpha$.

In other words, a partial world w satisfies a $\mathcal{L}$-formula α iff any way one may 'complete' w to get a complete world w', the latter satisfies α. One can easily check that if w itself is a complete world, then w $\Vdash \alpha$ iff w $\models \alpha$. In the particular case where $\mathcal{L}$ is finite, the above definition comes down to:

w $\Vdash \alpha$ iff for every complete world w' such that w' $\models \delta$(w), w' $\models \alpha$, *i.e.* iff δ(w) $\vdash \alpha$.

Note that:

. w $\Vdash \alpha \wedge \beta$ iff (w $\Vdash \alpha$ and w $\Vdash \beta$);

. (w $\Vdash \alpha$ or w $\Vdash \beta$) entails w $\Vdash \alpha \vee \beta$, but

. w $\Vdash \alpha \vee \beta$ *does not* entail (w $\Vdash \alpha$ or w $\Vdash \beta$)

For example, let w be such that $\boldsymbol{w}(p)$ is undefined: then w $\Vdash p \vee \neg p$, but w $\nVdash p$ and w $\nVdash \neg p$.

Let $\mathcal{U} \subseteq \mathcal{W}_{\mathcal{L}}^{p}$ be a set of partial worlds. For any set T of $\mathcal{L}$-formulas, we denote $\mathcal{W}_u(T)$ the (possibly empty) set of partial worlds w $\in \mathcal{U}$ such that for any formula $\alpha \in T$, w $\Vdash \alpha$. When $T = \{\alpha\}$, we omit the brackets and simply write $\mathcal{W}_u(\alpha)$.

We define on $\mathcal{L}$ a *$\mathcal{U}$-consequence relation* $\|\!\sim_{\mathcal{U},\mathcal{L}}$:

$\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$ iff $\mathcal{W}_u(\alpha) \subseteq \mathcal{W}_u(\beta)$.

It is immediate that $\|\!\sim_{\mathcal{U},\mathcal{L}}$ is transitive. It is also supra-classical (*i.e.* $\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$ whenever $\alpha \vdash \beta$): indeed, suppose that $\alpha \vdash \beta$ and let w $\in \mathcal{W}_u(\alpha)$: every complete world w' that satisfies Lit(w) also satisfies α and thus also satisfies β, so w $\Vdash \beta$, *i.e.* w $\in \mathcal{W}_u(\beta)$. On the other hand, $\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$ does *not* imply $\alpha \vdash \beta$: for instance, if w $= (\{p, q\}, \varnothing)$ and $\mathcal{U} = \{$w$\}$, we have $p \|\!\sim_{\mathcal{U},\mathcal{L}} q$, though $p \nvdash q$. Moreover, $\|\!\sim_{\mathcal{U},\mathcal{L}}$ doesn't meet contraposition (*i.e.* $\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$ doesn't entail

$\neg\beta \mid\!\sim_{\mathcal{U},\mathcal{L}} \neg\alpha$): for instance, if w = ($\{p, q\}$, $\varnothing$), w' = ($\varnothing$, $\{q\}$) and $\mathcal{U} = \{$w, w'$\}$, we have $p \mid\!\sim_{\mathcal{U},\mathcal{L}} q$, but $\neg q \not\mid\!\sim_{\mathcal{U},\mathcal{L}} \neg p$, since w' $\models \neg q$ and w' $\not\models \neg p$. Finally, for any $\mathcal{L}$-formula α, we have $\alpha \mid\!\sim_{\mathcal{U},\mathcal{L}} \bot$ iff $\mathcal{W}_u(\alpha) = \varnothing$.

We also define on $\mathcal{L}$ a $\mathcal{U}$-*equivalence relation* $\cong_{\mathcal{U},\mathcal{L}}$:

$\alpha \cong_{\mathcal{U},\mathcal{L}} \beta$ iff $\alpha \mid\!\sim_{\mathcal{U},\mathcal{L}} \beta$ and $\beta \mid\!\sim_{\mathcal{U},\mathcal{L}} \alpha$, that is,
iff $\mathcal{W}_u(\alpha) = \mathcal{W}_u(\beta)$.

Supra-classicality of $\mid\!\sim_{\mathcal{U},\mathcal{L}}$ entails that $\alpha \cong_{\mathcal{U},\mathcal{L}} \beta$ whenever $\alpha \equiv \beta$.

2.2 A brief reminder of Kraus, Lehmann & Magidor's framework.

We remind the definitions of cumulative and preferential models and their corresponding induced inference relations, as stated by Kraus, Lehmann and Magidor in [Kraus et al., 1990]. As usual, if $E' \subseteq E$ are sets and $<$ is a binary relation on E, we say that x is $<$-*minimal in* E' iff $x \in E'$ and for any y in E', $y \not< x$. We use $\mathcal{P}$ to denote the power set operator.

Given a propositional language $\mathcal{L}$, a *cumulative model* is a triple $\mathcal{M} = < S, l, \prec >$ where $\mathcal{S}$ is a set (intuitively interpreted as a set of states), $l : \mathcal{S} \longrightarrow \mathcal{P}(\mathcal{W}_\mathcal{L}) - \{\varnothing\}$ is a function that labels every state with a (non-empty) [2] set of complete worlds, and $\prec$ is a binary relation (dubbed 'preference

2 The definition of cumulative models in [Kraus et al., 1990] (def. 5 p. 16) explicitly excludes the possibility of 'empty labels', that is, of states s such that l(s) = $\varnothing$. Yet the completeness' proof (p. 19, between lemmas 10 and 11) makes use of a state s that is labelled with the empty set (s is the equivalence class of $\bot$). It appears that the restriction to non-empty labels is useless, as every cumulative model containing such labels can be turned into an equivalent model without empty labels (we shall not prove this here). So we could modify KLM's definition 5 to allow empty labels. But one can alternatively keep definition 5 as it is and modify instead the completeness' proof p. 19, taking $\mathcal{S} = (\mathcal{L}/\!\sim)$ – $\bot$'s equivalence class in place of $\mathcal{S} = (\mathcal{L}/\!\sim)$. Since the construction suggested by KLM is such that for any state s $\in (\mathcal{L}/\!\sim)$ – $\bot$'s equivalence class, s $\prec \bot$'s equivalence class, one easily checks that suppressing $\bot$'s equivalence class in the model leaves the induced inference relation unaffected. For convenience reasons, it is this second solution that we adopt here.

relation') over states, that satisfies the *smoothness* condition (defined below).

Satisfaction of a formula α by a state s is denoted $s \models\!\!\models \alpha$ and defined by:

$s \models\!\!\models \alpha$ iff for any w in $l(s)$, w $\models \alpha$.

The *smoothness* condition boils down to the following : for any s in $\mathcal{S}$ and any formula α, if $s \models\!\!\models \alpha$ and s isn't $\prec$-minimal in $\{s \in \mathcal{S} / s \models\!\!\models \alpha\}$, then there is an s' in $\mathcal{S}$ such that $s' \prec s$ and s' is $\prec$-minimal in $\{s \in \mathcal{S} / s \models\!\!\models \alpha\}$. For briefness' sake, when s is $\prec$-minimal in $\{s \in \mathcal{S} / s \models\!\!\models \alpha\}$, we say that s is *$\prec$-minimal for α*.

The *inference relation* $\mid\!\!\sim_{\mathcal{M}}$ induced on $\mathcal{L}$ by $\mathcal{M}$ is defined by:

$\alpha \mid\!\!\sim_{\mathcal{M}} \beta$ iff any s $\prec$-minimal for α satisfies β.

KLM show that any inference relation $\mid\!\!\sim_{\mathcal{M}}$ induced by a cumulative model $\mathcal{M}$ satisfies the rules:

Reflexivity	$\alpha \mid\!\!\sim_{\mathcal{M}} \alpha$
Left Equivalence	if $\alpha \equiv \beta$ and $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, then $\beta \mid\!\!\sim_{\mathcal{M}} \gamma$
Right Weakening	if $\alpha \vdash \beta$ and $\gamma \mid\!\!\sim_{\mathcal{M}} \alpha$, then $\gamma \mid\!\!\sim_{\mathcal{M}} \beta$
Cut	if $\alpha \wedge \beta \mid\!\!\sim_{\mathcal{M}} \gamma$ and $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$, then $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$
Cautious Monotony	if $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$ and $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, then $\alpha \wedge \beta \mid\!\!\sim_{\mathcal{M}} \gamma$

and that conversely, any inference relation $\mid\!\!\sim$ that satisfies these rules admits a cumulative model. This set of rules is named *system C*, and relations on the language that satisfy all the rules in C are called *cumulative inference relations*. The authors also give a number of rules that can be derived from C, among which :

And	if $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$ and $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, then $\alpha \mid\!\!\sim_{\mathcal{M}} \beta \wedge \gamma$
Equivalence	if $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$, $\beta \mid\!\!\sim_{\mathcal{M}} \alpha$ and $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, then $\beta \mid\!\!\sim_{\mathcal{M}} \gamma$

Moreover, one readily gets

Supra-classicality	if $\alpha \vdash \beta$, then $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$

from *Reflexivity* and *Right Weakening*.

A *preferential model* is a cumulative model $\mathcal{M} = < S, l, \prec >$ such that states are labelled by singletons and $\prec$ is a strict partial order. KLM show that in addition to the rules from C, the following rule *Or* is valid in all preferential models:

Or if $\alpha \mathrel{|\!\sim_{\mathcal{M}}} \gamma$ and $\beta \mathrel{|\!\sim_{\mathcal{M}}} \gamma$, then $\alpha \vee \beta \mathrel{|\!\sim_{\mathcal{M}}} \gamma$,

and that conversely, any inference relation $|\!\sim$ satisfying the rules in C $\cup \{Or\}$ admits a preferential model. The set of rules C $\cup \{Or\}$ is called *system P*, and the inference relations satisfying all the rules in P are called *preferential relations*.

2.3 Partial worlds models and induced inference relations

Let $\mathcal{L}$ be as above, $\mathcal{U} \subseteq \mathcal{W}_{\mathcal{L}}^{\mathrm{P}}$ and $<$ a binary relation on $\mathcal{U}$. We call $\mathcal{M} = (\mathcal{U}, <)$ a *partial worlds model*.

In the particular case where $\mathcal{L}$ is finite, $\mathcal{W}_{\boldsymbol{L}}^{\mathrm{P}}$ is finite and so are the partial worlds in $\mathcal{W}_{\boldsymbol{L}}^{\mathrm{P}}$. So in this case, $\mathcal{U}$ is a finite set of finite partial worlds. We call partial worlds models such that $\mathcal{U}$ satisfies this condition *finite partial worlds models*.

We say that a partial world w $\in \mathcal{U}$ is $<$-*minimal* for α if w is $<$-minimal in $\mathcal{W}_u(\alpha)$. We define the inference relation $\mathrel{|\!\sim_{\mathcal{M}}}$ induced on $\mathcal{L}$ by $\mathcal{M}$ by:

$\alpha \mathrel{|\!\sim_{\mathcal{M}}} \beta$ iff every w $<$-minimal for α satisfies β.

We say that $<$ is $\mathcal{L}$-*smooth* iff for any w $\in \mathcal{U}$ and any $\mathcal{L}$-formula α, the following holds:

if w $\|\!\models \alpha$ and w isn't $<$-minimal for α, then there is a w' $\in \mathcal{U}$ such that w' $<$ w and w' is $<$-minimal for α.

$\mathcal{L}$-smoothness is the transposition of usual *smoothness* in partial worlds' context. If $\mathcal{M} = (\mathcal{U}, <)$ is a partial worlds model such that $<$ is $\mathcal{L}$-smooth, we call $\mathcal{M}$ a $\mathcal{L}$-*smooth partial worlds model*.

It turns out that any $\mathcal{L}$-smooth partial worlds model can be seen as a cumulative model. Indeed, let $\mathcal{M} = (\mathcal{U}, <)$ be a $\mathcal{L}$-smooth partial worlds model, and $\mid\!\sim_{\mathcal{M}}$ the inference relation induced by $\mathcal{M}$. We may build a cumulative model $\mathcal{M}' = < S, l, \prec >$ such that $\mathcal{S} = \mathcal{U}$, $\prec = <$, and $\mid\!\sim_{\mathcal{M}'} = \mid\!\sim_{\mathcal{M}}$. For this we just have to define l by: for any w $\in \mathcal{S}$, l(w) = {w' $\in \mathcal{W}_{\mathcal{L}}$ / w' $\models$ *Lit*(w)}. Note that l is injective since $\mathcal{U}$ may contain at most one copy of each partial world. We check that for any w $\in \mathcal{S}$ and any formula α, we have w $\|\!\models \alpha$ iff w $\models\!\!\models \alpha$:

w $\|\!\models \alpha$ iff (by $\|\!\models$'s def.) for any w' $\in \mathcal{W}_{\mathcal{L}}$ such that w' $\models$ *Lit*(w), w' $\models \alpha$, *i.e.*

iff for any w' $\in l$(w), w' $\models \alpha$, *i.e.*

iff w $\models\!\!\models \alpha$.

Consequently, since $\prec = <$, w is $<$-minimal for α iff it is $\prec$-minimal for α. So, since $<$ is $\mathcal{L}$-smooth, $\prec$ satisfies *smoothness*, thus $\mathcal{M}'$ is a cumulative model. Moreover, since for any α, {w $\in \mathcal{U}$/w is $<$-minimal for α} = {w $\in \mathcal{S}$/w is $\prec$-minimal for α}, we have $\mid\!\sim_{\mathcal{M}'} = \mid\!\sim_{\mathcal{M}}$.

It follows that if $\mathcal{M} = (\mathcal{U}, <)$ is a $\mathcal{L}$-smooth partial worlds model, then $\mid\!\sim_{\mathcal{M}}$ is a cumulative inference relation, so KLM's system C is sound for $\mathcal{L}$-smooth partial worlds models.

However C is not complete for these same models. For example, the following rule ***(*)*** is valid in $\mathcal{L}$-smooth partial worlds models but can't be derived from C:

(*) For any p and q from $Var(\mathcal{L})$, ($p \mid\!\sim_{\mathcal{M}} \bot$ and $q \mid\!\sim_{\mathcal{M}} \bot$) iff $p \vee q \mid\!\sim_{\mathcal{M}} \bot$

Indeed, for any $\mathcal{L}$-smooth partial worlds model $\mathcal{M} = (\mathcal{U}, <)$ and any propositional variables p and q, we have $p \mid\!\sim_{\mathcal{M}} \bot$ and $q \mid\!\sim_{\mathcal{M}} \bot$ iff $\mathcal{W}_u(p) = \mathcal{W}_u(q) = \varnothing$, iff $\mathcal{W}_u(p \vee q) = \varnothing$, iff $p \vee q \mid\!\sim_{\mathcal{M}} \bot$. That ***(*)*** can't be derived from C follows from the fact that is not valid in cumulative models. For example the cumulative model $\mathcal{M}' = < S, l, \prec >$ such that $\mathcal{S}$ = {s}, l(s) = {({p}, {q}), ({q}, {p})}, and $\prec = \varnothing$ induces an inference relation $\mid\!\sim_{\mathcal{M}'}$ such that $p \mid\!\sim_{\mathcal{M}'} \bot$, $q \mid\!\sim_{\mathcal{M}'} \bot$ and $p \vee q \not\mid\!\sim_{\mathcal{M}'} \bot$ (since s $\not\models p$ and s $\not\models q$, but s $\models\!\!\models p \vee q$).

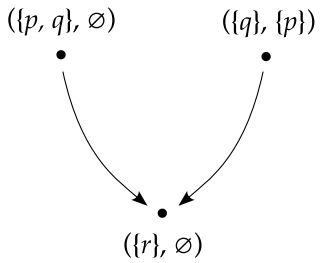

Figure 1: $\mathcal{M}_1 = (\mathcal{U}_1, <_1)$

Yet obviously the rule *Or* is not valid in $\mathcal{L}$-smooth partial worlds models. For example, consider the model $\mathcal{M}_1 = (\mathcal{U}_1, <_1)$, where $\mathcal{U}_1 = \{(\{p, q\}, \varnothing), (\{q\}, \{p\}), (\{r\}, \varnothing)\}$, and $<_1 = \{((\{r\}, \varnothing), (\{p, q\}, \varnothing)), ((\{r\}, \varnothing), (\{q\}, \{p\}))\}$, which may be graphically represented as in Figure 1: we have $p \mathrel{|\!\sim_{\mathcal{M}}} q$ and $\neg p \mathrel{|\!\sim_{\mathcal{M}}} q$, but $p \vee \neg p \not|\!\sim_{\mathcal{M}} q$, since $(\{r\}, \varnothing)$ is $<$-minimal for $p \vee \neg p$.

The question arises of the rules we would need to add to C to get a sound and complete system for $\mathcal{L}$-smooth partial worlds models. Unfortunately, no complete set of rules could be found so far, and in fact there are reasons to believe that reaching completeness is impossible, at least working within a standard propositional language. These reasons shall be discussed in next section. Before that we mention a few rules that even though they don't provide us with completeness, are useful in this context. The following rules are valid in $\mathcal{L}$-smooth partial worlds models:

$\mathcal{U}$-Consistence	if $\alpha \mathrel{|\!\sim_{\mathcal{M}}} \bot$, then $\alpha \mathrel{\|\!\sim_{\mathcal{U},\mathcal{L}}} \bot$
$\mathcal{U}$-Left Equivalence	if $\alpha \cong_{\mathcal{U},\mathcal{L}} \beta$ and $\alpha \mathrel{|\!\sim_{\mathcal{M}}} \gamma$, then $\beta \mathrel{|\!\sim_{\mathcal{M}}} \gamma$
$\mathcal{U}$-Right Weakening	if $\alpha \mathrel{\|\!\sim_{\mathcal{U},\mathcal{L}}} \beta$ and $\gamma \mathrel{|\!\sim_{\mathcal{M}}} \alpha$, then $\gamma \mathrel{|\!\sim_{\mathcal{M}}} \beta$

The proof for $\mathcal{U}$-*Consistence* is immediate using $\mathcal{L}$-smoothness and the definitions of $\mathrel{|\!\sim_{\mathcal{M}}}$ and $\mathrel{\|\!\sim_{\mathcal{U},\mathcal{L}}}$; $\mathcal{U}$-*Left Equivalence* follows from the fact that $\alpha \cong_{\mathcal{U},\mathcal{L}} \beta$ iff $\mathcal{W}_u(\alpha) = \mathcal{W}_u(\beta)$, and $\mathcal{U}$-*Right Weakening* follows from the fact that $\alpha \mathrel{\|\!\sim_{\mathcal{U},\mathcal{L}}} \beta$ iff $\mathcal{W}_u(\alpha) \subseteq \mathcal{W}_u(\beta)$. Since $\mathrel{\|\!\sim_{\mathcal{U},\mathcal{L}}}$ is supra-classical, $\mathcal{U}$-*Left Equivalence* and $\mathcal{U}$-*Right Weakening* are slightly stronger than the original *Left Equivalence* and *Right Weakening*, and can be used in place or these. Similarly, the partial-worlds

version of *Supra-classicality*,

Supra-$\mathcal{U}$-Consequence if $\alpha \parallel\!\!\sim_{\mathcal{U},\mathcal{L}} \beta$, then $\alpha \mathrel{|\!\sim_{\mathcal{M}}} \beta$

that can be derived from *Reflexivity* and *$\mathcal{U}$-Right Weakening*, is slightly stronger than the original *Supra-classicality*.

Chapter 3

The issue of completeness in partial worlds models' context

To investigate the issue of completeness in partial worlds models' context we shall take inspiration from two papers from Gabbay and Schlechta, namely [Gabbay and Schlechta, 2008] and [Gabbay and Schlechta, 2009]. Although the following discussion does not properly speaking make a proof, it nevertheless gives reasons to believe that no sound and complete set of rules can be found for inference relations induced on $\mathcal{L}$ by partial worlds models.

In these papers, the authors use a (not necessarily finite) propositional language $\mathcal{L}$. $\mathcal{M}_\mathcal{L}$ is the set of classical models of $\mathcal{L}$, *i.e.*, in our terminology, of complete worlds for $\mathcal{L}$ (thus Gabbay and Schlechta's $\mathcal{M}_\mathcal{L}$ is our $\mathcal{W}_\mathcal{L}$). For any set T of $\mathcal{L}$-formulas, $M_{(T)} \subseteq \mathcal{M}_\mathcal{L}$ is the set of models of T. $D_\mathcal{L}$ is the set of definable subsets of $\mathcal{M}_\mathcal{L}$, that is $D_\mathcal{L} = \{M_{(T)} / T$ is a set of $\mathcal{L}$-formulas$\}$. A *preferential structure* is a pair $\mathcal{M} = < U, \prec >$, where U is a set and $\prec$ is a binary relation on U. Given a set $\mathcal{Y}$ of sets, a function μ of domain $\mathcal{Y}$ is defined. In the general case where U may contain several 'copies' of some elements (*i.e.*, U is in fact a set of pairs $< x\,,\, i >$, with i an index), μ is defined by: for any $X \in \mathcal{Y}$, $\mu(X) = \{x \in X \,/\, \exists < x\,,\, i > \in U$ and $\neg\exists < x'$, $j > \in U$ s.t. $(x' \in X$ and $< x'$, $j > \prec < x\,,\, i >)\}$. In the particular case where U contains at most one copy of each element, and so we may drop the indexes, $\mu(X)$ is simply

defined as $\{x/x$ is $\prec$-minimal in $X \cap U\}$. The intended interpretation of μ is that of a minimalization function (note however that $\mu(X)$ may be empty, notably in the case where $X \cap \{x /\exists\ i$ s.t. $< x, i > \in U\}$ is). When $\mathcal{Y} \subseteq \mathcal{P}(\mathcal{M}_\mathcal{L})$, μ gives rise to an inference relation $\mid\sim$ in the usual way: for any set T of $\mathcal{L}$-formulas and any $\mathcal{L}$-formula α, $T \mid\sim \alpha$ if and only if $\mu(M_{(T)}) \subseteq M_{(\{\alpha\})}$. Therefore, preferential structures are a generalization of KLM's preferential models. The authors study the algebraic properties of preferential structures and, in cases where $\mathcal{Y} \subseteq \mathcal{P}(\mathcal{M}_\mathcal{L})$, the correspondence between these properties and the usual rules for inference relations (for instance, the rule *Reflexivity* corresponds to the fact that for any $X \in (\mathcal{Y} \cap D_\mathcal{L})$, $\mu(X) \subseteq X$). Among these properties, the closure properties of the domain of μ reveal themselves of special importance. In particular, in [Gabbay and Schlechta, 2008] the authors show that when *smoothness* is not satisfied and the domain of μ is not closed under finite unions, the algebraic counterpart of cumulativity breaks up into an infinity of non-equivalent properties, which jeopardize the chances of getting a representation theorem.

Of course, we can't straightforwardly use their results for our own purposes, since the models we consider satisfy *smoothness* and moreover they are built out of partial worlds. But we can take inspiration from their approach, and in particular we can check for definability and closure properties issues. This requires that we first accommodate their tools to partial worlds models, and first of all that we set forth a suitable notion of definability, which we shall do as follows:

If $\mathcal{L}$ is a (not necessarily finite) propositional language, $\mathcal{U} \subseteq \mathcal{W}_\mathcal{L}^{\mathrm{P}}$ and $A \subseteq \mathcal{U}$, we shall say that A is *$\mathcal{L}$-definable in $\mathcal{U}$* if and only if there is a $\mathcal{L}$-formula α such that $A = \mathcal{W}_u(\alpha)$. For instance, $\{(\{p, q\}, \varnothing), (\{p\}, \{q\})\}$ is $\mathcal{L}$-definable in $\mathcal{U} = \{(\{p, q\}, \varnothing), (\{p\}, \{q\}), (\{r\}, \varnothing)\}$, but not in $\mathcal{U}' = \{(\{p, q\}, \varnothing), (\{p\}, \{q\}), (\{p, r\}, \varnothing)\}$, since for any $\mathcal{L}$-formula α such that $(\{p, q\}, \varnothing) \Vdash \alpha$ and $(\{p\}, \{q\}) \Vdash \alpha$, we also have $(\{p, r\}, \varnothing) \Vdash \alpha$.

We shall denote $D_{\mathcal{U}, \mathcal{L}}$ the set of $\mathcal{L}$-definable (in $\mathcal{U}$) subsets of $\mathcal{U}$. So, form-

ally, $D_{\mathcal{U},\mathcal{L}} = \{A \subseteq \mathcal{U} \ / \ \exists\, \alpha \in \mathcal{L}$ such that $A = \mathcal{W}_u(\alpha)\}$. As we shall always consider the definability of a given subset of $\mathcal{U}$ in $\mathcal{U}$ itself, we shall, for short, speak of $\mathcal{L}$-*definable subsets of* $\mathcal{U}$. $D_{\mathcal{U},\mathcal{L}}$ is closed under finite intersections, since for any $\mathcal{L}$-formulas α and β, $\mathcal{W}_u(\alpha) \cap \mathcal{W}_u(\beta) = \mathcal{W}_u(\alpha \wedge \beta)$. But $D_{\mathcal{U},\mathcal{L}}$ is generally *not* closed under finite unions. For instance, if $\mathcal{U} = \{(\{p, q\}, \varnothing), (\{p\}, \{q\}), (\{p, r\}, \varnothing)\}$, then $\mathcal{W}_u(p \wedge q) \cup \mathcal{W}_u(p \wedge \neg q) = \{(\{p, q\}, \varnothing), (\{p\}, \{q\})\} \notin D_{\mathcal{U},\mathcal{L}}$.

It is well known that a preferential model $< S, l, \prec >$ can be seen as a pair $(U, \prec')$, where $U \subseteq (\mathcal{W}_{\mathcal{L}} \times I)$ (with I an index set) and $\prec'$ is a preference relation on U: one just has to replace each state s in S with a suitably indexed copy w_i of the complete world w such that $\mathrm{w} = l(s)$, and let $\prec' = \prec$. Similarly, a cumulative model $\mathcal{M} = < S, l, \prec >$ can be seen as a pair $(V, \prec'')$, where $V \subseteq \mathcal{P}(\mathcal{W}_{\mathcal{L}})$ and $\prec''$ is a preference relation on V. One just has to consider the cumulative relation $\mid\!\sim_{\mathcal{M}}$ induced by $\mathcal{M}$ and to follow the instructions given by KLM in their representation theorem [1] to build the corresponding cumulative model $\mathcal{M}' = < S', l', \prec' >$ of $\mid\!\sim_{\mathcal{M}}$. Then, since by construction l' is injective, one just has to replace each state s in S' with $l'(s)$, and to let $\prec'' = \prec'$ to get $(V, \prec'')$.

It should also be recalled that KLM's demand that $\prec$ be a strict partial order in preferential models only intends to make things 'nicer' but adds logically nothing, so that in preferential models as in cumulative ones, all that is finally required for $\prec$ is that it satisfies *smoothness*. It follows that what essentially distinguishes preferential from cumulative models is that in the former the preference relation can be seen as defined over a set of (indexed) complete worlds, while in the latter it can be seen as defined over a set of *sets of* complete worlds.

Finally, it is also well known that on the logical side this difference corresponds to the addition of the rule *Or* to C. *Or* can therefore be seen as allowing

1 [Kraus et al., 1990], section 3.5.

the existence of models such that $\prec$ is defined over a set of worlds.

Now, an essential characteristic of partial worlds models is precisely that the relation $<$ holds between partial worlds rather than between sets of (partial) worlds. A complete set of rules for the inference relations they induce should therefore contain some rule(s) corresponding to this property. But as we saw, *Or* is not valid in partial worlds models.

Gabbay and Schlechta's algebraic approach sheds some light on how the rule *Or* produces its effect in the context of preferential relations. In [Gabbay and Schlechta, 2009], they show that the algebraic counterpart of the rule *Or* is the property (μOr):

(μOr) For any X and Y in the domain, $\mu(X \cup Y) \subseteq \mu(X) \cup \mu(Y)$

where μ is the minimalization function. More precisely, they show that for any relation $\mid\sim$ on $\mathcal{L}$ satisfying *Left Equivalence* and *Right Weakening*, the function μ of domain $D_{\mathcal{L}}$ defined by: $\mu(M_{(T)}) = M_{(\overline{\overline{T}})}$ (where $\overline{\overline{T}} = \{\mathcal{L}\text{-formulas } \alpha \,/\, T \mid\sim \alpha\}$) satisfies (μOr) if and only if $\mid\sim$ satisfies *Or* [2].

Furthermore, one easily shows that for any set Z of pairs $< x\,,\, i >$, any binary relation $\prec$ on Z and any set $\mathcal{Y}$ of sets, if μ is a function of domain $\mathcal{Y}$ defined by: for any $X \in \mathcal{Y}$, $\mu(X) = \{x \in X \,/\, \exists < x\,,\, i > \in Z$ and $\neg\exists < x', j > \in Z$ s.t. $(x' \in X$ and $< x', j > \prec < x\,,\, i >)\}$, then μ satisfies (μOr) provided that $\mathcal{Y}$ is closed under finite unions. Indeed let X and Y be in $\mathcal{Y}$, and suppose that there is some $x \in \mu(X \cup Y)$ such that $x \notin \mu(X) \cup \mu(Y)$. By μ's definition, $\mu(X \cup Y) \subseteq X \cup Y$, thus $x \in X$ or $x \in Y$. If $x \in X$, then since $x \notin \mu(X)$, by μ's definition again there is a pair $< x', j >$ in Z such that $x' \in X$ and $< x', j > \prec < x\,,\, i >$. $x' \in X \cup Y$, thus $x \notin \mu(X \cup Y)$: a contradiction. And similarly if $x \in Y$. It follows that for any set of pairs Z, any set $\mathcal{Y}$ of setsand any function μ of domain $\mathcal{Y}$, if μ *falsifies* (μOr), then there can be no binary relation $\prec$ on Z such that for any $X \in \mathcal{Y}$, $\mu(X) = \{x \in X \,/\, \exists < x\,,\, i > \in Z$ and $\neg\exists < x', j > \in Z$ s.t. $(x' \in X$ and $< x', j > \prec < x\,,\, i >)\}$. Simply put, this

2 [Gabbay and Schlechta, 2009], proposition 3.8

means that if μ falsifies (μOr), there can be no relation $\prec$ on Z such that μ is the minimalization function induced by $\prec$. Note that in all cases the closure of $\mathcal{Y}$ under finite unions is required because otherwise $\mu(X \cup Y)$ may not be defined, and then the definition of (μOr) does not make sense. These results can be easily transposed to the particular case where Z contains at most one copy of each element, and thus to the even more particular case where Z is a set of partial worlds.

In the particular case of partial worlds models, Z is $\mathcal{U}$ and $\prec$ is $<$, thus by the above it indeed holds that for any $\mathcal{L}$-formulas α and β, $\{\text{w/w is } <\text{-minimal in } \mathcal{W}_u(\alpha) \cup \mathcal{W}_u(\beta)\} \subseteq \{\text{w/w is } <\text{-minimal in } \mathcal{W}_u(\alpha)\} \cup \{\text{w/w is } <\text{-minimal in } \mathcal{W}_u(\beta)\}$. But because $\mathrel{|\!\sim_{\mathcal{M}}}$ can only grasp $\mathcal{L}$-definable subsets of $\mathcal{U}$, it seems that the relation $\mathrel{|\!\sim_{\mathcal{M}}}$ is unable to express this property. Indeed, any function μ defined by the means of $\mathrel{|\!\sim_{\mathcal{M}}}$ (using $\mu(\mathcal{W}_u(\alpha)) = \mathcal{W}_u(\{\overline{\overline{\alpha}}\})$ with $\{\overline{\overline{\alpha}}\} = \{\beta/\alpha \mathrel{|\!\sim_{\mathcal{M}}} \beta\}$, or any definition we would like) will always have $D_{\mathcal{U},\mathcal{L}}$ for domain. But $D_{\mathcal{U},\mathcal{L}}$ need not be closed under finite unions, and in cases it is not (μOr) will not be defined for the whole $D_{\mathcal{U},\mathcal{L}}$ but only for the subsets of $D_{\mathcal{U},\mathcal{L}}$ that are closed under finite unions. It can be assumed that a suitably defined function μ will indeed satisfy (μOr) within each of these subsets, but still there will be several ways in which to extend μ so as to close $D_{\mathcal{U},\mathcal{L}}$ under finite unions, some of which do not satisfy (μOr). In other words, μ will not 'say enough' to rule out the possibility that no relation $<$ exists on $\mathcal{U}$ such that μ is a fragment of the minimalization function induced by $<$. It thus seems that whatever the valid rules we may find for $\mathrel{|\!\sim_{\mathcal{M}}}$, these will not, in the general case, be enough to ensure the existence of a suitable relation $<$ on $\mathcal{U}$, hence to get a representation theorem. The existence of a complete set of rules for the inference relations induced on $\mathcal{L}$ by partial worlds models therefore appears very doubtful.

Chapter 4

Augmenting the language

The absence of closure of $D_{\mathcal{U},\mathcal{L}}$ under finite unions can be seen as the consequence of a lack of expressivity of $\mathcal{L}$ with regard to partial worlds. Indeed, using $\mathcal{L}$ one can express the fact that a partial world w satisfies $\alpha \vee \beta$, but not that: it satisfies α or it satisfies β. In the context of complete worlds, the two are equivalent, but not in that of partial worlds. A simple way to close $D_{\mathcal{U},\mathcal{L}}$ under finite unions is thus to add a suitable new connective to the language, so as to allow it to express the fact that a partial world w satisfies α or satisfies β. This is what we shall do in what follows. For simplicity, and because for our purposes we only need finite languages, in doing so we shall only consider finite languages.

Using a finite propositional language $\boldsymbol{L}$ as a basis (so $\boldsymbol{L}$ is a particular instance of the language $\mathcal{L}$ we used so far), in section 4.1 we supplement it with two new binary connectives [1] $\|$ and $\curlywedge$ (it shall be shown in further sections that $\curlywedge$ brings in fact nothing to the expressivity of the augmented language, but is added for convenience). In 4.2 we transpose the previously defined notions to the resulting language $\boldsymbol{L}^{\|}$, and in 4.3 we provide a number of valid rules for inference relations induced on $\boldsymbol{L}^{\|}$ by $\boldsymbol{L}^{\|}$-smooth partial worlds models. Finally in section 4.4 we introduce a couple of useful additional definitions.

1 A symbol $\|$ is also used by Gabbay and Schlechta, with an algebraic meaning. The use we shall make here of this symbol is different, and should not be confused with the one made by these authors (anyway the context of use suffices to disambiguate, since here we use it as a connective).

4.1 The language $\boldsymbol{L}^{\|}$

Let $\boldsymbol{L}$ be a *finite* propositional language. We define the language $\boldsymbol{L}^{\|}$ by:

. If α is a $\boldsymbol{L}$-formula, then α is a $\boldsymbol{L}^{\|}$-formula (*i.e.* $\boldsymbol{L} \subseteq \boldsymbol{L}^{\|}$);

. If α and β are $\boldsymbol{L}^{\|}$-formulas, then $\alpha \curlywedge \beta$ and $\alpha \parallel \beta$ are $\boldsymbol{L}^{\|}$-formulas;

. Nothing else is a $\boldsymbol{L}^{\|}$-formula.

Satisfaction of a $\boldsymbol{L}^{\|}$-formula α by a partial world w is denoted w $\|\!\models \alpha$ as before. It is defined by the following clauses:

. If α is a $\boldsymbol{L}$-formula, then w $\|\!\models \alpha$ iff $\delta(\mathrm{w}) \vdash \alpha$ (as before);

. If $\alpha = \beta \parallel \gamma$, then w $\|\!\models \alpha$ iff (w $\|\!\models \beta$ or w $\|\!\models \gamma$);

. If $\alpha = \beta \curlywedge \gamma$, then w $\|\!\models \alpha$ iff (w $\|\!\models \beta$ and w $\|\!\models \gamma$).

As before, we note $\mathcal{W}_u(\alpha)$ the set of partial worlds $\mathrm{w} \in \mathcal{U}$ such that w $\|\!\models \alpha$. In the particular case where w is a complete world, one easily checks that w $\|\!\models \beta \parallel \gamma$ iff w $\models \beta \vee \gamma$ iff w $\|\!\models \beta \vee \gamma$. In the general case, w $\|\!\models \beta \parallel \gamma$ implies w $\|\!\models \beta \vee \gamma$ but not conversely. Furthermore, $\parallel$ is associative: for any $\boldsymbol{L}^{\|}$-formulas α, β and γ, and any partial world w, w $\|\!\models \alpha \parallel (\beta \parallel \gamma)$ iff w $\|\!\models (\alpha \parallel \beta) \parallel \gamma$. Consequently in the sequel we may drop the brackets, and simply write $\alpha \parallel \beta \parallel \gamma$. If $X = \{\alpha_1, \alpha_2, \ldots \alpha_n\}$ is a set of $\boldsymbol{L}^{\|}$-formulas, we may write $\big\|\alpha_i/(\alpha_i \in X)$ as an abbreviation for the $\boldsymbol{L}^{\|}$-formula $\alpha_1 \parallel \alpha_2 \ldots \parallel \alpha_n$. Finally, it directly follows from the definition of $\parallel$ that $\mathcal{W}_u(\alpha \parallel \beta) = \mathcal{W}_u(\alpha) \cup \mathcal{W}_u(\beta)$.

As for the new connective $\curlywedge$, for now we may just remark that it operates on $\boldsymbol{L}^{\|}$-formulas exactly as the usual $\wedge$ does on $\boldsymbol{L}$-formulas, since for any $\boldsymbol{L}$-formulas α and β and any partial world w, w $\|\!\models \alpha \wedge \beta$ iff w $\|\!\models \alpha$ and w $\|\!\models \beta$. $\curlywedge$ can thus be seen as an extension of the usual $\wedge$ to formulas containing $\parallel$.

One may wonder whether the introduction of the connective $\parallel$ justifies itself by more than purely technical reasons, that is if $\parallel$ has some natural intuitive interpretation in the framework. It turns out that $\parallel$ is the syntactical counterpart of a mental operation that is commonly performed on information by

natural agents. Namely, $\|$-disjunctions appear to be the primary form in which disjunction takes place in living beings.

The fact is that each time an agent considers different cases or options (as, for instance, to decide which case is the most plausible, or which action is to be taken), it forms a disjunction of this kind. Indeed, to be uncertain whether a given object/situation satisfies the conjunction of features α or the conjunction of features β is nothing but to believe that it satisfies at least one of these two, that is, that it satisfies $\alpha \,\|\, \beta$. To decide which of these two possibilities is the more likely, the agent feeds its inferential system with the information $\gamma \curlywedge (\alpha \,\|\, \beta)$ (where γ is the conjunction of all the features the agent currently believes the object to satisfy otherwise), and checks for the outcome. Obviously this mental operation cannot be rendered using the usual disjunction $\vee$, since $\gamma \wedge (p \vee \neg\, p) \equiv \gamma \wedge (q \vee \neg\, q) \equiv \gamma$, while being uncertain whether a given object satisfies the feature represented by p rather than its negation $\neg\, p$ is not the same as being uncertain whether it satisfies the feature represented q rather than its negation $\neg\, q$, nor is it the same as not having such hesitations.

It should be stressed that realizing this kind of disjunctions does not require any particular cognitive ability from the agent, since entertaining the idea that a given object satisfies $\alpha \,\|\, \beta$ is simply not to have decided between two competing mental representations of the object, the one according to which it satisfies α and the other according to which it satisfies β. According to the neural model suggested in section 1.1.2, $\|$-disjunctions are straightforwardly implemented in the agent's inferential system as unions of sets (assemblies) of neurons. More specifically, if A is the set of neurons that represent the information α in the agent's inferential system (that is, if A is the set of its concept neurons that support mental representations of objects that, in its opinion, satisfy α) and if similarly B is the set of its concept neurons that represent the information β, then $\alpha \,\|\, \beta$ is simply represented by $A \cup B$. By contrast, $\vee$-disjunctions cannot be represented by set-theoretical operations over concept

neurons assemblies, which makes it very likely that they can only be implemented at the language level, provided of course that the considered agent has language abilities. It is thus probable that $\|$-disjunctions are the primary form in which disjunction occurs in natural agents, and that $\vee$-disjunctions are on the contrary a secondary construct relying on higher cognitive abilities. This renders the use of the connective $\|$ very natural in our context.

Interestingly, it turns out that our satisfaction clauses for $\boldsymbol{L}^{\|}$-formulas are similar to the satisfaction clauses for $\boldsymbol{L}$-formulas given by Kripke for intuitionistic models [2]. More precisely, a partial world $\mathrm{w} \in \mathcal{W}_{\boldsymbol{L}}^{\mathrm{P}}$ can be mimicked using a Kripke model $\Phi\colon (Var(\boldsymbol{L}) \times \mathbf{K}) \longrightarrow \{\mathrm{T}, \mathrm{F}\}$ defined on a model structure $(\mathbf{G}, \mathbf{K}, \boldsymbol{R})$ such that $\mathbf{G} = \mathrm{w}$, $\mathbf{K} = \{\mathrm{w}\} \cup \{\mathrm{w'} \in \mathcal{W}_{\boldsymbol{L}} \ / \ \delta(\mathrm{w'}) \vdash \delta(\mathrm{w})\}$, $\boldsymbol{R}$ is the reflexive closure of $\{(\mathrm{w}, \mathrm{w'}) \ / \ \mathrm{w'} \in \mathbf{K}\}$ and for any $p \in Var(\boldsymbol{L})$ and any $\mathrm{w'} \in \mathbf{K}$, $\Phi(p, \mathrm{w'}) = \mathrm{T}$ iff $\boldsymbol{w}'(p) = 1$, and $\Phi(p, \mathrm{w'}) = \mathrm{F}$ otherwise. Indeed it is immediate that for any $p \in Var(\boldsymbol{L})$, $\mathrm{w} \|\!\models p$ iff $\Phi(p, \mathrm{w}) = \mathrm{T}$, and using Kripke's clauses to extend Φ to formulas one gets that if α is a $\boldsymbol{L}$-formula that contain no other connectives than $\neg$ and $\wedge$, then $\mathrm{w} \|\!\models \alpha$ iff $\Phi(\alpha, \mathrm{w}) = \mathrm{T}$ (the proof is by induction on the construction of formulas. Then using our satisfaction clause for $\|$ on the one side and Kripke's clause for $\vee$ on the other, one gets that if β is a $\boldsymbol{L}$-formula that contain no other connectives than $\neg$ and $\wedge$, then $\mathrm{w} \|\!\models \alpha \ \| \ \beta$ iff $\Phi(\alpha \vee \beta, \mathrm{w}) = \mathrm{T}$. Thus the disjunction denoted by $\|$ can be seen in a certain sense as 'intuitionistic'. However, and contrary to what is the general case in Kripke models, here it holds that $\Phi(\alpha, \mathrm{w}) = \mathrm{T}$ iff $\Phi(\neg\neg\alpha, \mathrm{w}) = \mathrm{T}$, that is, $\mathrm{w} \|\!\models \alpha$ iff $\mathrm{w} \|\!\models \neg\neg\alpha$.

4.2 Transposition of the previous definitions in $\boldsymbol{L}^{\|}$-context

We shall now transpose the previously defined notions — $\mathcal{U}$-consequence, definability, *smoothness*, inference relations and so on — to $\boldsymbol{L}^{\|}$-context.

2 [Kripke, 1965] pp. 94 – 100. This similarity was suggested by K. Schlechta in a personal communication.

Let $\mathcal{U} \subseteq \mathcal{W}_{\boldsymbol{L}}^{\mathrm{P}}$. Since $\boldsymbol{L}$ is finite, $\mathcal{U}$ is a finite set of finite partial worlds. We extend the relation of *$\mathcal{U}$-consequence* to $\boldsymbol{L}^{\shortparallel}$:

For any $\boldsymbol{L}^{\shortparallel}$-formulas α and β, $\alpha \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \beta$ iff $\mathcal{W}_u(\alpha) \subseteq \mathcal{W}_u(\beta)$

It is immediate that $\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}$ is transitive, and that $\|\!\!\sim_{\mathcal{U},\boldsymbol{L}} \subseteq \|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}$. Note that:

- $\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}$ is supra-classical, since $\|\!\!\sim_{\mathcal{U},\boldsymbol{L}}$ is.
- For any $\boldsymbol{L}^{\shortparallel}$ formula α, $\alpha \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \bot$ iff $\mathcal{W}_u(\alpha) = \varnothing$.
- For any $\boldsymbol{L}^{\shortparallel}$ formulas α, β and γ, $\alpha \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \gamma$ and $\beta \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \gamma$ iff $\alpha \parallel \beta \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \gamma$.
- For any $\boldsymbol{L}^{\shortparallel}$ formulas α, β and γ, $\alpha \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \beta$ and $\alpha \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \gamma$ iff $\alpha \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \beta \curlywedge \gamma$.
- For any $\mathrm{w} \in \mathcal{U}$ and any $\boldsymbol{L}^{\shortparallel}$ formula α, $\delta(\mathrm{w}) \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \alpha$ iff $\mathrm{w} \Vdash \alpha$ (see proof in Appendix A). Thus, in the particular case where α is a $\boldsymbol{L}$-formula, $\delta(\mathrm{w}) \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \alpha$ iff $\mathrm{w} \Vdash \alpha$ iff $\delta(\mathrm{w}) \vdash \alpha$.

Similarly, we extend the relation of *$\mathcal{U}$-equivalence* to $\boldsymbol{L}^{\shortparallel}$:

For any $\boldsymbol{L}^{\shortparallel}$-formulas α and β, $\alpha \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \beta$ iff $\alpha \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \beta$ and $\beta \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}\, \alpha$, that is, iff $\mathcal{W}_u(\alpha) = \mathcal{W}_u(\beta)$.

As before for $\cong_{\mathcal{U},\boldsymbol{L}}$, supra-classicality of $\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}}$ entails that $\alpha \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \beta$ whenever $\alpha \equiv \beta$. Furthermore, one easily checks that:

- $(\alpha \parallel \beta) \parallel (\beta \parallel \gamma) \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \alpha \parallel \beta \parallel \gamma$.
- $(\alpha \parallel \beta \parallel \gamma) \curlywedge (\, \beta \parallel \gamma) \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \beta \parallel \gamma$.
- $\bot \parallel \alpha \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \alpha$.
- If $\alpha \not\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \bot$, then $\alpha \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \big\|\delta(\mathrm{w})/(\mathrm{w} \in \mathcal{W}_u(\alpha))$.
- If $\alpha \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \alpha$', then $\alpha \parallel \beta \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \alpha' \parallel \beta$ ('substitution of $\mathcal{U}$-equivalent disjuncts').
- If α and β are $\boldsymbol{L}$-formulas, then $\alpha \curlywedge \beta \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \alpha \wedge \beta$.
- $(\alpha_1 \parallel \alpha_2) \curlywedge \beta \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} (\alpha_1 \curlywedge \beta) \parallel (\alpha_2 \curlywedge \beta)$

Note that the last three facts taken together entail that any $\boldsymbol{L}^{\shortparallel}$-formula α containing $\curlywedge$ is $\mathcal{U}$-equivalent to a $\boldsymbol{L}^{\shortparallel}$-formula α' that doesn't contain $\curlywedge$, but contains the usual $\wedge$ instead.

Given $A \subseteq \mathcal{U}$, we shall say that A is $\boldsymbol{L}^{\shortparallel}$-*definable* (*in* $\mathcal{U}$) iff there is a $\boldsymbol{L}^{\shortparallel}$-formula α such that $A = \mathcal{W}_u(\alpha)$. We denote $D_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}$ the set of $\boldsymbol{L}^{\shortparallel}$-definable (in $\mathcal{U}$) subsets of $\mathcal{U}$. As we shall always consider the $\boldsymbol{L}^{\shortparallel}$-definability of a given subset of $\mathcal{U}$ in $\mathcal{U}$ itself, we shall, for short, speak of $\boldsymbol{L}^{\shortparallel}$-*definable subsets of* $\mathcal{U}$. Since $\boldsymbol{L} \subseteq \boldsymbol{L}^{\shortparallel}$, we have $D_{\mathcal{U}, \boldsymbol{L}} \subseteq D_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}$. Note that even though $\mathcal{U}$ is finite, generally $D_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}} \neq \mathcal{P}(\mathcal{U})$. For instance, let $\mathcal{U} = \{(\{p,\, q\},\, \varnothing),\, (\{p\},\, \varnothing)\}$; then $\{(\{p\},\, \varnothing)\} \notin D_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}$, since any $\boldsymbol{L}^{\shortparallel}$ formula that is satisfied by $(\{p\},\, \varnothing)$ is also satisfied by $(\{p,\, q\},\, \varnothing)$. But contrary to $D_{\mathcal{U}, \boldsymbol{L}}$, $D_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}$ is closed under finite unions: if A and $B \in D_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}$, that is if there are some $\boldsymbol{L}^{\shortparallel}$-formulas α and β such that $A = \mathcal{W}_u(\alpha)$ and $B = \mathcal{W}_u(\beta)$, then $A \cup B = \mathcal{W}_u(\alpha) \cup \mathcal{W}_u(\beta) = \mathcal{W}_u(\alpha \parallel \beta)$, so $A \cup B \in D_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}$.

Moreover, since any $\boldsymbol{L}^{\shortparallel}$-formula α containing $\curlywedge$ is $\mathcal{U}$-equivalent to a $\boldsymbol{L}^{\shortparallel}$-formula α' that does not contain $\curlywedge$, any subset of $\mathcal{U}$ which is definable by a $\boldsymbol{L}^{\shortparallel}$-formula containing $\curlywedge$ is also definable by a $\boldsymbol{L}^{\shortparallel}$-formula not containing $\curlywedge$. Thus $\curlywedge$ adds nothing to the expressivity of $\boldsymbol{L}^{\shortparallel}$, and from a logical point of view it would have been strictly equivalent to define $\boldsymbol{L}^{\shortparallel}$ by the sole addition of $\parallel$. We did not for convenience reasons, since this will allow us to adapt to $\boldsymbol{L}^{\shortparallel}$ the usual rules for inference relations. Now, as $\curlywedge$ is but an extension of the usual $\wedge$ to formulas containing $\parallel$, in the sequel we shall drop the special writing $\curlywedge$ to denote conjunctions of formulas containing $\parallel$, and use indifferently $\wedge$ for all conjunctions of $\boldsymbol{L}^{\shortparallel}$ formulas.

Let $<$ be a binary relation on $\mathcal{U}$. We denote $\parallel\!\sim_{\mathcal{M}}$ the inference relation induced on $\boldsymbol{L}^{\shortparallel}$ by $\mathcal{M} = (\mathcal{U}, <)$ in the usual manner, that is:

> For any $\boldsymbol{L}^{\shortparallel}$-formulas α and β, $\alpha \parallel\!\sim_{\mathcal{M}} \beta$ iff every w $<$-minimal in $\mathcal{W}_u(\alpha)$ satisfies β.

We say that $<$ is $\boldsymbol{L}^{\shortparallel}$-*smooth* if and only if for any w in $\mathcal{U}$ and any $\boldsymbol{L}^{\shortparallel}$-formula α, the following holds:

> If w $\parallel\!\models \alpha$ and w isn't $<$-minimal in $\mathcal{W}_u(\alpha)$, then there is a w' in $\mathcal{U}$ such that w' $<$ w and w' is $<$-minimal in $\mathcal{W}_u(\alpha)$.

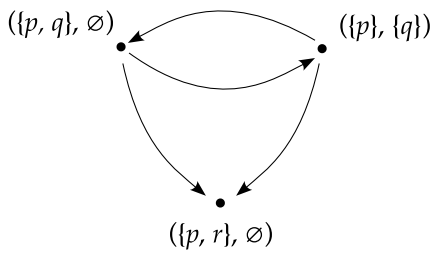

Figure 2: $\mathcal{M}_2 = (\mathcal{U}_2, <_2)$

Since $\boldsymbol{L} \subseteq \boldsymbol{L}^{\shortparallel}$, any $\boldsymbol{L}^{\shortparallel}$-smooth relation is also $\boldsymbol{L}$-smooth, but not conversely. For instance, consider the model $\mathcal{M}_2 = (\mathcal{U}_2, <_2)$ showed in figure 2: $<$ is $\boldsymbol{L}$-smooth, since any $\boldsymbol{L}$-formula satisfied by both $(\{p, q\}, \varnothing)$ and $(\{p\}, \{q\})$ is also satisfied by $(\{p, r\}, \varnothing)$. But $<$ is not $\boldsymbol{L}^{\shortparallel}$-smooth, since there is no $<$-minimal world for $(p \wedge q) \,\|\, (p \wedge \neg q)$.

If $\mathcal{M} = (\mathcal{U}, <)$ is a partial worlds model such that $<$ is $\boldsymbol{L}^{\shortparallel}$-smooth, we call $\mathcal{M}$ a $\boldsymbol{L}^{\shortparallel}$-*smooth partial worlds model*. Any $\boldsymbol{L}^{\shortparallel}$-smooth model is a $\boldsymbol{L}$-smooth model, but not conversely.

4.3 Inference relations induced on $\boldsymbol{L}^{\shortparallel}$ by $\boldsymbol{L}^{\shortparallel}$-smooth partial worlds models

We say that a $\boldsymbol{L}^{\shortparallel}$-formula α is *maximal-consistent in* $\mathcal{U}$ if

i) $\alpha \in \boldsymbol{L}$,

ii) $\alpha \,\|\!\not\sim_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}} \bot$, and

iii) For any $\boldsymbol{L}$-formula β such that $\alpha \nvdash \beta$, $\alpha \wedge \beta \,\|\!\sim_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}} \bot$.

For any formula α that is maximal-consistent in $\mathcal{U}$, there is a partial world $\mathrm{w} \in \mathcal{U}$ such that $\delta(\mathrm{w}) \equiv \alpha$. Indeed, if α is maximal-consistent in $\mathcal{U}$, then by the clause ii) there is a $\mathrm{w} \in \mathcal{U}$ such that $\mathrm{w} \,\|\!\models \alpha$, iff (by the first clause) $\delta(\mathrm{w}) \vdash \alpha$. So $\alpha \wedge \delta(\mathrm{w}) \equiv \delta(\mathrm{w}) \,\|\!\not\sim_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}} \bot$, which by clause iii) entails that $\alpha \vdash \delta(\mathrm{w})$.

Let $\mathcal{M} = (\mathcal{U}, <)$ be a $\boldsymbol{L}^{\shortparallel}$-smooth partial worlds model. $\|\!\sim_{\mathcal{M}}$ satisfies the rules:

Reflexivity	$\alpha \mid\!\!\sim_{\mathcal{M}} \alpha$
(***L***$^{\|}$) ***$\mathcal{U}$-Left Equivalence***	if $\alpha \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \beta$ and $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, then $\beta \mid\!\!\sim_{\mathcal{M}} \gamma$
(***L***$^{\|}$) ***$\mathcal{U}$-Right Weakening***	if $\alpha \mid\!\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \beta$ and $\gamma \mid\!\!\sim_{\mathcal{M}} \alpha$, then $\gamma \mid\!\!\sim_{\mathcal{M}} \beta$
Cut	if $\alpha \wedge \beta \mid\!\!\sim_{\mathcal{M}} \gamma$ and $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$, then $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$
Cautious Monotony	if $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$ and $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, then $\alpha \wedge \beta \mid\!\!\sim_{\mathcal{M}} \gamma$
(***L***$^{\|}$) ***$\mathcal{U}$-Consistence***	if $\alpha \mid\!\!\sim_{\mathcal{M}} \bot$, then $\alpha \mid\!\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \bot$
||-Or	if $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$ and $\beta \mid\!\!\sim_{\mathcal{M}} \gamma$, then $\alpha \,\|\, \beta \mid\!\!\sim_{\mathcal{M}} \gamma$
Injectivity	for any formula α maximal-consistent in $\mathcal{U}$, if $\alpha \,\|\, \beta \,\|\, \gamma \mid\!\!\sim_{\mathcal{M}} \beta \,\|\, \gamma$ and $\alpha \,\|\, \beta \not\mid\!\!\sim_{\mathcal{M}} \beta$, then $\alpha \,\|\, \gamma \mid\!\!\sim_{\mathcal{M}} \gamma$

The proof for *Reflexivity* is trivial. Those for $\mathcal{U}$*-Left Equivalence* and $\mathcal{U}$*-Right Weakening* are similar to those given in section 2.3 above for $\mid\!\!\sim_{\mathcal{M}}$ (one just needs to replace $\mathcal{L}$ with $\boldsymbol{L}^{\|}$, and generally $\mathcal{L}$-notions by the corresponding $\boldsymbol{L}^{\|}$-notions). Those for *Cut* and *Cautious Monotony* are similar to those given in [Kraus et al., 1990][3]. The proof for $\mathcal{U}$*-Consistence* is immediate using $\boldsymbol{L}^{\|}$*-smoothness* and the definitions of $\mid\!\!\sim_{\mathcal{M}}$ and $\mid\!\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}$. For $\|$*-Or*, let w be $<$-minimal for $\alpha \,\|\, \beta$: either w $\|\!\models \alpha$, or w $\|\!\models \beta$. If w $\|\!\models \alpha$, then it is $<$-minimal for α (for suppose not, then there is w' $<$ w such that w' $\|\!\models \alpha$, so w is not $<$-minimal for $\alpha \,\|\, \beta$: contradiction). Since by hypothesis $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, w $\|\!\models \gamma$. Similarly, if w $\|\!\models \beta$, then w $\|\!\models \gamma$. The proof for *Injectivity* is given in Appendix B.

Injectivity is thus named because it is not valid in structures $\mathcal{S} = (\mathcal{V}, <)$ such that $\mathcal{V}$ contains several copies of a same partial world, which corresponds in KLM's framework to the fact that the labelling function is not injective. For example, let $\mathcal{V} = \{(\{p, q\}, \varnothing)_1, (\{p, q\}, \varnothing)_2, (\{r\}, \varnothing), (\{s\}, \varnothing)\}$. Consider the structure $\mathcal{S} = (\mathcal{V}, <)$ in Figure 3, and define $\mid\!\!\sim_{\overline{\mathcal{V},\boldsymbol{L}}}$, $\mid\!\!\sim_{\mathcal{S}}$, ... etc. as expected. $p \wedge q$ is maximal-consistent in $\mathcal{V}$, yet we have $(p \wedge q) \,\|\, r \,\|\, s \mid\!\!\sim_{\mathcal{S}} r \,\|\, s$ and $(p \wedge q) \,\|\, r \not\mid\!\!\sim_{\mathcal{S}} r$, but $(p \wedge q) \,\|\, s \not\mid\!\!\sim_{\mathcal{S}} s$.

3 [Kraus et al., 1990], p. 18

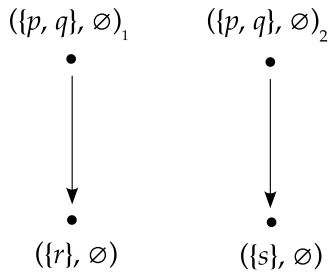

Figure 3: $\mathcal{S} = (\mathcal{V}, <)$

The meaning of *Injectivity* is best seen considering the equivalent writing of the rule:

For any formula α maximal-consistent in $\mathcal{U}$,

if $\alpha \parallel \beta \mathrel{|\!\sim_{\mathcal{M}}} \beta$ and $\alpha \parallel \gamma \mathrel{|\!\sim_{\mathcal{M}}} \gamma$, then $\alpha \parallel \beta \parallel \gamma \mathrel{|\!\sim_{\mathcal{M}}} \beta \parallel \gamma$

Since α is maximal-consistent in $\mathcal{U}$, there is some partial world w in $\mathcal{U}$ such that $\alpha \equiv \delta(\mathrm{w})$. According to the intended interpretation of $\mathcal{M}$, w stands for some mental representation a in the agent's mind, so w can be denoted w_a. $\delta(\mathrm{w}_a) = \delta(\mathrm{w})$ stands for the conjunction of the features that, in the agent's opinion, are satisfied by the objects/situations of which a is the mental representation. Similarly, let $b_1, \ldots, b_n$ be the agent's mental representations such that $\mathcal{W}_u(\beta) = \{\mathrm{w}_{b1}, \ldots, \mathrm{w}_{bn}\}$, and $c_1, \ldots, c_m$ its mental representations such that $\mathcal{W}_u(\gamma) = \{\mathrm{w}_{c1}, \ldots, \mathrm{w}_{cm}\}$. If $\beta \not\cong_{\mathcal{U},\mathbf{L}^{\|}} \bot$ and $\gamma \not\cong_{\mathcal{U},\mathbf{L}^{\|}} \bot$, then $\beta \cong_{\mathcal{U},\mathbf{L}^{\|}} \big\|\delta(\mathrm{w}_{bi})/(\mathrm{w}_{bi} \in \mathcal{W}_u(\beta))$ and $\gamma \cong_{\mathcal{U},\mathbf{L}^{\|}} \big\|\delta(\mathrm{w}_{cj})/(\mathrm{w}_{cj} \in \mathcal{W}_u(\gamma))$, so using substitution of $\mathcal{U}$-equivalent $\|$-disjuncts (see p. 59), *$\mathcal{U}$-Left Equivalence* and *$\mathcal{U}$-Right Weakening*, one gets that

if $\delta(\mathrm{w}_a)\| \delta(\mathrm{w}_{b1})\| \ldots \| \delta(\mathrm{w}_{bn}) \mathrel{|\!\sim_{\mathcal{M}}} \delta(\mathrm{w}_{b1})\| \ldots \| \delta(\mathrm{w}_{bn})$
and $\delta(\mathrm{w}_a)\| \delta(\mathrm{w}_{c1})\| \ldots \| \delta(\mathrm{w}_{cm}) \mathrel{|\!\sim_{\mathcal{M}}} \delta(\mathrm{w}_{c1})\| \ldots \| \delta(\mathrm{w}_{cm})$,
then $\delta(\mathrm{w}_a)\| \delta(\mathrm{w}_{b1})\| \ldots \| \delta(\mathrm{w}_{bn}) \| \delta(\mathrm{w}_{c1})\| \ldots \| \delta(\mathrm{w}_{cm}) \mathrel{|\!\sim_{\mathcal{M}}} \| \delta(\mathrm{w}_{b1})\| \ldots$
$\ldots \| \delta(\mathrm{w}_{bn}) \| \delta(\mathrm{w}_{c1})\| \ldots \| \delta(\mathrm{w}_{cm})$

which can be interpreted as follows: if, when the information that the agent considers is compatible with its mental representations a, $b_1, \ldots, b_n$, it does not overlook a as a possibility, and if when the information that it considers

is compatible with its mental representations a, c_1, ... , c_m it does not overlook a as a possibility, then, when the information it considers is compatible with a, b_1, ..., b_n, c_1, ..., c_m, it does not overlook a as a possibility either. If $\beta \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$, then $\{w_{b1}, \ldots, w_{bn}\} = \varnothing$ and by substitution of $\mathcal{U}$-equivalent $\|$-disjuncts, *$\mathcal{U}$-Left Equivalence* and *$\mathcal{U}$-Right Weakening*, the meaning of *Injectivity* gets trivial. Similarly if $\gamma \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$.

We shall call PW (for 'partial worlds') the above set of rules. Here are some additional rules that can be usefully derived from PW:

($\boldsymbol{L}^{\|}$)***Supra-$\mathcal{U}$-Consequence***

If $\alpha \,|\!|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \beta$, then $\alpha \,|\!|\!\sim_{\mathcal{M}} \beta$

(from *Reflexivity* and *$\mathcal{U}$-Right Weakening*).

Equivalence

If $\alpha \,|\!|\!\sim_{\mathcal{M}} \beta$, $\beta \,|\!|\!\sim_{\mathcal{M}} \alpha$ and $\alpha \,|\!|\!\sim_{\mathcal{M}} \gamma$, then $\beta \,|\!|\!\sim_{\mathcal{M}} \gamma$

(from *Cautious Monotony*, *$\mathcal{U}$-Left Equivalence* and *Cut*).

And

If $\alpha \,|\!|\!\sim_{\mathcal{M}} \beta$ and $\alpha \,|\!|\!\sim_{\mathcal{M}} \gamma$, then $\alpha \,|\!|\!\sim_{\mathcal{M}} \beta \wedge \gamma$

(from *Reflexivity*, *$\mathcal{U}$-Right Weakening*, *Cautious Monotony* and *Cut*).

$\|$-Disjunct Equivalence

If $\alpha_1 \,|\!|\!\sim_{\mathcal{M}} \alpha_2$, $\alpha_2 \,|\!|\!\sim_{\mathcal{M}} \alpha_1$ and $\alpha_1 \,\|\, \beta \,|\!|\!\sim_{\mathcal{M}} \gamma$,then $\alpha_2 \,\|\, \beta \,|\!|\!\sim_{\mathcal{M}} \gamma$

(from *Reflexivity*, *$\mathcal{U}$-Right Weakening*, *Cautious Monotony*, *$\mathcal{U}$-Left Equivalence* and *Equivalence*).

$\|$-Transitivity

If $\alpha_1 \,|\!|\!\sim_{\mathcal{M}} \alpha_2$, ..., $\alpha_{n-1} \,|\!|\!\sim_{\mathcal{M}} \alpha_n$, then $\alpha_1 \,\|\, \alpha_n \,|\!|\!\sim_{\mathcal{M}} \alpha_n$

(from *Reflexivity*, *$\|$-Or*, *$\mathcal{U}$-Right Weakening*,*Cautious Monotony*, *$\mathcal{U}$-Left Equivalence* and *$\|$-Disjunct Equivalence*).

The proof for *Supra-$\mathcal{U}$-Consequence* is immediate. Those for *Equivalence*

and *And* are similar to those given in [Kraus et al., 1990][4]. Those for $\|\sim$-*Disjunct Equivalence* and $\|\sim$-*Transitivity* are given in Appendix C and Appendix D.

4.4 A few additional definitions

Finally, we provide the two additional definitions that shall be useful in the following chapter 5.

4.4.1 The formula $C_{\|\sim}(\alpha)$

Let $\boldsymbol{L}^{\|}/_{\cong_{\mathcal{U},\boldsymbol{L}^{\|}}}$ be the set of all equivalence classes of $\boldsymbol{L}^{\|}$ under $\cong_{\mathcal{U},\boldsymbol{L}^{\|}}$, and g an arbitrary choice function on $\boldsymbol{L}^{\|}/_{\cong_{\mathcal{U},\boldsymbol{L}^{\|}}}$ that selects a representative for each equivalence class. Since $\boldsymbol{L}^{\|}$ is finite, so is $\boldsymbol{L}^{\|}/_{\cong_{\mathcal{U},\boldsymbol{L}^{\|}}}$. For any $\boldsymbol{L}^{\|}$-formula α, we denote $C_{\|\sim}(\alpha)$ the conjunction of all the $\boldsymbol{L}^{\|}$-formulas β such that: $\alpha \,\|\!\sim \beta$ and β is the representative of its equivalence class under $\cong_{\mathcal{U},\boldsymbol{L}^{\|}}$. By a series of applications of the derived rule *And*, one readily shows that for any $\boldsymbol{L}^{\|}$-formula α, $\alpha \,\|\!\sim C_{\|\sim}(\alpha)$. Moreover, for any $\boldsymbol{L}^{\|}$-formulas α and β, $\alpha \,\|\!\sim \beta$ iff $C_{\|\sim}(\alpha) \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \beta$: indeed, if $\alpha \,\|\!\sim \beta$, then β is $\mathcal{U}$-equivalent to some conjunct in $C_{\|\sim}(\alpha)$, so any partial world in $\mathcal{U}$ that satisfies $C_{\|\sim}(\alpha)$ also satisfies β, *i.e.* $C_{\|\sim}(\alpha) \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \beta$. Conversely, suppose that $C_{\|\sim}(\alpha) \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \beta$. Since $\alpha \,\|\!\sim C_{\|\sim}(\alpha)$, by $\mathcal{U}$-*Right Weakening* one gets $\alpha \,\|\!\sim \beta$.

4.4.2 Precisifications and precisification-free partial worlds models

Given $\mathrm{w} \in \mathcal{W}^{\mathrm{p}}_{\boldsymbol{L}}$, a *precisification* of w is a partial world w' $\in \mathcal{W}^{\mathrm{p}}_{\boldsymbol{L}}$ such that $\boldsymbol{w} \subset \boldsymbol{w}$' (where $\boldsymbol{w}$ and $\boldsymbol{w}$' are the truth assignments which generate respectively w and w', and $\subset$ is the strict inclusion). In other words, w' is a precisification of w iff $\delta(\mathrm{w}') \vdash \delta(\mathrm{w})$ and $\delta(\mathrm{w}) \nvdash \delta(\mathrm{w}')$, or equivalently, iff $\delta(\mathrm{w}') \,\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \delta(\mathrm{w})$ and $\delta(\mathrm{w}) \,\|\!\!\not\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \delta(\mathrm{w}')$ (since $\delta(\mathrm{w})$ and $\delta(\mathrm{w})$ are $\boldsymbol{L}$-formulas, see p. 59 above).

4 [Kraus et al., 1990], (p. 14).

We shall say that a set of partial worlds $\mathcal{U}$ is *precisification-free* if and only if for any w and w' in $\mathcal{U}$, w' is *not* a precisification of w.

If $\mathcal{U}$ is precisification-free, then it is immediate that for any w in $\mathcal{U}$, δ(w) is maximal-consistent in $\mathcal{U}$. As (*modulo* classical equivalence) $\{\alpha/\alpha$ is maximal-consistent in $\mathcal{U}\} \subseteq \{\delta(\mathrm{w})/\mathrm{w} \in \mathcal{U}\}$ (cf. p. 61), it follows that, if $\mathcal{U}$ is precisification-free, then (*modulo* classical equivalence) $\{\alpha/\alpha$ is maximal-consistent in $\mathcal{U}\} = \{\delta(\mathrm{w})/\mathrm{w} \in \mathcal{U}\}$.

Moreover, if $\mathcal{U}$ is precisification-free then $D_{\mathcal{U},\boldsymbol{L}^{\|}} = \mathcal{P}(\mathcal{U})$. Indeed, suppose that there is a set $A = \{\mathrm{w}_1 , \ldots , \mathrm{w}_n\} \subseteq \mathcal{U}$ which is not $\boldsymbol{L}^{\|}$-definable. Then, for any $\boldsymbol{L}^{\|}$-formula α such that any w_i in A satisfies α, there is a w' in $\mathcal{U}$ such that w' $\notin A$ and w' $\|\!\models \alpha$. In particular, there is a w' such that w' $\|\!\models \big\|\delta(\mathrm{w}_i)/(\mathrm{w}_i \in A)$ and w' $\notin A$. This implies that there is a $\mathrm{w}_i \in A$ such that w' $\|\!\models \delta(\mathrm{w}_i)$, that is, such that δ(w') $\vdash \delta(\mathrm{w}_i)$. Since w' $\notin A$, w' $\neq \mathrm{w}_i$, thus w' is a precisification of w_i.

We shall say that a partial worlds model $\mathcal{M} = (\mathcal{U}, <)$ is *precisification-free* if and only if $\mathcal{U}$ is precisification-free.

Chapter 5

Two representation theorems

We shall now give two representation theorems. The first one states that PW is sound and complete for inference relations induced on $\boldsymbol{L}^{\|}$ by finite $\boldsymbol{L}^{\|}$-smooth precisification-free models. For the second one, we first introduce a notion of *ranked* partial worlds models, inspired from Lehmann and Magidor's *ranked models* [1], and an additional rule *Rankedness*. Then we show that PW $\cup$ $\{$*Rankedness*$\}$ is sound and complete for inference relations induced on $\boldsymbol{L}^{\|}$ by finite ranked precisification-free partial worlds models [2].

5.1 Representation theorem for finite $\boldsymbol{L}^{\|}$-smooth precisification-free partial worlds models

In section 4.3, we showed that PW is sound for inference relations induced on $\boldsymbol{L}^{\|}$ by finite $\boldsymbol{L}^{\|}$-smooth partial worlds models. *A fortiori* it is sound for those induced by precisification-free finite $\boldsymbol{L}^{\|}$-smooth models. We shall see now that

1 [Lehmann and Magidor, 1992].

2 It has been remarked that both these results would hold just as well if $\mathcal{U}$ was more generally defined as a subset of $\mathcal{P}(\mathcal{W}_{\boldsymbol{L}})$ — we would then have to define a precisification of an element $s \in \mathcal{U}$ as a strict subset of s, $\delta(s)$ as $\bigvee\delta(\mathrm{w})/\mathrm{w} \in s$, and the other notions as expected. Indeed, PW and PW $\cup$ $\{$*Rankedness*$\}$ would respectively be sound and complete for inference relations induced on $\boldsymbol{L}^{\|}$ by the resulting finite precisification-free $\boldsymbol{L}^{\|}$-smooth and ranked models (what would change are the properties of $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$). Yet such models would not be usable for our purposes, since elements in $\mathcal{U}$ are meant to stand for mental representations of objects and situations, which should be represented by partial worlds.

PW is also complete for the latter. For this we shall show that any relation $\|\!\sim$ satisfying PW and such that $\|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}$ is induced by a precisification-free subset $\mathcal{U}$ of $\mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$ admits a precisification-free finite $\boldsymbol{L}^{\|}$-smooth partial worlds model.

Let $\boldsymbol{L}$, $\boldsymbol{L}^{\|}$ and $\mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$ be as above, and let $\mathcal{U}$ be a precisification-free subset of $\mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$, $\|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}$ the corresponding $\mathcal{U}$-consequence relation on $\boldsymbol{L}^{\|}$, and $\|\!\sim$ a binary relation on $\boldsymbol{L}^{\|}$ satisfying PW. Since $\boldsymbol{L}$ is finite, $\mathcal{U}$ is a finite set of finite partial worlds. We define a relation $<$ on $\mathcal{U}$ by: for any w and w' in $\mathcal{U}$, w $<$ w' iff w $\neq$ w' and δ(w')$\|$ δ(w) $\|\!\sim \delta$(w). It is immediate that $\mathcal{M} = (\mathcal{U}, <)$ is a finite precisification-free partial worlds model. We show that $<$ is $\boldsymbol{L}^{\|}$-smooth:

Lemma 1 *For any* w $\in \mathcal{U}$, w $\|\!\models C_{\|\!\sim}(\delta(\mathrm{w}))$

Proof:

Let w $\in \mathcal{U}$:

i) δ(w) $\|\!\not\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \bot$, so by $\mathcal{U}$-*Consistence*,
δ(w) $\|\!\not\sim \bot$, iff
$C_{\|\!\sim}(\delta(\mathrm{w}))$ $\|\!\not\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \bot$, iff
$\exists$ w' $\in \mathcal{U}$ such that w' $\|\!\models C_{\|\!\sim}(\delta(\mathrm{w}))$.

ii) By *Reflexivity*, δ(w) $\|\!\sim \delta$(w), iff
$C_{\|\!\sim}(\delta(\mathrm{w}))$ $\|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \delta$(w), thus by i),
w' $\|\!\models \delta$(w).
Thus w' $\|\!\models \delta(\mathrm{w}) \wedge \delta(\mathrm{w}')$, thus
$\delta(\mathrm{w}) \wedge \delta(\mathrm{w}')$ $\|\!\not\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \bot$.

iii) By hypothesis $\mathcal{U}$ is precisification-free, so δ(w) and δ(w') both are maximal-consistent. By definition of maximal-consistency, ii) implies that $\delta(\mathrm{w}) \equiv \delta(\mathrm{w}')$, iff w = w'. By i), this entails that w $\|\!\models C_{\|\!\sim}(\delta(\mathrm{w}))$.

□

Lemma 2 *$<$ is $\boldsymbol{L}^{\|}$-smooth.*

Proof:

i) By its definition, $<$ is irreflexive.

ii) $<$ is transitive: indeed, suppose that there are w_1, w_2 and w_3 in $\mathcal{U}$ such that $w_3 < w_2 < w_1$. Then:

1. By definition of $<$, $\delta(w_1)\| \delta(w_2) \mid\!\sim \delta(w_2)$, and $\delta(w_2)\| \delta(w_3) \mid\!\sim \delta(w_3)$.

2. By *Reflexivity*, $\delta(w_1) \mid\!\sim \delta(w_1)$, so by *$\mathcal{U}$-Right Weakening* $\delta(w_1) \mid\!\sim \delta(w_1)\| \delta(w_2)$.
Similarly, $\delta(w_2) \mid\!\sim \delta(w_2)\| \delta(w_3)$.

3. By $\|$-*Transitivity* on 1 and 2, we get $\delta(w_1)\| \delta(w_3) \mid\!\sim \delta(w_3)$.

4. Furthermore, $w_1 \neq w_3$. Indeed suppose that $w_1 = w_3$:

a) Then $\delta(w_1) = \delta(w_3)$, so by 1, $\delta(w_3)\| \delta(w_2) \mid\!\sim \delta(w_2)$.
So by 2 + *Equivalence*, $C_{\mid\!\sim}(\delta(w_2)) = C_{\mid\!\sim}(\delta(w_2)\| \delta(w_3))$.

b) By *Reflexivity*, $\delta(w_3) \mid\!\sim \delta(w_3)$, so by *$\mathcal{U}$-Right Weakening*, $\delta(w_3) \mid\!\sim \delta(w_2)\| \delta(w_3)$.
By 1 + *Equivalence*, $C_{\mid\!\sim}(\delta(w_3)) = C_{\mid\!\sim}(\delta(w_2)\| \delta(w_3))$.
So by a), $C_{\mid\!\sim}(\delta(w_2)) = C_{\mid\!\sim}(\delta(w_3))$.

c) By *Reflexivity*, $C_{\mid\!\sim}(\delta(w_2)) \|_{\overline{\mathcal{U},\boldsymbol{L}}} \delta(w_2)$ and $C_{\mid\!\sim}(\delta(w_3)) \|_{\overline{\mathcal{U},\boldsymbol{L}}} \delta(w_3)$.

d) By Lemma 1, $\delta(w_2) \|_{\overline{\mathcal{U},\boldsymbol{L}}} C_{\mid\!\sim}(\delta(w_2))$, so by b) + c) + transitivity of $\|_{\overline{\mathcal{U},\boldsymbol{L}}}$, $\delta(w_2) \|_{\overline{\mathcal{U},\boldsymbol{L}}} \delta(w_3)$, iff (since $\delta(w_3)$ is a $\boldsymbol{L}$-formula) $\delta(w_2) \vdash \delta(w_3)$.

e) Similarly, $\delta(w_3) \|_{\overline{\mathcal{U},\boldsymbol{L}}} C_{\mid\!\sim}(\delta(w_3)) = C_{\mid\!\sim}(\delta(w_2)) \|_{\overline{\mathcal{U},\boldsymbol{L}}} \delta(w_2)$, so $\delta(w_3) \vdash \delta(w_2)$.
By d), $\delta(w_2) \equiv \delta(w_3)$, iff $w_2 = w_3$

f) By hyp. $w_3 < w_2$, thus by def. of $<$, $w_2 \neq w_3$, which contradicts e).

5. By 3 and 4, $w_3 < w_1$.

iii) By i) and ii), $<$ is acyclic.

iv) Since $\mathcal{U}$ is finite, it follows from iii) that for any subset $A \subseteq \mathcal{U}$, there is at least one $w \in A$ such that w is $<$-minimal in A. Since $<$ is transitive,

for any w' in A that is not $<$-minimal in A, there is a w $\in A$ such that w is $<$-minimal in A and w $<$ w'. This applies in particular to subsets $A \subseteq \mathcal{U}$ such that A is $\boldsymbol{L}^{\|}$-definable.

□

It follows from Lemma 2 that $\mathcal{M}$ is a finite $\boldsymbol{L}^{\|}$-smooth precisification-free partial worlds model. We now show that $\mathcal{M}$ is a model of $\|\sim$:

Let $\|\sim_{\mathcal{M}}$ be the inference relation induced by $\mathcal{M}$ on $\boldsymbol{L}^{\|}$. We shall show that $\|\sim_{\mathcal{M}} = \|\sim$. As seen in section 4.3, $\|\sim_{\mathcal{M}}$ satisfies PW and its derived rules. Let α be a $\boldsymbol{L}^{\|}$-formula.

Lemma 3 *If $\mathcal{W}_u(\alpha) = \varnothing$, then, for any $\boldsymbol{L}^{\|}$-formula β, we have both $\alpha \|\sim_{\mathcal{M}} \beta$ and $\alpha \|\sim \beta$.*

Proof:

Immediate, using the fact that $\mathcal{W}_u(\alpha) = \varnothing$ iff $\alpha \|\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \bot$, *Supra-$\mathcal{U}$-Consequence* and *$\mathcal{U}$-Right Weakening*.

□

For the remainder of the proof, we shall assume that $\mathcal{W}_u(\alpha) \neq \varnothing$.

Let $X = \{\text{w} \in \mathcal{W}_u(\alpha) \,/\, \forall \text{ w'} \in \mathcal{W}_u(\alpha) \text{ s.t. w'} \neq \text{w}, \delta(\text{w}) \,\|\, \delta(\text{w'}) \not\|\sim \delta(\text{w'})\}$, and $Y = \mathcal{W}_u(\alpha) - X$. By the definition of $<$, X is the set of $<$-minimal elements in $\mathcal{W}_u(\alpha)$. By Lemma 2, $X \neq \varnothing$. Say $X = \{\text{w}_1, \ldots, \text{w}_n\}$.

For each $\text{w}_i \in X$, let $\text{y}_i = \{\text{w} \in Y / \delta(\text{w})\| \, \delta(\text{w}_i) \, \|\sim \delta(\text{w}_i)\}$. By the definition of $<$, y_i is the set of elements of $\mathcal{W}_u(\alpha)$ that are minimized by w_i (y_i may be empty). By Lemma 2, $\bigcup \text{y}_i/(\text{w}_i \in X) = Y$.

Lemma 4 $C_{\|\sim}(\alpha) \|\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \big\|\delta(\text{w}_k)/(\text{w}_k \in X)$

Proof:

i) For any $\text{w}_i \in X$, using successive applications of *$\|$-Or + $\mathcal{U}$-Left Equivalence*, we get $(\big\|\delta(\text{w}_k)/(\text{w}_k \in \text{y}_i)) \, \|\delta(\text{w}_i) \, \|\sim \delta(\text{w}_i)$.

Thus by *$\mathcal{U}$-Right Weakening*, for any $\text{w}_i \in X$, we get:

$(\big\|\delta(\text{w}_k)/(\text{w}_k \in \text{y}_i)) \, \| \, \delta(\text{w}_i) \, \|\sim \big\|\delta(\text{w}_k)/(\text{w}_k \in X)$.

ii) By a series of applications of $\|$-*Or* + $\mathcal{U}$-*Left Equivalence* on i), we get $(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in Y)) \parallel (\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)) \mid\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$, that is, $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \mathcal{W}_u(\alpha)) \mid\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$.

iii) By hypothesis $\mathcal{W}_u(\alpha) \neq \varnothing$, thus $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \mathcal{W}_u(\alpha)) \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \alpha$.
Thus by $\mathcal{U}$-*Left Equivalence* on ii), $\alpha \mid\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$, iff $C_{\mid\sim}(\alpha) \mid\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$.

□

Lemma 5 $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \mid\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} C_{\mid\sim}(\alpha)$

Proof:

Suppose the contrary, *i.e.* that there is a $\mathrm{w}_k \in X$ such that $\delta(\mathrm{w}_k) \mid\!\not\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} C_{\mid\sim}(\alpha)$. For simplicity's sake, say w_k is w_1.

i) By the derived rule *And*, $\alpha \mid\!\sim C_{\mid\sim}(\alpha)$.

ii) For any $\mathrm{w}_k \in X$, $\mathrm{w}_k \models \alpha$, iff $\delta(\mathrm{w}_k) \mid\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \alpha$.
Thus $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \mid\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \alpha$, so by *Supra-*$\mathcal{U}$*-Consequence*, $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \mid\!\sim \alpha$.

iii) In Lemma 4's proof (item iii) we showed that $\alpha \mid\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$.
Thus by the derived rule *Equivalence* on i) and ii), $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \mid\!\sim C_{\mid\sim}(\alpha)$.

iv) By *Reflexivity*, $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \mid\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$, thus by *And* on iii), $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \mid\!\sim (\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)) \wedge C_{\mid\sim}(\alpha)$.

v) $(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)) \wedge C_{\mid\sim}(\alpha) \cong_{\mathcal{U},\boldsymbol{L}^{\|}}$
$(\delta(\mathrm{w}_1) \wedge C_{\mid\sim}(\alpha)) \parallel (\delta(\mathrm{w}_2) \wedge C_{\mid\sim}(\alpha)) \parallel \ldots \parallel (\delta(\mathrm{w}_n) \wedge C_{\mid\sim}(\alpha))$.

vi) By hypothesis, $\delta(\mathrm{w}_1) \mid\!\not\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} C_{\mid\sim}(\alpha)$, iff $\mathrm{w}_1 \mid\!\not\models C_{\mid\sim}(\alpha)$.
By hypothesis $\mathcal{U}$ is precisification-free, which implies that there is no w in $\mathcal{U}$ such that $\mathrm{w} \models \delta(\mathrm{w}_1) \wedge C_{\mid\sim}(\alpha)$.
Thus $\delta(\mathrm{w}_1) \wedge C_{\mid\sim}(\alpha) \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$.

vii) Necessarily, there is a $\mathrm{w}_2 \in X$ such that $\mathrm{w}_2 \neq \mathrm{w}_1$: for suppose not, then $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) = \delta(\mathrm{w}_1)$, thus by $\mathcal{U}$-*Right Weakening* on iv) and vi), $\delta(\mathrm{w}_1) \mid\!\sim \bot$, so by $\mathcal{U}$-*Consistence*, $\delta(\mathrm{w}_1) \mid\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \bot$, iff $\mathrm{w}_1 \notin \mathcal{U}$, which is absurd.

viii) By v) and vi), $(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)) \wedge C_{\|\sim}(\alpha) \cong_{\mathcal{U},\boldsymbol{L}^{\|}}$
$(\delta(\mathrm{w}_2) \wedge C_{\|\sim}(\alpha))\| \ldots \| (\delta(\mathrm{w}_n) \wedge C_{\|\sim}(\alpha)) \cong_{\mathcal{U},\boldsymbol{L}^{\|}}$
$(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_2, \ldots, \mathrm{w}_n\})) \wedge C_{\|\sim}(\alpha)$.

ix) By $\mathcal{U}$-*Right Weakening* on iv) and viii),
$\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \ \|\!\sim (\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_2, \ldots, \mathrm{w}_n\})) \wedge C_{\|\sim}(\alpha)$,
thus by $\mathcal{U}$-*Right Weakening*,
$\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \ \|\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_2, \ldots, \mathrm{w}_n\})$.

x) $\mathrm{w}_1 \in X$, so by definition of X, $\delta(\mathrm{w}_1) \| \delta(\mathrm{w}_2) \ \|\!\not\sim \delta(\mathrm{w}_2)$.
So by ix), there is a $\mathrm{w}_3 \in X$ such that $\mathrm{w}_1 \neq \mathrm{w}_3 \neq \mathrm{w}_2$.

xi) By hypothesis $\mathcal{U}$ is precisification-free, so for any $\mathrm{w}_i \in X$, $\delta(\mathrm{w}_i)$ is maximal-consistent.
Thus, since $\delta(\mathrm{w}_1)\| \delta(\mathrm{w}_2) \ \|\!\not\sim \delta(\mathrm{w}_2)$, by *Injectivity* on ix) and x),
$\delta(\mathrm{w}_1) \| (\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_3, \ldots, \mathrm{w}_n\})) \ \|\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_3, \ldots, \mathrm{w}_n\})$.
Since for any i $(3 \leq i \leq \mathrm{n})$, $\delta(\mathrm{w}_1) \| \delta(\mathrm{w}_i) \ \|\!\not\sim \delta(\mathrm{w}_i)$, successive applications of the rule *Injectivity* lead to a contradiction.
Thus $\delta(\mathrm{w}_1) \ \|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}}} C_{\|\sim}(\alpha)$.

A similar reasoning can be made for any $\mathrm{w}_k \in X$ (one just has to change the indexation). So, for any w_k in X, $\delta(\mathrm{w}_k) \ \|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} C_{\|\sim}(\alpha)$, iff
$\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \ \|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} C_{\|\sim}(\alpha)$.

□

Lemma 6 *If $\mathcal{W}_u(\alpha) \neq \varnothing$, then, for any $\boldsymbol{L}^{\|}$-formula β, we have $\alpha \ \|\!\sim \beta$ iff $\alpha \ \|\!\sim_{\mathcal{M}} \beta$.*

Proof:

Assume $\mathcal{W}_u(\alpha) \neq \varnothing$. By Lemmas 4 and 5, $C_{\|\sim}(\alpha) \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$. For any $\boldsymbol{L}^{\|}$-formula β, we have:

$\alpha \ \|\!\sim \beta$ iff
$C_{\|\sim}(\alpha) \ \|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \beta$, iff
$\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \ \|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \beta$, iff
for any w_k such that $\mathrm{w}_k \in X$, $\mathrm{w}_k \ \|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \beta$, iff

for any w_k such that w_k is $<$-minimal for α, $w_k \Vdash \beta$, iff
$\alpha \mid\!\!\sim_{\mathcal{M}} \beta$

□

Lemma 7 *$\mathcal{M} = (\mathcal{U}, <)$ is a finite $\boldsymbol{L}^{\|}$-smooth precisification-free partial worlds model such that $\mid\!\!\sim_{\mathcal{M}} = \mid\!\!\sim$.*

Proof:

Immediate, from Lemmas 2-6.

□

Theorem 1 *If $\mid\!\!\sim$ is a relation on $\boldsymbol{L}^{\|}$ satisfying PW and such that $\mid\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$ is induced by a precisification-free subset $\mathcal{U}$ of $\mathcal{W}_{\boldsymbol{L}}^{\mathrm{p}}$, then $\mid\!\!\sim$ admits a finite $\boldsymbol{L}^{\|}$-smooth precisification-free partial worlds model.*

Proof:

Immediate, from Lemma 6.

□

Corollary 1 *PW is sound and complete for inference relations induced on $\boldsymbol{L}^{\|}$ by finite $\boldsymbol{L}^{\|}$-smooth precisification-free partial worlds models.*

Remark: $\|$-*Transitivity* is valid in any $\boldsymbol{L}^{\|}$-smooth model, whether $<$ is transitive or not. However, $\|$-*Transitivity* forces the above construction to be transitive. Thus, for any finite $\boldsymbol{L}^{\|}$-smooth precisification-free model $\mathcal{M}$, there is a transitive (and irreflexive) finite precisification-free model $\mathcal{M}'$, such that $\mid\!\!\sim_{\mathcal{M}} = \mid\!\!\sim_{\mathcal{M}'}$. On the other hand, any strict ordered model is $\boldsymbol{L}^{\|}$-smooth, provided that $\mathcal{U}$ is finite. Thus, although finite strict ordered precisification-free partial worlds models form a strict subclass of finite $\boldsymbol{L}^{\|}$-smooth ones, they give rise to the same set of induced inference relations (contrary to what is the case in the KLM framework, where *smoothness* and strict ordering generate two different sets of induced inference relations, even in the finite case).

5.2 Representation theorem for finite precisification-free ranked models

A binary relation $<$ on a set E is a *modular order* iff $<$ is a strict partial order that satisfies the property [3]

Modularity *(1st version)*:
For any x, y and z in E, if $x < y$, then $z < y$ or $x < z$.

If we write $y \leq z$ to mean that $z \not< y$, then an equivalent formulation of *Modularity* is:

Modularity *(2nd version)*:
For any x, y and z in E, if $x < y$ and $y \leq z$, then $x < z$.

In practice, *Modularity* orders E's elements into ranks.

If $\mathcal{M} = (\mathcal{U}, <)$ is a partial worlds model such that $<$ is a modular order, we say that $\mathcal{M}$ is a *ranked* partial worlds model.

Let $\boldsymbol{L}$ be a finite propositional language, $\mathcal{U} \subseteq \mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$, and $\mathcal{M} = (\mathcal{U}, <)$ a ranked partial worlds model. Since $\mathcal{U}$ is finite and $<$ is a strict partial order, $\mathcal{M}$ is a finite $\boldsymbol{L}^{\shortparallel}$-smooth partial worlds model. Thus $\|\!\sim_{\mathcal{M}}$ satisfies the rules from PW. In addition, $\|\!\sim_{\mathcal{M}}$ satisfies the following rule *Rankedness*:

Rankedness For any $\boldsymbol{L}^{\shortparallel}$-formulas α_1, α_2 and α_3 (all $\not\equiv$) maximal-consistent in $\mathcal{U}$,
if $\alpha_1 \parallel \alpha_2 \|\!\sim_{\mathcal{M}} \alpha_2$ and $\alpha_1 \parallel \alpha_3 \|\!\not\sim_{\mathcal{M}} \alpha_3$, then $\alpha_2 \parallel \alpha_3 \|\!\sim_{\mathcal{M}} \alpha_2$

(see Appendix E for proof). *A fortiori*, if $\mathcal{M} = (\mathcal{U}, <)$ is a precisification-free finite ranked partial worlds model, then $\|\!\sim_{\mathcal{M}}$ satisfies PW $\cup$ {*Rankedness*}.

3 Definition taken from [Lehmann and Magidor, 1992] p. 19.

Conversely, if $\boldsymbol{L}$ is a finite propositional language, $\mathcal{U}$ a precisification-free subset of $\mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$, $\|\!\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$ the $\mathcal{U}$-consequence relation induced on $\boldsymbol{L}^{\|}$ by $\mathcal{U}$ and $\|\!\!\sim$ a relation on $\boldsymbol{L}^{\|}$ satisfying PW $\cup$ $\{$*Rankedness*$\}$, then using the fact that when $\mathcal{U}$ is precisification-free (and *modulo* classical equivalence) $\{\alpha \in \boldsymbol{L}^{\|}/\alpha$ is maximal-consistent in $\mathcal{U}\} = \{\delta(\mathrm{w})/\mathrm{w} \in \mathcal{U}\}$, one easily checks that the construction suggested in section 5.1 above gives rise to a finite ranked precisification-free model. So PW $\cup$ $\{$*Rankedness*$\}$ is sound and complete for inference relations induced on $\boldsymbol{L}^{\|}$ by finite ranked precisification-free partial worlds models.

Chapter 6

Modelling automatic inferences and learning

Finally, we shall give precisions on how partial worlds models can be used to model automatic inferences and learning. First (section 6.1), we shall provide a formal version of the modelling of automatic inferences sketched in section 1.2 above. Then (section 6.2) we shall investigate the learning processes attached to automatic inferences and show how they can be modelled within our framework.

6.1 Modelling automatic inferences

Let $\mathcal{A}$ be the considered cognitive agent. We show how to chose a propositional language $\boldsymbol{L}$ and build a partial worlds model $\mathcal{M} = (\mathcal{U}, <)$ such that for any $\boldsymbol{L}^{\shortparallel}$-formulas α and β, $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$ if and only if $\mathcal{A}$ is disposed to infer β from α, in the sense defined on p. 35 above.

Let $\mathcal{F}^+$ be the set of all the features that $\mathcal{A}$ is physiologically able to perceive. We chose a propositional language $\boldsymbol{L}$ such that there is a bijection σ: $\mathcal{F}^+ \longrightarrow Var(\boldsymbol{L})$. Since $\mathcal{F}^+$ is finite [1], $\boldsymbol{L}$ is finite.

As remarked in section 1.2 above (p. 31), agents are able to conceive the negation of the features they can perceive, and to reason over these negated

1 See p. 31 above.

features just as if they were (negative) features. Let η be the function which to each feature f in $\mathcal{F}^+$ associates its negation *not-f*, and $\mathcal{F}^-$ the image of $\mathcal{F}^+$ by η. $\mathcal{F} = \mathcal{F}^+ \cup \mathcal{F}^-$ is the set of all the features that $\mathcal{A}$ is able to mentally handle.

Now we define a function $\rho\colon \mathcal{F} \longrightarrow \{\lambda \in \boldsymbol{L} \ / \ \lambda \text{ is a literal}\}$:

If $f \in \mathcal{F}^+$, then $\rho(f) = \sigma(f)$, otherwise, $\rho(f) = \neg\sigma(\eta^{-1}(f))$

It is immediate that ρ is a bijection. It defines a relation of representation between the literals from $\boldsymbol{L}$ and the features from $\mathcal{F}$, in the sense that the literal λ shall stand in the model for the feature f if and only if $\lambda = \rho(f)$. One readily checks that for any $p \in Var(\boldsymbol{L})$ and any $f \in \mathcal{F}^+$, $p = \rho(f)$ if and only if $\neg p = \rho(\textit{not-}f)$.

Let $\mathcal{R}$ be the set of $\mathcal{A}$'s most precise mental representations. As specified in section 1.2, we regard $\mathcal{A}$'s mental representations as sets of positive and negative features, which means that for any r in $\mathcal{R}$, $\mathsf{r} \subseteq \mathcal{F}$. In addition, since $\mathcal{R}$ contains only $\mathcal{A}$'s most precise mental representations, it holds that for any r and r' in $\mathcal{R}$, $\mathsf{r} \not\subset \mathsf{r}'$ (where $\subset$ denotes the strict inclusion). We also suppose that $\mathcal{A}$'s mental representations are consistent, *i.e.* that for any r in $\mathcal{R}$ and any f in $\mathcal{F}^+$, $\{f, \textit{not-}f\} \not\subseteq \mathsf{r}$. Obviously, $\mathcal{R}$ is finite, since $\mathcal{F}^+$, hence also $\mathcal{F}$, are.

Let $\rho'\colon \mathcal{R} \longrightarrow \mathcal{W}_{\boldsymbol{L}}^{\mathrm{p}}$ be the function which to each r in $\mathcal{R}$ associates the partial world w such that $\delta(\mathrm{w}) = \bigwedge \rho(f)/f \in \mathsf{r}$. ρ' defines a relation of representation between partial worlds for $\boldsymbol{L}$ and $\mathcal{A}$'s most precise mental representations. Namely, a partial world $\mathrm{w} \in \mathcal{W}_{\boldsymbol{L}}^{\mathrm{p}}$ shall stand for r in the model if and only if $\mathrm{w} = \rho'(\mathsf{r})$.

For any r in $\mathcal{R}$, we assume that a measure of the vividness of r in $\mathcal{A}$'s mind is available, and can be expressed by a real number from the interval $]\,0,\,1\,[$ (see p. 34 above). We define a function $v : \mathcal{R} \longrightarrow]\,0,\,1\,[$ which to each $\mathsf{r} \in \mathcal{R}$ associates the measure of its vividness. We use the open real interval because vividness of memories cannot be arbitrarily high, although no precise maximal value can be found.

We define $\mathcal{U} = \{\text{w} \in \mathcal{W}_{\boldsymbol{L}}^{\text{p}} \,/\, \exists\ r \in \mathcal{R} \text{ s.t. } \text{w} = \rho'(r)\}$. It is immediate that $\mathcal{U}$ is finite and precisification-free. $\mathcal{U}$ will represent $\mathcal{R}$ in the model. Given a partial world w, the fact that $\text{w} \in \mathcal{U}$ means that $\mathcal{A}$ knows some class of objects/situations such that $r = \rho'^{-1}(\text{w})$ is the mental representation of this/these object(s)/situation(s) in its mind. Given a literal λ from $\boldsymbol{L}$, the fact that $\text{w} \Vdash \lambda$ means that according to this mental representation (*i.e.*, in $\mathcal{A}$'s view), the corresponding object(s)/situation(s) satisfy/ies the feature $f = \rho^{-1}(\lambda)$.

As previously said, biological neural networks are rather noisy, and it is likely that small differences in vividness are blurred by neural noise. Hence, a small difference in vividness between two memories will probably not be enough to allow one memory to take advantage over the other in the competition-for-recall process. As a consequence, memories with similar vividness will be equally recalled — or, on the contrary, equally inhibited by a fairly more vivid one. To account for this phenomenon, we shall use a rounding function $rd : \,]\,0,\,1\,[\longrightarrow \,]\,0,\,1\,[$ to erase small differences in vividness. Let us say that rd is the function that rounds to the first n decimals (but other rounding functions might be just as well suited to our needs).

We then define a binary relation $<$ on $\mathcal{U}$:

For any w and w' in $\mathcal{U}$, $\text{w} < \text{w'}$ iff $rd(v(\rho'^{-1}(\text{w}))) >_{\mathbb{R}} rd(v(\rho'^{-1}(\text{w'})))$

where $>_{\mathbb{R}}$ is the usual order on $\,]\,0,\,1\,[$. In other words, $\text{w} < \text{w'}$ if and only if the mental representation w stands for is 'more vivid enough' than the one w' stands for. It is immediate that $<$ is a modular order, thus $\mathcal{M} = (\mathcal{U}, <)$ is a finite precisification-free ranked partial worlds model. $\mathcal{M}$ can be seen as representing the agent's worldview, or, depending on the perspective we take, as representing its inferential system.

Using $\boldsymbol{L}$ as a basis, we define $\boldsymbol{L}^{\Vdash}$ as specified in section 4.1, the meaning of $\Vdash$ being as discussed in this same section. The inference relation $\Vdash_{\mathcal{M}}^{\sim}$ induced by $\mathcal{M}$ on $\boldsymbol{L}^{\Vdash}$ is to be interpreted as the set of $\mathcal{A}$'s dispositions to infer, *i.e.*, as its general knowledge about things. More specifically, if α and β are $\boldsymbol{L}^{\Vdash}$-formulas,

the fact that $\alpha \mid\!\sim_{\mathcal{M}} \beta$ represents the fact that $\mathcal{A}$ is disposed to infer β from α, in the sense specified p. 35 above. As shown in section 5, $\mid\!\sim_{\mathcal{M}}$ is closed under the rules from PW $\cup$ {*Rankedness*}. These rules are properties of the inference relation $\mid\!\sim_{\mathcal{M}}$, hence they should be interpreted as rules that, according to the model, structure the agent's general knowledge about things. They are a by-product of the posited process, and thus they will necessarily hold in any brain running such a process.

6.2 Modelling learning

We now turn to the modelling of learning. In general terms, learning can be defined as a revision of one's knowledge, so as to take into account newly acquired information. In the context of automatic inferences, it boils down to a revision of the agent's dispositions to draw automatic inferences, that is, to a revision of the inference relation $\mid\!\sim_{\mathcal{M}}$. In practice, this means that the information entering the agent's inferential system triggers both the drawing of automatic inferences according to its current dispositions to infer, and a subsequent modification of these same dispositions. This revision can take different forms, depending on how incoming information incorporates with the agent's current knowledge. In this section we shall first review the different kinds of learning, and try for each of them to identify the learning process at work, first at the mental level of description, and then, albeit more cursorily, at the neural level. Then we shall propose a formal modeling of the hypothesised processes.

6.2.1 Kinds of learning

Relative to automatic inferences, learning can be separated into two broad categories, according to whether or not the content of the incoming information fits into the agent's current mental representations. If it does, then it appears to some extent as familiar to the agent, and yields a kind of learning which we shall call *repetition-induced learning*. This kind of learning essentially consists

in the reinforcement of memories through repetition of experience. If on the contrary the content of the incoming information does not fit into the agent's current mental representations, then it appears as new to it, which causes its surprise and triggers a kind of learning we shall call *novelty-induced learning*. We shall examine these two cases in turn.

Repetition-induced learning

Repetition-induced learning occurs when an agent's repeated experience of a same object or situation strengthens the vividness of the corresponding representation in its mind. For example, suppose that the agent knows of white and black swans, but that in its environment white swans outnumber black ones, so that it encounters white swans much more often than black ones. Presumably, when thinking about swans, it is white swans that should come to its mind in the first place, and so the agent should be disposed to infer 'white' from 'swan'. But now suppose that the number of black swans gradually increases over time, and that finally black swans become more numerous than white ones. One may expect the agent's representation of black swans to become more and more vivid, and finally to become more vivid than its representation of white swans. Accordingly, its disposition to infer 'white' from 'swan' should gradually fade away, and be eventually replaced by a disposition to infer 'black' from 'swan'.

This kind of learning mainly relies on the general fact that the reactivation (recalling) of a representation in an agent's mind increases it's vividness, while non-reactivated representations tend on the contrary to progressively decay. At the neural level, this most likely results from the previously described phenomena of long term potentiation (LTP) and long term depression (LTD) [2], which are believed to underlie a number of memory functions [3]. More precisely, one may imagine that the reactivation of a neural assembly might strengthen trough LTP the neural connections between the neurons from the assembly,

2 See pp. 23 and 25 above.

3 See for example [Izquierdo, 1993].

thus rendering the assembly as a whole more robust to neural noise and inhibition. It might also cause the neural assembly to expand, by allowing it aggregate through LTP and LTD an number of accidentally active neurons in a manner somewhat similar to that suggested on p. 25 above (which would again reinforce its robustness to noise and inhibition). Finally, reactivation might also strengthen the neural connections between the input neurons and those in the assembly, thus rendering the latter more responsive to the input. The loss of vividness of non-reactivated representations might for its part result from the loss by the corresponding assemblies of the neurons that are captured by reactivated assemblies, and/or from the weakening of the neural connections between the input neurons and those in the assembly.

It should be stressed, however, that the frequency with which an agent encounters a given object/situation is not the only criteria that determines the vividness of its representation in its mind. Another key factor is the agent's interest for this object/situation, that is, the emotional weight of the corresponding informational content in its mind. Indeed it is well known that contents with higher emotional value (whether positive or negative) are better memorized than less emotional ones, and less vulnerable to forgetting [4]. In mammals, this psychological observation is corroborated at the neural level by the well attested fact that the amygdala, which is known to be a brain center for emotions [5], plays a decisive role in memorization processes by modulating the action of LTP, notably in the medial temporal lobe [6]. It is likely that similar mechanisms occur in other taxonomic groups, as value assignation systems can be found throughout animal kingdom, and even in arthropods [7]. From the evolutionary standpoint, this makes sense if one considers that pre-

4 See for example [Cahill and McGaugh, 1995].

5 See pp. 21 and 23 above.

6 See for example [McGaugh, 2000] and [Phelps and LeDoux, 2005] p. 177.

7 See [Giurfa, 2007] pp. 811-812 for value-encoding neurons in the honeybee brain, and their critical role in learning.

diction errors can have drastically different costs for natural agents, and that the best adapted agent is not the one that makes the fewer prediction errors (*i.e.*, whose predictions best follow mathematical probabilities), but the one that best avoids errors with higher costs, even if this implies making more prediction errors overall. For example, consider a bird pecking seeds on the ground while a cat is resting nearby. Each time the cat moves, the bird needs to guess whether or not it will jump at it and try to catch it. For the bird, the cost of guessing incorrectly that the cat will jump is rather low, as it simply leads it to fly away needlessly. But the cost of guessing incorrectly that the cat will not jump is very high, as this could be fatal to the bird. Therefore, even if the cat finally jumps only one out of a hundred times, the bird has better chances of survival if it expects the cat to jump than if it expects it not to jump. Now, because the jumping cat situation is critical for the bird's survival, we may assume that it has a very high emotional value in its mind, and in fact a much more higher one than that of the resting cat, which is not so critical. As more emotional contents are better memorized, it is likely that even though the bird actually observes the cat resting much more often than jumping, its mental representation of the jumping cat situation will be more vivid in its mind than its representation of the resting cat situation, causing it to expect the cat to jump and improving its chances of survival. The modulation of memorization according to the value (or 'interest') assigned by the agent to experienced objects and situations thus appears to provide a significant adaptive advantage [8]. As value-assignation systems probably exist in all brained animals, this makes it likely that such a modulation of memorization actually occurs, at least to some extent, in most if not all of them.

8 Similar views can be found for example in [McGaugh, 2000] p. 248 and [Phelps and LeDoux, 2005] p. 177.

Novelty-induced learning

Novelty-induced learning comes in several kinds. A first one is when the agent learns the existence of a given class of objects/situations. For example, suppose that 'black', 'white' and 'swan' are features [9] the agent is physiologically able to perceive, and suppose that it knows of white swans, but ignores that there are also black ones. According to our analysis, this corresponds to the fact that it has a mental representation of white swans but no mental representation of black ones, or, put in another way, that one (at least) of its mental representations satisfies both the features 'white' and 'swan', but none satisfies both 'black' and 'swan'. Suppose that now the agent sees a black swan, that is, that its perceptual system detects the co-occurring features 'black' and 'swan', and sends the corresponding information to its inferential system. Obviously, this information does not fit into any of its mental representations, in the sense that none of them satisfies both these two features. This discrepancy between the agent's current observation and its worldview is properly what is called 'astonishment'. But soon the agent gets over its surprise, and admits that black swans do, in fact, exist. In other words, it revises its worldview by adding a 'black swan' element to the set of its mental representations, so as to make it consistent with the collected information. Astonishment works here as an error signal that triggers the revision process. The vividness of this new representation in the agent's mind will mainly depend on its interest for black swans: the greater this interest, the higher the vividness of the new representation. This interest in turn will presumably depend on the agent's previous interest for the (conjunctions of) features that compose the new representation (here, swans and black things). Yet we shall not investigate this issue any further, as this would bring us too far from our present concerns.

A second kind of novelty-induced learning is when the agent supplements

9 Again, for simplicity, we regard these as features, although 'swan' is certainly in fact the conjunction of many features (see footnote n°42 page 33).

its current knowledge of a given class of objects/situations with additional information. For instance, suppose that the agent knows of white swans, but never had the opportunity to observe their legs, and thus has no idea whatsoever about what colour they are. Consequently, its mental representation of white swans says nothing about the colour of their legs, which means that it neither satisfies nor falsifies any of the legs-colour features that the agent is physiologically able to perceive. Suppose that now the agent sees a white swan walking on the riverside, and observes that its legs are in fact black. Again, this does not fit into any of its current mental representations, since none of them satisfies at the same time the features 'white', 'swan' and 'black legs'. So again, this causes its astonishment, although perhaps not exactly of the same kind and not so strong as in the previous case. And again, this triggers the revision of its worldview according to the collected information. But this time, the revision consists in supplementing its knowledge about white swans by adding the feature 'black legs' to its previous mental representation of these. The vividness of the thus supplemented mental representation will not only depend on the agent's interest for the incoming information, but also on the vividness of its prior representation of white swans. Indeed, if the agent's interest for the observed content is low, then since the observation of a white swan implies the recognition of the considered object as being a white swan, and so the reactivation of the agent's mental representation of white swans, it seems that the vividness of the supplemented representation should be that of the agent's previous representation augmented by the gain of vividness resulting from its reactivation, just as in the repetition-induced learning case. But if the agent's interest for the observed content is strong enough, then it seems that it will affect the vividness the supplemented representation, just as in the first case of novelty-induced learning. Therefore, it seems that in this case the vividness of the supplemented representation will mainly depend on the strength of this interest.

A variant of this supplementation case is the merging of two or more representations into a single one. For example, suppose that the agent knows of blackbirds but does not know how they sing, and that on the other hand, it knows some particular bird song, but does not know which bird sings this way. It thus has two distinct representations, one for blackbirds and the other for this particular bird song. In addition, suppose that now the agent observes a blackbird singing this song. Again, this causes astonishment, since none of its mental representations satisfies at the same time the features that compose its representation of blackbirds and those that compose its representation of the song; and again, this launches a learning process. But this time, the agent merges its previous representations into one. This is a supplementation case in the sense that in doing so, the agent supplements at the same time its representation of blackbirds with the features that pertain to its representation of the song, and its representation of the song with the features that pertain to its representation of blackbirds. In this case we may assume that if the agent's interest for the incoming information is not strong enough to influence the vividness of the new representation, then the later will mainly depend on that of the most vivid of the prior representations, in other words that the vividness of the new representation will be equal to that of the most vivid of the prior representations, increased by the gain of vividness resulting from its reactivation. But if on the contrary the agent's interest for the incoming information is strong enough, then we may assume that the vividness of the new representation will depend as before on the strength of its interest for this information.

In all cases, the vividness of the agent's other mental representations (*i.e.*, those that have not been reactivated by the observation) should slightly decrease, because of the progressive forgetting that affects non-reactivated representations. It should however be noted that these should not be equally affected by forgetting, since those the content of which is of stronger interest for the agent will probably be more robust to forgetting than those the content of which is of lesser interest.

On the neural side, the supplementation of a given mental representation might simply result from the binding of features into conjunctions described on pages 23–24 above. More specifically, the neurons that support the to-be supplemented representation might simply come to bind the supplementary feature(s), in addition to those that they were already binding. As for the creation of a new representation, it requires the formation of a new assembly of concept-neurons. This might be achieved through the re-allocation of a number of neurons from other assemblies, that might be rapidly rewired through LTP and LTD. More precisely, one may imagine that some neurons supporting more general concepts that subsume the to-be-created representation might bind the currently observed features, just as in the supplementation case. Another possibility, that is not exclusive with the previous one, is that a number of neurons supporting similar enough concepts might be captured by the new assembly. Indeed, because the input they are wired to respond to is quite similar to the one that gives rise to the new representation, neural noise might bring a number of these neurons to get accidentally activated. The combined action of LTP and LTD might then change these neurons connections so as to make them respond to the present input, in a manner similar to that suggested on page 25. The difference between the present case and that considered on page 25 would lie in the agent's interest for the content represented by the input. If this interest is strong enough, then the new representation would be able to consolidate, otherwise it would rapidly dissolve as evoked on p. 25.

It should however be specified that not all the mental representations an agent may have need to have been acquired through learning. On the contrary, it is likely that natural agents are born with a number of innate representations, that can then either be used as they are or suitably supplemented. In species in which the youngs are born with immature brains such as humans, these might be very rudimentary and very few, and largely reshaped through subsequent learning. There might however be some, as for example the one that allows the new-born mammal to recognise a teat to feed from. But in

species were the youngs are born more autonomous, these might be more numerous and more elaborated, and form the substrate of what we call 'instinct' in usual language.

6.2.2 Formal modelling of learning

The above described learning processes all boil down to changes in the set $\mathcal{R}$ of the agent's most precise mental representations, and/or changes in the vividness $v(r)$ of these mental representations. Therefore, they can be represented in the framework by the corresponding modifications of the partial world model $\mathcal{M} = (\mathcal{U}, <)$ that represents the agent's worldview. In what follows we give a formal version of these learning processes, and we show how they induce a revision of the model $\mathcal{M}$.

Let $\mathcal{A}$, $\mathcal{F}^+$, $\mathcal{F}^-$, $\mathcal{F}$, σ, $\boldsymbol{L}$, ρ and rd be as defined in section 6.1, and let $T = \langle t_1, ..., t_n \rangle$ be a sequence of time instants. Given an object/situation $\mathfrak{s}$ and an instant $t_i \in T$ such that $\mathcal{A}$ considers $\mathfrak{s}$ at time t_i, the set $\{f \in \mathcal{F} / \ \mathcal{A}$ observes f at time $t_i\}$ is called $\mathcal{A}$'s *total observation (of $\mathfrak{s}$) at time t_i*. Note that such a total observation does not need to contain all the features that $\mathcal{A}$ could in principle observe given $\mathfrak{s}$, but only those that it actually notices at this precise moment [10].

Let $\mathcal{O}$ be the set of $\boldsymbol{L}$-formulas o such that:

i) o is a literal or a conjunction of literals,

ii) if f and f' are mutually exclusive features (in the sense specified on p. 31 above) and $\lambda_1 = \rho(f)$ occurs in o, then $\lambda_2 = \neg\rho(f')$ also occurs in o,

iii) $o \not\equiv \bot$.

$\mathcal{O}$ is the set of $\boldsymbol{L}$-formulas that represent the informational content of theoretically possible total observations by $\mathcal{A}$.

10 This notion freely inspired from Leitgeb's work [Leitgeb, 2004].

Given some situation $\mathfrak{s}$ that $\mathcal{A}$ considers at time t_i, we denote o_i the informational content of the corresponding total observation. o_i should be interpreted as the information that is sent by $\mathcal{A}$'s sensory areas to its inferential system at time t_i. We denote $\mathcal{F}_{o_i}$ the corresponding total observation, that is, $\mathcal{F}_{o_i} = \{f \in \mathcal{F} / \rho(f) \text{ occurs in } o_i\} = \{f \in \mathcal{F} / \mathcal{A} \text{ observes } f \text{ at time } t_i\}$. Note that we need not to assume that for each instant t_i in T there is a corresponding observation $\mathcal{F}_{o_i}$, since an agent needs not be always attentive to its own perceptions.

For any $o \in \mathcal{O}$, we assume that a measure of how much $\mathcal{A}$ would be interested if it were to observe o is available, and that it is given by a real number from $]\,0,\,1\,[$. For simplicity, we also assume that this measure does not change over time. Let $\mathcal{I} : \mathcal{O} \longrightarrow]\,0,\,1\,[$ be the function which to each $o \in \mathcal{O}$, associates the measure of $\mathcal{A}$'s interest for o.

For any $t_i \in T$, we shall denote $\mathcal{R}_i$ the set of $\mathcal{A}$'s most precise mental representations at time t_i, and v_i the function which to each $r \in \mathcal{R}_i$ associates the measure of its vividness in $\mathcal{A}$'s mind at this same time, as stated on p. 78 above.

Suppose that for some $i \in \{\,1,\,...,\,n\,\}$ we know $\mathcal{R}_i$ and v_i. Following the indications given in section 6.1, we may define:

- $\rho'_i : \mathcal{R}_i \longrightarrow \mathcal{W}_L^{\mathrm{P}}$, the function which to each $r \in \mathcal{R}_i$ associates the partial world w such that $\delta(\mathrm{w}) = \bigwedge \rho(f)/f \in r$,
- $\mathcal{U}_i = \{\mathrm{w} \in \mathcal{W}_L^{\mathrm{P}} / \exists\, r \in \mathcal{R}_i \text{ s.t. } \mathrm{w} = \rho'_i(r)\}$,
- $<_i$, the binary relation on $\mathcal{U}_i$ such that for any w and w' in $\mathcal{U}_i$, w $<_i$ w' iff $rd(v_i({\rho'_i}^{-1}(\mathrm{w}))) >_{\mathbb{R}} rd(v_i({\rho'_i}^{-1}(\mathrm{w'})))$,
- $\mathcal{M}_i = (\mathcal{U}_i, <_i)$.

$\mathcal{M}_i$ is a finite ranked precisification-free partial worlds model, and the model of the agent's worldview at time t_i, in the sense of the section 6.1.

Now suppose that at time t_i, $\mathcal{A}$ considers some situation $\mathfrak{s}$ and makes the total observation $\mathcal{F}_{o_i}$. According to the analysis provided in section 6.2.1, this may result in two possible learning cases, depending on whether or not there is some mental representation $r \in \mathcal{R}_i$ such that r satisfies the observed content o_i. We shall successively consider these two cases. For clarity's sake, we shall address them in the reverse order to that of section 6.2.1.

1. *Novelty-induced learning*

If for any $r \in \mathcal{R}_i$, $\mathcal{F}_{o_i} \not\subseteq r$, that is, if no mental representation in $\mathcal{A}$'s mind satisfies o_i, then the content of information o_i is new to it. This is rendered in the model $\mathcal{M}_i$ by the fact that no partial world in $\mathcal{U}_i$ satisfies o_i. On the syntactic side, we get that $o_i \,\|\!\!\sim_{\overline{\mathcal{U}_i, \mathbf{L}^i}} \bot$, which by *Supra-$\mathcal{U}$-Consequence* entails $o_i \,\|\!\!\sim_{\mathcal{M}_i} \bot$. This is to be interpreted as the fact that $\mathcal{A}$ is astonished, since it means that the content of its current observation conflicts with its knowledge about things, bringing it to the conclusion that 'something is wrong'. This discrepancy between $\mathcal{A}$'s current observation and its worldview triggers a learning process.

As explained in section 6.2.1, the latter basically consists in the addition of a new representation $r' = \mathcal{F}_{o_i}$ to $\mathcal{R}_i$, so as to form the set $\mathcal{R}_{i+1}$ of $\mathcal{A}$'s most precise mental representations at time t_{i+1}. But here again, we need to consider two cases:

- If for any r in $\mathcal{R}_i$, $r \not\subseteq r'$, that is, if the to-be-added new representation does not supplement any of $\mathcal{A}$'s already existent mental representations, then we are in the first case of novelty-induced learning described above (p. 84), and r' is simply added to $\mathcal{R}_i$. That is, $\mathcal{R}_{i+1} = \mathcal{R}_i \cup \{r'\}$.
- If on the contrary there are some $r_1, \ldots, r_k \in \mathcal{R}_i$ such that for all j $(1 \leq_{\mathbb{N}} j \leq_{\mathbb{N}} k)$, $r_j \subseteq r'$, that is, if the to-be-added new representation r' supplements some of $\mathcal{A}$'s pre-existing mental representations, then we are in the supplementation case of novelty-induced learning (see p. 85). In this case, the new representation r' will replace in

$\mathcal{R}_{i+1}$ all the mental representations it supplements, that is, all the $r \in \mathcal{R}_i$ such that $r \subseteq r'$.

Putting these two cases together, we get that $\mathcal{R}_{i+1} = (\mathcal{R}_i - \{r \in \mathcal{R}_i \ / \ r \subseteq r'\}) \cup \{r'\}$.

We now turn to the issue of the vividness of the elements of $\mathcal{R}_{i+1}$ in $\mathcal{A}$'s mind at time t_{i+1}. According to the analysis given in section 6.2.1, the vividness of the new representation r' at time t_{i+1} should depend on $\mathcal{I}(o_i)$, and also, if there are some, on the vividness of the representations from $\mathcal{R}_i$ that r' supplements. Let $\{r_1, \ ... \ , r_k\}$ be the (possibly empty) set of such representations, that is, $\{r_1, \ ... \ , r_k\} = \{r \in \mathcal{R}_i \ / \ r \subseteq r'\}$. In addition, let $x = max(\{0, \ v_i(r_1), \ ... \ , \ v_i(r_k)\})$, and let $\epsilon \in \ ]\,0, \ 1\,[$ be the agent's learning rate[11]. We shall assume that $v_{i+1}(r') = max(\{\mathcal{I}(o_i), \ x+\epsilon.\mathcal{I}(o_i)(1-x)\})$. This choice can be motivated as follows:

- If $\{r_1, \ ... \ , r_k\} = \varnothing$, then $x = 0$, so $v_{i+1}(r') = max(\{\mathcal{I}(o_i), \ \epsilon.\mathcal{I}(o_i)\})$, thus, since $\epsilon <_{\mathbb{R}} 1$, $v_{i+1}(r') = \mathcal{I}(o_i)$. In other words, in case the to-be-added new representation r' does not supplement any of $\mathcal{A}$'s pre-existing mental representations (which is the first case of novelty induced learning considered in section 6.2.1), then we assume that the vividness of r' will simply depend on $\mathcal{A}$'s interest for o_i.
- If $\{r_1, \ ... \ , \ r_k\} \neq \varnothing$, then we get that $v_{i+1}(r') = \mathcal{I}(o_i)$ if $\mathcal{I}(o_i) \geq_{\mathbb{R}} x+\epsilon.\mathcal{I}(o_i)(1-x)$, and $v_{i+1}(r') = x+\epsilon.\mathcal{I}(o_i)(1-x)$ otherwise. This means that in the supplementation case of novelty induced learning, we assume that if $\mathcal{A}$'s interest for o_i is 'high enough' com-

11 The notion of learning rate comes from neuromodelling and artificial neural networks, where a learning rate is a very small real number that controls the updating function by which the weights of the connections between a network's neurons are revised over time. In this context, a learning rate is taken to account for the rate at which an agent's neural connections increase their strength as a result of some neural mechanism for synaptic plasticity such as, *e.g.*, the previously mentioned long-term potentiation mechanism. More generally, an agent's learning rate can be seen as a measure of the lability of its knowledge: the higher its learning rate, the more sensitive the agent to incoming information. It is in this broader sense that we use the term here.

pared to the vividness of the most vivid of the to-be-supplemented representations, then the vividness of r' will fully depend on the strength of this interest, but that otherwise it will largely depend on the vividness of the most vivid of the to-be-supplemented representations. More precisely, we assume that in this latter case the vividness of r' will be that of the most vivid of the to-be-supplemented representations, increased by some gain of vividness $\epsilon.\mathcal{I}(o_i)(1-x)$ resulting from its reactivation. Furthermore, we assume that this gain will depend on the one hand on the agent's learning rate ϵ, and on the other hand on its interest $\mathcal{I}(o_i)$ for the informational content to be learned. The $(1-x)$ component for its part accounts for the fact the vividness of a representation cannot increase indefinitely, by making $\epsilon.\mathcal{I}(o_i)(1-x)$ progressively decrease as x grows. Since $\epsilon.\mathcal{I}(o_i) <_{\mathbb{R}} 1$, we will always have $\epsilon.\mathcal{I}(o_i)(1-x) <_{\mathbb{R}} (1-x)$. Thus $v_{i+1}(r')$ will tend to 1 as x grows, and $v_{i+1}(r')$ will always be in $]\,0,\,1\,[$.

As for the other mental representations in $\mathcal{R}_{i+1}$, we may assume that they have not been reactivated by the observation $\mathcal{F}_{o_i}$, since they do not satisfy o_i. Let $\epsilon' \in]\,0,\,1\,[$ be the agent's 'forgetting rate', that is, a measure of the general tendency of its mental representations to lose vividness when they are not reactivated. Following the analysis provided in section 6.2.1, we shall assume for any $r \in \mathcal{R}_{i+1} - \{r'\}$, $v_{i+1}(r) = v_i(r) - \epsilon'.v_i(r)(1-\mathcal{I}(o_i))$. The $(1-\mathcal{I}(o_i))$ component causes $\epsilon'.v_i(r)(1-\mathcal{I}(o_i))$ to decrease as $\mathcal{I}(o_i)$ grows, accounting for the greater robustness to forgetting of mental representations in the content of which the agent is more interested. The $v_i(r)$ component makes that $\epsilon'.v_i(r)(1-\mathcal{I}(o_i))$ decreases as $v_i(r)$ does, and so that $v_{i+1}(r)$ tends to 0 as $v_i(r)$ decreases, ensuring that $v_{i+1}(r) \in]\,0,\,1\,[$.

2. *Repetition-induced learning*

If there is some $r \in \mathcal{R}_i$ such that $\mathcal{F}_{o_i} \subseteq r$, then the agent knows at least

one object/situation that, in its view, satisfies o_i. In the model $\mathcal{M}_i$, this is rendered by the fact that there is at least one partial world w in $\mathcal{U}_i$ such that w $\Vdash o_i$. So $o_i \not\Vdash_{\overline{\mathcal{U}_i, \boldsymbol{L}^{\Vdash}}} \bot$, and by ($\boldsymbol{L}^{\Vdash}$) $\mathcal{U}$-*Consistence*, $o_i \not\Vdash_{\mathcal{M}_i} \bot$, which corresponds to the fact that o_i causes no astonishment to $\mathcal{A}$. Therefore, the inferential process can run it course, and the agent draws the automatic inferences it is disposed to draw.

Moreover, since there is at least one r in $\mathcal{R}_i$ that satisfies o_i, the set of objects and situations $\mathcal{A}$ knows about is unchanged by the observation $\mathcal{F}_{o_i}$, which means that $\mathcal{R}_{i+1} = \mathcal{R}_i$.

According to the analysis given in section 1.1.2, the inferential process triggered by $\mathcal{F}_{o_i}$ comes down to the fact that only the most vivid of $\mathcal{A}$'s mental representations from $\mathcal{R}_i$ that satisfy o_i are reactivated by $\mathcal{F}_{o_i}$. Yet here we need to consider not the exact vividness values but rather the rounded ones, because, as previously mentioned, neural noise blurs small differences in vividness. Let $y = max(\{rd(v_i(r))/r \in \mathcal{R}_i$ and $\mathcal{F}_{o_i} \subseteq r\})$, and $\mathcal{Y} = \{r \in \mathcal{R}_i \ / \ \mathcal{F}_{o_i} \subseteq r$ and $rd(v_i(r)) = y\}$. $\mathcal{Y}$ is the set of $\mathcal{A}$'s most precise mental representations that, according to our analysis, are reactivated at time t_i by the observation $\mathcal{F}_{o_i}$. Taking up the previous estimates of the gain or loss of vividness of representations following their reactivation or their non-reactivation, we shall assume that:

- For any $r \in \mathcal{Y}$, $v_{i+1}(r) = v_i(r) + \epsilon.\mathcal{I}(o_i)(1 - v_i(r))$, and
- For any $r \in \mathcal{R}_{i+1} - \mathcal{Y}$, $v_{i+1}(r) = v_i(r) - \epsilon'.v_i(r)(1 - \mathcal{I}(o_i))$.

At this stage we know $\mathcal{R}_{i+1}$ both in the case where there is some $r \in \mathcal{R}_i$ such that $\mathcal{F}_{o_i} \subseteq r$, and in the case where there is none, and for each of these cases we have an estimate of v_{i+1}. Therefore, we are able to define ρ'_{i+1}, $\mathcal{U}_{i+1}$ and $<_{i+1}$ as shown on page 89. $\mathcal{M}_{i+1} = (\mathcal{U}_{i+1}, <_{i+1})$ is a finite and precisification-free ranked partial worlds model, and the model of the agent's worldview at time t_{i+1}. The inference relation induced by $\mathcal{M}_{i+1}$ represents the agent's general knowledge at time t_{i+1}. Since in the above proceeding, a finite precisification-free ranked partial worlds $\mathcal{M}_i$ is always revised into another,

uniquely defined, finite precisification-free ranked partial worlds model $\mathcal{M}_{i+1}$, the revision process thus defined is indefinitely iterable.

However, once the intended interpretation of the logical framework is clarified, one may wish to get rid as much as possible of the non-logical elements of the above modelling. This can be done in the following manner. Suppose that for some $i \in \{1, ..., n\}$ we know $\mathcal{R}_i$ and v_i, and that we have built $\mathcal{M}_i = (\mathcal{U}_i, <_i)$ as suggested on page 89. We define a function $v*_i : \mathcal{U}_i \longrightarrow]0, 1[$ by: for any w $\in \mathcal{U}_i$, $v*_i(\text{w}) = v_i({\rho'}_i^{-1}(\text{w}))$. Now let o_i be content of $\mathcal{A}$'s total observation at time t_i. The above two cases of learning boil down to the following:

1. If $o_i \not\|\!\!\sim_{\overline{\mathcal{U}_i, \boldsymbol{L}}} \bot$:

Let w' $\in \mathcal{W}_{\boldsymbol{L}}^{\text{p}}$ be the partial world such that $o_i = \delta(\text{w'})$, and

$x^* = max(\{0\} \cup \{v*_i(\text{w}) \,/ o_i \vdash \delta(\text{w})\})$. We define:

. $\mathcal{U}_{i+1} = (\mathcal{U}_i - \{\text{w} \in \mathcal{U}_i \,/ o_i \vdash \delta(\text{w})\}) \cup \{\text{w'}\}$,

. $v*_{i+1}(\text{w'}) = max(\{\mathcal{I}(o_i),\, x^* + \epsilon.\mathcal{I}(o_i)(1 - x^*)\})$,

. For any w $\in (\mathcal{U}_{i+1} - \{\text{w'}\})$, $v*_{i+1}(\text{w}) = v*_i(\text{w}) - \epsilon'.v*_i(\text{w})(1 - \mathcal{I}(o_i))$.

2. If $o_i \not|\!\!\sim_{\overline{\mathcal{U}_i, \boldsymbol{L}}} \bot$, then $\{\text{w} \in \mathcal{U}_i \,/\, \text{w} \Vdash o_i\} \neq \varnothing$.

Let $y^* = max(\{rd(v*_i(\text{w})) \,/\, \text{w} \in \mathcal{U}_i \text{ and w} \Vdash o_i\})$, and

$\mathcal{Y}^* = \{\text{w} \in \mathcal{U}_i \,/\, \text{w} \Vdash o_i \text{ and } rd(v*_i(\text{w})) = y^*\}$. We define:

. $\mathcal{U}_{i+1} = \mathcal{U}_i$,

. For any w $\in \mathcal{Y}^*$, $v*_{i+1}(\text{w}) = v*_i(\text{w}) + \epsilon.\mathcal{I}(o_i)(1 - v*_i(\text{w}))$,

. For any w $\in (\mathcal{U}_{i+1} - \mathcal{Y}^*)$, $v*_{i+1}(\text{w}) = v*_i(\text{w}) - \epsilon'.v*_i(\text{w})(1 - \mathcal{I}(o_i))$.

In both cases, we define $<_{i+1}$ by: for any w and w' $\in \mathcal{U}_{i+1}$, w $<_{i+1}$ w' iff $rd(v*_{i+1}(\text{w})) >_{\mathbb{R}} rd(v*_{i+1}(\text{w'}))$. $\mathcal{M}_{i+1} = (\mathcal{U}_{i+1}, <_{i+1})$ is a finite and precisification-free ranked partial worlds model, and the model of the agent's worldview at time t_{i+1}.

Conclusion and perspectives

In this work, we took inspiration in cognitive sciences to address the issue of the logical modelling of reasoning and learning, and we developed a bio-inspired model of a very simple kind of inferences we dubbed *automatic inferences*. This model relies on a number of hypotheses about the functioning of brains, and for this reason it cannot be claimed to be true, but only to be plausible enough. One may expect that further knowledge about brain organization and processes will necessitate a number of changes in the suggested model. However the main hypotheses it relies on appear to be well-established enough to give us reasons to believe that these adjustments will only be of details.

For example, a more precise understanding of the neural factors that underlie the phenomenon of vividness of mental representations in an agent's mind might call for some changes in the formulas that define the function that updates the measure of vividness of representations in the modelling of learning suggested in section 6.2.2. Or, a better understanding of the neural mechanisms of feature-binding and of their role in the formation of mental representations might render it desirable to drop the simplifying assumption we made that all the features that take part in a given representation are equally important, and therefore that mental representations can be seen as plain conjunctions of features. If this should happen, then it would not be possible to figure mental representations by partial worlds any more, and we would have to replace them in the model by more complex semantical structures. This would obviously change the logical apparatus in a drastic way, but not the general idea behind the model nor its global dynamics.

Furthermore, it should be borne in mind that the here proposed modelling of automatic inferences and of the associated learning is but a first stone for a more general project of modelling of reasoning and learning. As stressed at the beginning of this dissertation, reasoning in natural agents incorporates a wide range of cognitive abilities, of which the ability to draw automatic inferences is presumably one of the most elementary. A natural extension of the present work would thus be to articulate a modelling of some of these other abilities to that of automatic inferences. Among those that might be fruitfully investigated, the first one that comes to mind is the ability to use general concepts in reasoning. For example, an agent that observes a black swan for the first time while having some previous knowledge of other bird species may be able to expect it to be able to fly, despite the fact that it has no previous experience of black swans (this of course can only occur once the agent's surprise is over, that is, once it has admitted that black swans do exist). As far as we can tell at this stage of our thinking, it seems that to do so the agent first searches for the most precise of its general concepts that are consistent with the incoming information, and then retrieves their common informational content to use it as a premise for automatic inferences. The ability to climb back to more general concepts in reasoning is unevenly distributed among species, and most probably depends on the species' ability to have a ample set of general concepts.

Another interesting case to study would be that of reasoning processes that involve ordered strings of mental representations instead of single representations. Acknowledging Eichenbaum's hypothesis that episodic memories are encoded by the hippocampus as ordered strings of discrete mental representations [1], one may imagine an abstraction process by which the common features of similar enough strings could be abstracted, in a way that would be somewhat akin to that by which the content of similar single mental representations

1 [Eichenbaum, 2004]. See also p. 27 above.

is abstracted into more general concepts. Yet mathematically speaking, a set of ordered strings is nothing but the graph of a relation, hence such an ability to represent the common features shared by the elements of a set of strings might come down to the ability to represent relations. This of course is a mere hypothesis and would need to be further checked, but the fact that the ability to represent relations seems to exist only in species with relatively high cognitive abilities, just as the ability to have rich and detailed episodic memories, supports its plausibility.

Finally, it might be interesting to investigate the consequences of self-awareness in natural agents. A rather common view is that self-awareness lies in the ability to collect information about one's own neural operations, and to process this information exactly in the same way as sensory information from the outer world. If this view is correct, then a self-aware agent might be able to form mental representations of (some of) its own neural operations, and then to abstract the common features of these so as to form representations of classes of operations. Among the classes of operations that might be abstracted and represented in this manner are the conjunction, the negation and the ||-disjunction of informational contents. Indeed, operations of these kinds are continuously performed by cognitive agents [2], hence we may expect the corresponding classes to be abstracted in the first place. Note that since from a mathematical point of view logical connectives are functions and functions are relations, the ability to represent relations should be first required. On the other hand, such an explicit representation of logical connectives — which is to be contrasted with the implicit realisation of connectives that prevails in other conditions [3] — might in turn allow for the emergence of syntactically structured verbal representation, which is the hallmark of superior language abilities and reasoning.

2 Neural operations plausibly supporting conjunctions have been briefly described on p. 23 above. To get an insight into what the neural operations corresponding to negations and ||-disjunctions might be like, see p. 31 (negation), and p. 57 (||-disjunctions).

3 See p. 36 above.

Therefore, starting with the very basic ability to form mental representations and to use them as a support for inferences and decision making, we might by progressive adjunction of additional abilities be able to reconstruct significant parts of human thinking. In such a modelling, reasoning and learning would be naturally articulated, and iteration of learning would not yield difficulties any more.

Appendices

Appendix A

For any w $\in \mathcal{U}$ and any $\boldsymbol{L}\Vdash$-formula $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \boldsymbol{\alpha}$ iff w $\Vdash\models \boldsymbol{\alpha}$.

Proof:

$\Longrightarrow$) Let w $\in \mathcal{U}$ and let α be a $\boldsymbol{L}\Vdash$-formula such that δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \alpha$.
w $\Vdash\models \delta$(w), so by the definition of $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash}$, w $\Vdash\models \alpha$.

$\Longleftarrow$) Let w $\in \mathcal{U}$ and let α be a $\boldsymbol{L}\Vdash$-formula such that w $\Vdash\models \alpha$. We proceed by induction on the construction of $\boldsymbol{L}\Vdash$-formulas:

i) (Base case) If α is a $\boldsymbol{L}$-formula, then δ(w)$\vdash \alpha$ (see section 2.1). Thus by supra-classicality of $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash}$, δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \alpha$.

ii) (Induction step) If α is not a $\boldsymbol{L}$-formula, then, either $\alpha = \beta \parallel \gamma$ or $\alpha = \beta \curlywedge \gamma$. Suppose that we have already proven that for any w' $\in \mathcal{U}$, w' $\Vdash\models \beta$ entails δ(w') $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \beta$ and w' $\Vdash\models \gamma$ entails δ(w') $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \gamma$.

1. If $\alpha = \beta \parallel \gamma$, then by hypothesis
w $\Vdash\models \beta \parallel \gamma$, iff (by def. of $\parallel$)
w $\Vdash\models \beta$ or w $\Vdash\models \gamma$, thus (by induction hypothesis)
δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \beta$ or δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \gamma$.

- If δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \beta$, then by the def. of $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash}$,
  δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \beta \parallel \gamma$, *i.e.* δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \alpha$.
- Otherwise, δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \gamma$, so by the def. of $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash}$,
  δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \beta \parallel \gamma$, *i.e.* δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}}\Vdash} \alpha$.

2. If $\alpha = \beta \curlywedge \gamma$, then by hypothesis
w $\Vdash\models \beta \curlywedge \gamma$, iff (by def. of $\curlywedge$)

w $\|\!\models \beta$ and w $\|\!\models \gamma$, thus (by induction hypothesis)
δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \beta$ and δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \gamma$, iff
δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \beta \curlywedge \gamma$, *i.e.* δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \alpha$.

□

Appendix B

***Injectivity*:**

For any formula α maximal-consistent in $\mathcal{U}$,

if $\alpha \parallel \beta \parallel \gamma \,\|\!\sim_{\mathcal{M}} \beta \parallel \gamma$ and $\alpha \parallel \beta \;\; \|\!\not\sim_{\mathcal{M}} \beta$, then $\alpha \parallel \gamma \,\|\!\sim_{\mathcal{M}} \gamma$.

Proof:

Let α, β and γ be $\boldsymbol{L}^{\|}$-formulas such that α is maximal-consistent in $\mathcal{U}$, $\alpha \parallel \beta \parallel \gamma \,\|\!\sim_{\mathcal{M}} \beta \parallel \gamma$ and $\alpha \parallel \beta \,\|\!\not\sim_{\mathcal{M}} \beta$, and let w $\in \mathcal{U}$ be $<$-minimal for $\alpha \parallel \gamma$ (if there is no such w, then, trivially, $\alpha \parallel \gamma \,\|\!\sim_{\mathcal{M}} \gamma$). Suppose that w $\|\!\not\models \gamma$:

i) Then w $\|\!\models \alpha$.

ii) By hypothesis α is maximal-consistent, so α is a $\boldsymbol{L}$-formula. Since w $\|\!\models \alpha$, δ(w) $\vdash \alpha$.

iii) w $\|\!\models \alpha \wedge \delta$(w), so $\alpha \wedge \delta$(w) $\|\!\not\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$. By def. of maximal-consistent formulas, $\alpha \vdash \delta$(w). So by ii), $\alpha \equiv \delta$(w).

iv) By hypothesis $\alpha \parallel \beta \,\|\!\not\sim_{\mathcal{M}} \beta$, iff
$\exists$ w' $\in \mathcal{U}$ s.t. w' is $<$-minimal for $\alpha \parallel \beta$ and w' $\|\!\not\models \beta$. Thus w' $\|\!\models \alpha$, iff δ(w') $\vdash \alpha$. Since α is maximal-consistent, $\alpha \vdash \delta$(w'), so $\alpha \equiv \delta$(w').

v) By iii) and iv), w = w', thus by iv), w is $<$-minimal for $\alpha \parallel \beta$ and w $\|\!\not\models \beta$.

vi) Since w $\|\!\models \alpha$, w $\|\!\models \alpha \parallel \beta \parallel \gamma$. We show that w is $<$-minimal for $\alpha \parallel \beta \parallel \gamma$:
Suppose the contrary, then there is a w" $<$ w such that w" $\|\!\models \alpha \parallel \beta \parallel \gamma$. By $\boldsymbol{L}^{\|}$-smoothness, w $\not<$ w, so w" $\neq$ w, thus since α is maximal-consistent, w" $\|\!\not\models \alpha$. So w" $\|\!\models \beta \parallel \gamma$. If w" $\|\!\models \beta$, then w" $\|\!\models \alpha \parallel \beta$, so w is not $<$-minimal for $\alpha \parallel \beta$, which contradicts v). If w" $\|\!\models \gamma$, then w" $\|\!\models \alpha \parallel \gamma$, so w is not $<$-minimal for $\alpha \parallel \gamma$, which contradicts the hypothesis. Thus w is $<$-minimal for $\alpha \parallel \beta \parallel \gamma$.

vii) By hypothesis $\alpha \parallel \beta \parallel \gamma \mid\!\!\sim_{\mathcal{M}} \beta \parallel \gamma$, so by vi), w $\Vdash \beta \parallel \gamma$.
By v) w $\nVdash \beta$, so w $\Vdash \gamma$, which contradicts the hypothesis.

□

Appendix C

***‖-Disjunct Equivalence*:**

If $\alpha_1 \mid\!\!\sim_{\mathcal{M}} \alpha_2$, $\alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_1$ and $\alpha_1 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \gamma$, then $\alpha_2 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \gamma$.

Proof:

Let α_1, α_2, β and γ be $\boldsymbol{L}^{\parallel}$-formulas such that $\alpha_1 \mid\!\!\sim_{\mathcal{M}} \alpha_2$, $\alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_1$ and $\alpha_1 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \gamma$.

i) By hypothesis, $\alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_1$ so by *$\mathcal{U}$-Right Weakening*, $\alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.

ii) By *Reflexivity*, $\alpha_1 \mid\!\!\sim_{\mathcal{M}} \alpha_1$, so by *$\mathcal{U}$-Right Weakening*, $\alpha_1 \mid\!\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.
By *‖-Or* on i), $\alpha_1 \parallel \alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.

iii) By *Reflexivity*, $\alpha_1 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$, so by *‖-Or* on ii), $(\alpha_1 \parallel \alpha_2) \| (\alpha_1 \parallel \beta) \mid\!\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.
Since $(\alpha_1 \parallel \alpha_2) \| (\alpha_1 \parallel \beta) \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \alpha_1 \parallel \alpha_2 \parallel \beta$, by *$\mathcal{U}$-Left Equivalence*
$\alpha_1 \parallel \alpha_2 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.

iv) By *Reflexivity*, $\alpha_1 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$, so by *$\mathcal{U}$-Right Weakening*,
$\alpha_1 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \alpha_1 \parallel \alpha_2 \parallel \beta$.

v) By hypothesis $\alpha_1 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \gamma$, so by the derived rule *Equivalence*, on iii) and iv),
$\alpha_1 \parallel \alpha_2 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \gamma$.

vi) By hypothesis, $\alpha_1 \mid\!\!\sim_{\mathcal{M}} \alpha_2$ and by *Reflexivity* $\alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_2$, so by *‖-Or* $\alpha_1 \parallel \alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_2$.
So by *$\mathcal{U}$-Right Weakening*, $\alpha_1 \parallel \alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_2 \parallel \beta$.

vii) By *Reflexivity* $\beta \mid\!\!\sim_{\mathcal{M}} \beta$, so by *$\mathcal{U}$-Right Weakening*, $\beta \mid\!\!\sim_{\mathcal{M}} \alpha_2 \parallel \beta$.
So by *‖-Or* on vi), $\alpha_1 \parallel \alpha_2 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \alpha_2 \parallel \beta$.

viii) By *Cautious Monotony* on v) and vii), $(\alpha_1 \parallel \alpha_2 \parallel \beta) \wedge (\alpha_2 \parallel \beta) \mid\!\!\sim_{\mathcal{M}} \gamma$.
Since $(\alpha_1 \parallel \alpha_2 \parallel \beta) \wedge (\alpha_2 \parallel \beta) \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \alpha_2 \parallel \beta$, by *$\mathcal{U}$-Left Equivalence*,
$\alpha_2 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \gamma$.

□

Appendix D

$\|$-*Transitivity*:

If $\alpha_1 \mid\!\sim_{\mathcal{M}} \alpha_2$, ... , $\alpha_{n-1} \mid\!\sim_{\mathcal{M}} \alpha_n$, then $\alpha_1 \parallel \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_n$.

Proof:

Let α_1, α_2, ... , α_n be such that $\alpha_1 \mid\!\sim_{\mathcal{M}} \alpha_2$, ... , $\alpha_{n-1} \mid\!\sim_{\mathcal{M}} \alpha_n$.

i) By hypothesis, for any $i(1 \leq_{\mathbb{N}} i <_{\mathbb{N}} n)$, $\alpha_i \mid\!\sim_{\mathcal{M}} \alpha_{i+1}$, and, by *Reflexivity* $\alpha_{i+1} \mid\!\sim_{\mathcal{M}} \alpha_{i+1}$.
Thus by $\|$-*Or*, for any $i(1 \leq_{\mathbb{N}} i <_{\mathbb{N}} n)$, $\alpha_i \parallel \alpha_{i+1} \mid\!\sim_{\mathcal{M}} \alpha_{i+1}$.

ii) By *Reflexivity*, for any $i(1 \leq_{\mathbb{N}} i <_{\mathbb{N}} n)$, $\alpha_{i+1} \mid\!\sim_{\mathcal{M}} \alpha_{i+1}$, so by $\mathcal{U}$-*Right Weakening*, for any $i(1 \leq_{\mathbb{N}} i <_{\mathbb{N}} n)$, $\alpha_{i+1} \mid\!\sim_{\mathcal{M}} \alpha_i \parallel \alpha_{i+1}$.

iii) By i), $\alpha_{n-1} \parallel \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_n$, so by successive applications of the derived rule $\|$-*Disjunct Equivalence* on i) and ii), one gets $\alpha_1 \parallel \alpha_2 \parallel \ldots \parallel \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_n$.

iv) By $\mathcal{U}$-*Right Weakening* on iii), $\alpha_1 \parallel \alpha_2 \parallel \ldots \parallel \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \alpha_n$.

v) By *Cautious Monotony* on iii) and iv), $(\alpha_1 \parallel \alpha_2 \parallel \ldots \parallel \alpha_n) \wedge (\alpha_1 \parallel \alpha_n) \mid\!\sim_{\mathcal{M}} \alpha_n$.
$(\alpha_1 \parallel \alpha_2 \parallel \ldots \parallel \alpha_n) \wedge (\alpha_1 \parallel \alpha_n) \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \alpha_1 \parallel \alpha_n$, so by $\mathcal{U}$-*Left Equivalence*, $\alpha_1 \parallel \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_n$.

□

Appendix E

***Rankedness*:**

For any formulas α_1, α_2 and α_3 maximal-consistent in $\mathcal{U}$ (all $\not\equiv$), if $\alpha_1 \parallel \alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_2$ and $\alpha_1 \parallel \alpha_3 \not\mid\!\sim_{\mathcal{M}} \alpha_3$, then $\beta \parallel \gamma \mid\!\sim_{\mathcal{M}} \alpha_2$.

Proof:

Let $\mathcal{M} = (\mathcal{U}, <)$ be a finite ranked partial worlds model, and let α_1, α_2 and α_3 be (all $\not\equiv$) $\boldsymbol{L}^{\parallel}$-formulas maximal-consistent in $\mathcal{U}$ and such that $\alpha_1 \parallel \alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_2$ and $\alpha_1 \parallel \alpha_3 \not\mid\!\sim_{\mathcal{M}} \alpha_3$. In addition, let w $\in \mathcal{U}$ s.t. w is $<$-minimal for $\beta \parallel \gamma$ (if there is no such w in $\mathcal{U}$, then, trivially, $\beta \parallel \gamma \mid\!\sim_{\mathcal{M}} \alpha_2$). Suppose that w $\not\Vdash \alpha_2$:

i) Then w $\Vdash \alpha_3$.
Thus w $\Vdash \alpha_1 \parallel \alpha_3$.

ii) By hypothesis $\alpha_1 \parallel \alpha_3 \ \ \parallel\!\!\not\sim_{\mathcal{M}} \alpha_3$, iff
there is a w' in $\mathcal{U}$ such that w' is $<$-minimal for $\alpha_1 \parallel \alpha_3$ and w' $\parallel\!\not\models \alpha_3$.
Since $<$ is modular, by i), w' $\leq$ w.

iii) Since w' $\parallel\!\models \alpha_1 \parallel \alpha_3$ and w' $\parallel\!\not\models \alpha_3$, w' $\parallel\!\models \alpha_1$. Thus w' $\parallel\!\models \alpha_1 \parallel \alpha_2$.

iv) Since w' $\parallel\!\models \alpha_1$ and α_1 is maximal-consistent, $\alpha_1 \equiv \delta$(w').

v) We show that w' $\parallel\!\not\models \alpha_2$: suppose the contrary, then since α_2 is maximal-consistent, we get $\alpha_2 \equiv \delta$(w'), thus by iv) $\alpha_1 \equiv \alpha_2$, which contradicts the hypothesis.

vi) By hypothesis $\alpha_1 \parallel \alpha_2 \parallel\!\sim_{\mathcal{M}} \alpha_2$, so by iii) and v), w' is not $<$-minimal for $\alpha_1 \parallel \alpha_2$. Since $\mathcal{M}$ is finite, $<$ is $\boldsymbol{L}^{\parallel}$-smooth, thus there is w" $<$ w' such that w" is $<$-minimal for $\alpha_1 \parallel \alpha_2$. w" $\parallel\!\models \alpha_2$, so w" $\parallel\!\models \alpha_2 \parallel \alpha_3$.

vii) By vi) w" $<$ w', and by ii), w' $\leq$ w, so by modularity of $<$, w" $<$ w.

viii) By vi), w" $\parallel\!\models \alpha_2 \parallel \alpha_3$, so by vii) w is not $<$-minimal for $\alpha_2 \parallel \alpha_3$, which contradicts the hypothesis.

□

Index of symbols

The main symbols used in this work are listed below. References are given to the pages where a symbol is introduced or re-introduced, or where the corresponding notion is explained.

Bibliography

Alchourrón, C., Gardenfors, P., and Makinson, D. (1985). On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530.

Aminoff, E. M., Kveraga, K., and Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, 17(8):379–390.

Barnett, P. D., Nordström, K., and O'Carroll, D. C. (2007). Retinotopic organization of small-field-target-detecting neurons in the insect visual system. *Current Biology*, 17(7):569–578.

Bartos, M. (2008). Hunting prey with different escape potentials — alternative predatory tactics in a dune dwelling salticid. *The Journal of Arachnology*, 35:499–508.

Bayley, P. J., Hopkins, R. O., and Squire, L. R. (2006). The fate of old memories after medial temporal lobe damage. *The Journal of Neuroscience*, 26(51):13311–13317.

Bayley, P. J. and Squire, L. R. (2005). Failure to acquire new semantic knowledge in patients with large medial temporal lobe lesions. *Hippocampus*, 15(2):273–280.

Bear, M. F. and Abraham, W. C. (1996). Long-term depression in hippocampus. *Annual Review of Neuroscience*, 19:437–462.

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology*, 61:27–48.

Cahill, L. and McGaugh, J. L. (1995). A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and Cognition*, 4(4):410–421.

Chittka, L. and Niven, J. (2009). Are bigger brains better? *Current Biology*, 19:995–1008.

Dubois, D. (2008). On ignorance and contradiction considered as truth-values. *Logic Journal of the IGPL*, 16(2).

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience*, 1(1):41–50.

Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron*, 44(1):109–120.

Freund, M. (2004). On the revision of preferences and rational inference processes. *Artificial Intelligence*, 152(1):105–137.

Gabbay, D. M. and Schlechta, K. (2008). Cumulativity without closure of the domain under finite unions. *The Review of Symbolic Logic*, 1(3):372–392.

Gabbay, D. M. and Schlechta, K. (2009). Roadmap for preferential logics. *Journal of Applied Non-classical Logics*, 19(1):43–95.

Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., and Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, 322(5898):96–101.

Giurfa, M. (2007). Behavioral and neural analysis of associative learning in the honeybee: a taste from the magic well. *Journal of Comparative Physiology A*, 193(8):801–824.

Holdstock, J. S., Hocking, J., Notley, P., Devlin, J. T., and Price, C. J. (2009). Integrating visual and tactile information in the perirhinal cortex. *Cerebral Cortex*, 19(12):2993–3000.

Hollis, K. L. and Guillette, L. M. (2011). Associative learning in insects: Evolutionary models, mushroom bodies, and a neuroscientific conundrum. *Comparative Cognition & Behavior Reviews*, 6:24–45.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148(3):574–591.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195(1):215–243.

Izquierdo, I. (1993). Long-term potentiation and the mechanisms of memory. *Drug Development Research*, 30(1):1–17.

Jackson, R. R., Carter, C. M., and Tarsitano, M. S. (2001). Trial-and-error solving of a confinement problem by a jumping spider, *Portia fimbriata*. *Behaviour*, 138(10):1215–1234.

Jackson, R. R. and Cross, F. R. (2011). Spider cognition. *Advances in Insect Physiology*, 41:115–174.

Konieczny, S. and Pino Perez, R. (2002). Sur la représentation des états épistémiques et la révision itérée. In *Révision des croyances*, pages 181–202. Hermes Science Publications.

Kraus, S., Lehmann, D., and Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44. (Page numbers are those of the arXiv version).

Kripke, S. (1965). Semantical analysis of intuitionistic logic I. In Crossley, J. and Dummett, M., editors, *Formal systems and recursive functions*, pages 92–130. North-Holland Publishing Company.

Lehmann, D. and Magidor, M. (1992). What does a conditional knowledge base entail ? *Artificial Intelligence*, 55.

Leitgeb, H. (2004). *Inference on the low-level, an investigation into deduction, nonmomotonic reasoning and the philosophy of cognition*. Kluwer Academic Publishers.

Leopold, D. A. and Logothesis, N. K. (1999). Multistable phenomena: changing views in perception. *Trends in Cognitive Sciences*, 3(7):254–264.

Levy, D. A., Bayley, P. J., and Squire, L. R. (2004). The anatomy of semantic knowledge: Medial vs. lateral temporal lobe. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6710–6715.

Lin, L. et al. (2007). Neural encoding of the concept of nest in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14):6066–6071.

Manns, J. R., Hopkins, R. O., and Squire, L. R. (2003). Semantic memory and the human hippocampus. *Neuron*, 38(1):127–133.

Martin, A. and Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*, 11(2):194–201.

McGaugh, J. (2000). Memory — a century of consolidation. *Science*, 287(2):248–251.

Menzel, R. and Giurfa, M. (2001). Cognitive architecture of a mini-brain: the honeybee. *Trends in Cognitive Sciences*, 5(2):62–71.

O'Carroll, D. (1993). Feature-detection neurons in dragonflies. *Nature*, 362(6420):541–543.

Paulk, A. C., Dacks, A. M., and Gronenberg, W. (2009). Color processing in the medulla of the bumblebee (Apidae: *Bombus impatiens*). *The Journal of Comparative Neurology*, 513(5):441–456.

Paulk, A. C., Phillips-Portillo, J., Dacks, A. M., Fellous, J. M., and Gronenberg, W. (2008). The processing of color, motion, and stimulus timing are anatomically segregated in the bumblebee brain. *The Journal of Neuroscience*, 28(25):6319–6332.

Phelps, E. A. and LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, 48(2):175–187.

Preston, A. R. and Wagner, A. D. (2007). The medial temporal lobe and memory. In Kesner, R. and Martinez, J., editors, *Neurobiology of Learning and Memory*, pages 305–337. Elsevier. 2nd Edition.

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13:587–597.

Sanes, J. R. and Zipursky, S. L. (2010). Design principles of insect and vertebrate visual systems. *Neuron*, 66(1):15–36.

Schildberger, K. (1984). Multimodal interneurons in the cricket brain: properties of identified extrinsic mushroom body cells. *Journal of Comparative Physiology A*, 154(1):71–79.

Schwartz, J., Grimault, N., Hupé, J., Moore, B. C. J., and Pressnitzer, D. (2012). Multistability in perception: binding sensory modalities, an overview. *Philosophical Transactions of the Royal Society B*, 367(1591):896–905.

Shimamura, A. P. (2002). Relational binding theory and the role of consolidation in memory retrieval. In Squire, L. and Schacter, D., editors, *Neuropsychology of memory*, pages 61–72. The Guilford press. 3rd Edition.

Shimamura, A. P. (2010). Hierarchical relational binding in the medial temporal lobe: The strong get stronger. *Hippocampus*, 20(11):1206–1216.

Squire, L. R. and Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, 5(2):169–177.

Squire, L. R. and Bayley, P. J. (2007). The neuroscience of remote memory. *Current Opinion in Neurobiology*, 17(2):185–196.

Squire, L. R., Stark, C. E. L., and Clark, R. E. (2004). The medial temporal lobe. *Annual Review of Neurosciences*, 27:279–306.

Srinivasan, M. V. (2006). Honeybee vision: In good shape for shape recognition. *Current Biology*, 16(2):58–60.

Sterzer, P., Kleinschmidt, A., and Rees, G. (2009). The neural bases of multistable perception. *Trends in Cognitive Sciences*, 13(7):310–318.

Strausfeld, N. J., Hansen, L., Li, Y., Gomez, R. S., and Ito, K. (1998). Evolution, discovery, and interpretations of arthropod mushroom bodies. *Learning & Memory*, 5(1):11–37.

Suzuki, W. A. and Eichenbaum, H. (2000). The neurophysiology of memory. *Annals of the New York Academy of Sciences*, 911:175–191.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cerebral Cortex*, 13(1):90–99.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., and Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 103(21):8239–8244.

Tsunoda, K. et al. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4(8):832–838.

Wessnitzer, J. and Webb, B. (2006). Multimodal sensory integration in insects — towards insect brain control architectures. *Bioinspiration & Biomimetics*, 1(3).

Wilson, R. I. and Mainen, Z. (2006). Early events in olfactory processing. *Annual Review of Neuroscience*, 29:163–201.

Zhang, Y., Lu, H., and Bargmann, C. I. (2005). Pathogenic bacteria induce aversive olfactory learning in *Caenorhabditis elegans*. *Nature*, 438(10):179–184.

Version française

Introduction

La modélisation logique du raisonnement et de l'apprentissage butte depuis des années sur un certain nombre de difficultés tenaces. L'une d'elles est de savoir comment articuler la modélisation du raisonnement à celle de l'apprentissage. En effet, l'acquisition d'information par un agent cognitif déclenche à la fois chez celui-ci l'opération d'inférences et un apprentissage, c'est à dire une révision de ses dispositions à inférer. Les dispositions à inférer d'un agent devraient donc dépendre de la suite des informations qu'il a précédemment acquises, c'est à dire de son expérience passée. Un autre problème, étroitement lié au précédent, est qu'une modélisation satisfaisante de l'apprentissage devrait permettre de représenter des révisions successives, l'apprentissage étant essentiellement un processus continu. Cependant il n'y a aujourd'hui pas de consensus quant à la manière dont les dispositions à inférer d'un agent devraient être révisées [1].

Le présent mémoire introduit un formalisme logique pour la modélisation des inférences qui vise à ouvrir la voie à une solution de ces problèmes. Son idée maîtresse est que pour aborder ces questions il faudrait s'inspirer de la manière dont les agents naturels (c'est à dire les humains et les animaux) procèdent lorsqu'ils opèrent des inférences ou apprennent. Apprentissage et raisonnement sont en effet naturellement articulés chez ces agents, et une lo-

1 L'exemple le plus emblématique en est sans doute le long débat sur la manière dont il conviendrait d'itérer la modélisation de la révision des croyances proposée par Alchourrón, Gardenfors et Makinson ([Alchourrón et al., 1985]). Voir [Konieczny and Pino Perez, 2002] pour un récapitulatif des différentes tentatives en ce sens, ainsi que de leur inconvénients respectifs. Voir aussi [Freund, 2004] pour une proposition ultérieure.

gique qui simulerait leurs processus mentaux serait automatiquement exempte de ces problèmes. Dans une première tentative pour fonder une telle logique, on se concentre ici sur un type d'inférences très simples qui, soutiendra-t-on, peuvent être considérées comme formant le coeur du raisonnement chez les agents naturels. On identifie un processus sous-jacent plausible pour ces inférences, et on développe une famille de modèles logiques permettant de simuler ce processus. On s'attache ensuite à produire un ensemble de règles d'inférence caractérisant les relations d'inférence induites par ces modèles. Ces règles sont la conséquence du processus suggéré, et doivent donc être regardées comme des règles de raisonnement qui émergent du fonctionnement d'un cerveau exécutant ce processus. De ce point de vue, le formalisme introduit montre comment du raisonnement intelligent et de la logique peuvent être produits par un cerveau en marche. Pour finir, on analyse les processus d'apprentissage attachés aux inférences considérées, et on montre comment le formalisme proposé permet de modéliser ces processus.

Chapitre 1

Motivation et interprétation du formalisme logique proposé

Traditionnellement, l'approche utilisée pour la modélisation logique du raisonnement est plus ou moins tacitement une approche normative : on s'intéresse à comment des agent idéaux devraient raisonner, et non pas à comment les agents réels le font. De plus, il s'agit généralement d'une approche purement extérieure, au sens où on ne se préoccupe pas des processus internes (mentaux/neuraux) par lesquels le raisonnement est produit. À l'inverse, l'approche qui va être adoptée ici est une approche interne et descriptive, c'est à dire qu'elle va tenter de rendre compte de ce qui se passe réellement chez les agents naturels lorsqu'ils raisonnent, et de simuler leurs processus de raisonnement. Notre première tache est donc d'identifier ces processus, et pour cela nous allons naturellement nous tourner vers les sciences cognitives. Cette recherche constituera la première partie du présent chapitre. Une fois qu'on aura identifié un processus à la fois plausible et exploitable, on donnera dans la seconde partie un premier aperçu de la modélisation envisagée.

1.1 Le processus mental modélisé

Lorsqu'on considère le raisonnement chez les agents naturels, la première chose qu'on peut observer c'est que celui-ci ne constitue pas un processus unique et

homogène, mais qu'il est formé au contraire d'un grand nombre de facultés cognitives distinctes. Certaines d'entre elles ne se trouvent sans doute que chez les humains, comme par exemple la réflexivité et les capacités de langage supérieures, tandis que d'autres sont plus largement répandues et existent à des degrés divers chez les espèces dites 'supérieures' (comme par exemple la capacité à appréhender les relations), et que d'autres enfin sont omniprésentes à travers le règne animal. On peut supposer sans trop s'avancer que chacune de ces facultés est supportée au niveau cérébral par un processus neuronal sous-jacent. Il est par ailleurs évident que les capacités et processus impliqués dans le raisonnement humain sont extrêmement nombreux et entremêlés, et pour un grand nombre d'entre eux, encore mal connus. C'est pourquoi il ne serait pas raisonnable d'espérer modéliser d'un seul coup l'ensemble du raisonnement humain. Une stratégie plus réaliste est de rechercher un processus élémentaire, simple et suffisamment plausible, et de tenter ensuite de le modéliser.

1.1.1 Les inférences automatiques

Or, parmi les facultés cognitives communément rencontrées chez les agents naturels, il en est une qui semble être à la fois extrêmement basique, omniprésente et essentielle au raisonnement. Il s'agit de la capacité à opérer des inférences non-monotones à partir de l'information disponible.

Chez l'humain, la forme la plus élémentaire sous laquelle elle se manifeste est celle de ces inférences approximatives que nous opérons machinalement en nous fondant sur notre connaissance intuitive des choses et des situations, comme par exemple lorsque, observant un oiseau, nous nous attendons à ce qu'il s'envole si nous tentons de nous en approcher. Une caractéristique essentielle de ces inférences est qu'elles ne sont pas délibérées, mais qu'elles se produisent au contraire de manière automatique. Elles ne nécessitent d'ailleurs même pas notre conscience pour s'opérer, et elles demeurent généralement inconscientes, à moins que nous ne les amenions à notre conscience par un mouvement réflexif

sur nous-mêmes. Pour autant que l'introspection nous permette d'en juger, elles ne semblent pas non plus s'appuyer sur la représentation verbale, comme le suggère le fait que la plupart du temps nous nous contentons d'en 'sentir' les prémisses et les conclusions, sans nous donner la peine de poser des mots dessus. Enfin, elles semblent s'opérer presque sans effort, ce qui rend probable qu'elles reposent sur un processus très simple et peu coûteux. Ces inférences ne constituent bien sûr pas l'ensemble des capacités inférentielles humaines — il est clair que la réflexivité et la représentation verbale donnent lieu à des types de raisonnement bien plus sophistiqués — mais elles jouent un rôle dans la plupart des décisions que nous prenons et des jugements que nous faisons dans notre vie de tous les jours. En particulier, nous utilisons souvent leurs conclusions comme prémisses dans des raisonnements plus complexes, comme ceux justement qui impliquent la réflexivité et la représentation verbale. Elle se présentent donc comme un élément simple, autonome (c'est à dire ne reposant sur aucune autre faculté) et fondamental (c'est à dire sur laquelle d'autres facultés reposent) du raisonnement humain.

Cependant la capacité à opérer de telles inférences est loin d'être spécifique aux humains. L'opération d'inférences rejoint la prise de décision, et ce d'autant plus que les inférences considérées sont automatiques, inconscientes et non supportées verbalement. Mais tous les animaux prennent des décisions, et la question est de savoir si celles-ci résultent d'un réflexe rigide ou d'un processus plus flexible. Si le processus de décision est flexible, c'est à dire si les décisions de l'animal dépendent de manière non-monotone de l'information dont il dispose, alors celui-ci peut être vu comme exécutant un processus inférentiel non-monotone. Or il s'avère que la flexibilité de la prise de décision est la norme plutôt l'exception à travers tout le règne animal. Mêmes des animaux dotés de cerveaux minuscules comme les insectes ou des araignées, dont les capacités de raisonnement sont supposément extrêmement limitées, ont été observés prendre des décisions dépendant de manière non-monotone de

ce qu'ils perçoivent de la situation courante[1]. La capacité à opérer des inférences non-monotones, inconscientes et non supportées verbalement, est donc sans doute l'une des facultés cognitives les plus répandues qui soit.

J'appellerai *inférences automatiques* ce type d'inférences. En raison des similarités profondes qui, malgré d'évidentes différences superficielles, prévalent entre les espèces dans le fonctionnement général des cerveaux, il n'y a pas de raison de supposer que ce type d'inférences est réalisé de manière très différente d'une espèce à une autre. On peut donc imaginer que la capacité à opérer des inférences automatiques repose sur un processus autonome, suffisamment simple pour se loger dans les cerveaux les plus petits, et qu'elle forme le coeur du raisonnement chez les agents naturels. Par contraste, les autres facultés cognitives apparaitraient alors comme des modules additionnels, qui se branchent sur celui-ci pour compléter le raisonnement chez les espèces supérieures.

1.1.2 Un processus plausible pour les inférences automatiques

Pour identifier le processus par lequel s'opèrent les inférences automatiques, on peut s'appuyer sur l'intuition que nous avons de nos propres processus

1 Un cas bien étudié est celui des stratégies de chasse des araignées salticides relativement aux caractéristiques de leurs proies. Par exemple, lorsqu'elle approche une araignée du genre *Scytodes*, une salticide de l'espèce *Portia labiata* choisit normalement de le faire par l'arrière. Comme les différentes espèces de *Scytodes* capturent leurs proies en leur crachant dessus à distance une soie imprégnée de venin, cette stratégie minimise le risque pour *P. labiata* de se faire capturer et dévorer par sa proie. Mais les femelles *Scytodes* transportent généralement leurs oeufs dans des cocons qu'elles tiennent avec leurs pièces buccales, et cracher requiert qu'elles déposent d'abord leurs oeufs, ce qu'elles sont réticentes à faire. De ce fait, les femelles *Scytodes* transportant des oeufs sont moins enclines à cracher que les *Scytodes* ne transportant pas d'oeufs. Lorsqu'elle approche une femelle *Scytodes* transportant des oeufs, *P. labiata* renonce généralement à faire un détour, et choisit plutôt une approche frontale qui lui permet de mordre sa proie au niveau des régions céphaliques, assurant ainsi une meilleure efficacité à son venin ([Jackson and Cross, 2011] pp. 130-131). De nombreux exemples de 'stratégies conditionnelles' de ce type sont rapportées dans la littérature sur les salticides (voir par exemple [Bartos, 2008]). Un autre exemple frappant de prise de décision non-monotone chez les salticides (mais mettant sans doute en jeu plus qu'une simple prise de décision non-monotone) est celui de la recherche de solution par essais et erreurs. Typiquement, étant donné une situation et un but à atteindre (par ex., sortir d'une situation problématique), l'araignée va opter pour une stratégie A. Si elle réussit, elle réutilisera par la suite cette même stratégie si elle se trouve à nouveau placée dans une situation semblable. Mais si elle échoue, elle renoncera à la stratégie A et essayera à la place une stratégie alternative B ([Jackson et al., 2001].

mentaux (c'est à dire faire appel à l'introspection), ou examiner la machinerie neuronale qui les supporte. Ces deux méthodes ne sont pas exclusives, et elles se complètent même utilement.

L'introspection

L'introspection pour sa part suggère que pour réaliser des inférences automatiques, nous nous appuyons sur les représentations mentales que nous avons des objets et des situations que nous connaissons. Par exemple, si nous voulons décider si un oiseau que nous sommes en train d'observer va s'envoler à notre approche, nous faisons appel à notre expérience des oiseaux. Cela veut dire que nous cherchons dans notre mémoire des souvenirs de situations semblables, et que nous vérifions si dans ces situations l'oiseau s'est envolé ou non. S'il s'est envolé dans toutes les situations semblables dont nous pouvons nous rappeler, nous concluons qu'il en fera de même dans le cas présent.

Il faut noter cependant que ces souvenirs ne sont pas forcément des souvenirs individuels de chaque situation que nous avons vécue au cours de notre vie. Au contraire, le processus de mémorisation compile généralement les expériences similaires en une mémoire unique, laissant de côté les détails insignifiants. Par exemple, nous avons une idée de ce qu'est un moineau, sans pour autant nous souvenir de toutes situations dans lesquelles nous avons pu en observer un. Les souvenirs que nous avons des choses et sur lesquels nous nous appuyons pour réaliser nos inférences automatiques sont donc des représentations archétypales, c'est à dire que ce sont des sortes de concepts.

Mais ce ne sont pas n'importe quelle sorte de concepts. Comme on le sait, les concepts peuvent être plus ou moins généraux, les concepts les plus généraux subsumant les moins généraux et formant ainsi des chaines convergentes de concepts de plus en plus généraux. Un point crucial ici est que pour opérer des inférences automatiques, nous faisons appel seulement aux plus spécifiques de nos concepts. Ces derniers sont en effet les représentations les plus précises et les plus détaillées que nous avons des choses, et donc les mieux à même d'infor-

mer nos inférences. Nous examinons ces représentations les plus précises pour essayer d'en trouver qui correspondent à ce que nous percevons de la situation présente, c'est à dire qui satisfont toutes les caractéristiques que nous pouvons en observer, et nous vérifions si celles que nous avons trouvées satisfont aussi la caractéristique 's'envole'. Si toutes le font, alors nous concluons que l'oiseau va s'envoler.

Mais ce faisant nous ne vérifions pas tous nos concepts les plus spécifiques qui correspondent à ce que nous percevons de la situation actuelle, ce qui serait bien trop long et bien trop exigeant mentalement. Au lieu de cela, nous vérifions simplement les premiers qui nous viennent à l'esprit, c'est à dire ceux qui sont les plus prégnants dans notre mémoire. Si tous ceux-là satisfont la caractéristique 's'envole', alors nous concluons que l'oiseau va s'envoler. C'est cette remémoration partielle qui donne aux inférences automatiques leur caractère non-monotone. En effet, une information plus complète sur la situation courante réactiverait sans doute un ensemble de souvenirs différent, nous amenant à d'autres conclusions. Il faut cependant souligner que cette récupération partielle de nos souvenirs n'affecte pas significativement la pertinence de nos conclusions, dans la mesure où nos souvenirs les plus prégnants sont aussi généralement ceux qui sont les plus significatifs et les plus importants pour nous, et donc ceux qui sont les mieux à même de nous fournir des conclusions utiles.

Cependant il peut aussi arriver que nous ne trouvions en nous aucun souvenir d'objet ou de situation semblable, et que par conséquent le cas présent nous apparaisse comme nouveau. Dans ce cas, le processus décrit ci-dessus ne peut pas se réaliser, et ceci déclenche en retour des processus d'apprentissage. Ceux-ci consistent essentiellement en la création d'une représentation nouvelle, ou en la complétion de celles déjà existantes.

On objectera peut-être qu'un tel processus ne peut être supposé chez les animaux, parce qu'il repose de manière essentielle sur les représentations mentales et (plaidera-t-on peut-être) les animaux n'ont pas de représentations mentales.

De fait, l'idée que les animaux puissent avoir des représentations mentales a longtemps été déconsidérée, ceux-ci étant principalement regardés, notamment par les philosophes, comme étant de simples machines répondant de manière automatique aux stimuli reçus. Cependant cette vue a été sérieusement remise en cause au cours de ces dernières décennies par les études sur la cognition animale, et il est aujourd'hui largement admis par les chercheurs dans ce domaine que la plupart des animaux, y compris ceux dotés de cerveaux minuscules comme les arthropodes, ont des représentations mentales[2]. La réticence que le non-spécialiste peut avoir à imaginer les animaux, et plus particulièrement ceux à petits cerveaux, comme ayant des représentations mentales, vient sans doute du fait qu'il/elle a tendance à associer au sens de ces mots l'intuition qu'il/elle a de sa propre expérience subjective humaine, ce qui n'est pas nécessaire. Du point de vue cognitiviste — et ainsi par exemple que le formulent Jackson et Cross[3] — une représentation mentale est beaucoup plus simplement *'quelque chose qui ressemble davantage à un état interne véhiculant de l'information, et qui est utilisé lors de la prise de décision. Une idée clé est que les représentations sont utilisées dans des processus qui ont lieu plusieurs pas en retrait par rapport aux simples enchainements stimulus-réponse'*. C'est dans un sens dépassionné comme celui-ci qui nous prenons le terme.

Les neurosciences

Les neurosciences pour leur part permettent d'imaginer comment un processus comme celui envisagé ci-dessus pourrait être réalisé dans un cerveau. À première vue, les cerveaux présentent des différences notables d'une espèce à une autre, notamment en termes de taille ou de plan anatomique. Mais par delà ces différences évidentes, tous partagent un même schéma fonctionnel général.

2 Voir par exemple [Srinivasan, 2006], [Chittka and Niven, 2009] et [Jackson and Cross, 2011] sur les représentations mentales chez les arthropodes.

3 [Jackson and Cross, 2011] p. 120.

En particulier, chez tous les espèces dotées d'un cerveau la perception s'effectue par le biais d'un ensemble de canaux sensoriels parallèles, un pour chaque modalité sensorielle (vision, olfaction, etc.) présente dans l'espèce considérée. Chacun de ces canaux collecte de l'information en provenance du monde extérieur au moyen d'un ensemble de neurones récepteurs spécifiques (photo-récepteurs de la rétine pour le canal visuel, récepteurs chimiques dans le canal olfactif, etc.) et traite ensuite cette information pour en extraire des caractéristiques pertinentes. Par exemple, les bords, les couleurs ou les mouvements directionnels sont des caractéristiques typiquement extraites au sein des canaux visuels, tandis que les odeurs particulières sont des caractéristiques extraites dans les canaux olfactifs. Dans tous les cas, l'extraction d'une caractéristique donnée est réalisée par un ensemble de neurones dédiés, qui répondent sélectivement à cette caractéristique [4]. À mesure qu'on progresse à l'intérieur d'un canal sensoriel, les caractéristiques extraites peuvent se combiner entre elles pour former des caractéristiques de niveau supérieur, comme par exemple un arrangement angulaire particulier entre plusieurs bords, ou une texture particulière [5]. Les caractéristiques extraites dans un canal donné dépendent de l'organisation neuronale particulière de celui-ci, et donc le type, le nombre et la complexité des caractéristiques qui sont finalement perçues varie d'une espèce

4 Les travaux fondateurs dans ce domaine sont ceux de Hubel et Wiesel, qui vers les années soixante ont identifié des détecteurs de bords dans les cortex visuels primaires du chat et du singe ([Hubel and Wiesel, 1959, 1962, 1968]). Depuis, de nombreuses études ont été menées chez un grand nombre d'espèces, qui ont démontré l'existence de détecteurs de caractéristiques tant chez les vertébrés que chez les invertébrés, et ceci pour toutes les modalités sensorielles. Sur les caractéristiques visuelles chez les invertébrés, on peut consulter par exemple [O'Carroll, 1993] (détecteurs de bords chez la libellule), [Barnett et al., 2007] (détecteurs de mouvements chez la mouche) et [Paulk et al., 2008, 2009] (détecteurs de couleurs et de mouvement chez le bourdon). Pour une rapide comparaison entre les canaux visuels et olfactifs des vertébrés et des invertébrés, on peut consulter [Chittka and Niven, 2009] p. 996-997 ; pour une étude plus précise des canaux olfactifs des vertébrés et des invertébrés, voir [Wilson and Mainen, 2006] ; pour une analyse détaillée des homologies entre les premiers niveaux des systèmes visuels des insectes et des vertébrés, voir [Sanes and Zipursky, 2010].

5 Pour un aperçu rapide de l'extraction de caractéristiques complexes dans différents canaux sensoriels, voir [Taylor et al., 2006] p. 8239. Pour des exemples de caractéristiques visuelles complexes chez le macaque, voir [Tsunoda et al., 2001] et [Tanaka, 2003].

à l'autre, mais la règle générale dans toutes les espèces dotées d'un cerveau, des arthropodes aux humains, est que l'information sensorielle est analysée en caractéristiques.

Et de même, dans toutes les espèces pourvues d'un cerveau les canaux sensoriels débouchent finalement sur un certain nombre d'aires cérébrales centrales. Celles-ci sont diversement organisées — le plan anatomique du cerveau d'un insecte et celui d'un mammifère sont évidement très différents — mais toutes ont en commun deux caractéristiques principales. Premièrement, elles contiennent des neurones qui reçoivent des signaux en provenance de différents canaux sensoriels, et qui répondent donc à la co-occurence de caractéristiques ; et deuxièmement, elles sont impliquées dans la mémoire et l'apprentissage [6].

Chez les insectes par exemple, les protocerebrum latéral et médial et les mushroom bodies — trois structures neuronales fortement interconnectées situées dans le cerveau central — reçoivent des signaux directs ou indirects en provenance de tous les canaux sensoriels [7]. Il a été montré que certains neurones des mushroom bodies répondent à des stimuli relevant de plusieurs modalités sensorielles différentes, et même, ce qui est plus intéressant encore, à des com-

6 Le lecteur peu familier avec les études sur la cognition animale sera peut-être réticent à imaginer les animaux à petits cerveaux comme étant capables d'apprendre, l'idée contraire ayant pendant longtemps été considérée comme allant de soi par beaucoup. Cependant au cours de ces dernières décennies de nombreuses preuves de la capacité d'apprentissage de ces animaux ont été apportées. Même le minuscule vers nématode *C. elegans*, dont le système nerveux est composé d'exactement 302 neurones, s'est révélé être capable d'apprendre a éviter des stimuli associés à des environnements nocifs ([Zhang et al., 2005]). Si bien que l'idée qui fait son chemin aujourd'hui est plutôt que *'la capacité d'apprentissage pourrait être une propriété émergente des systèmes nerveux, et donc tous les animaux dotés d'un système nerveux devraient être capables d'apprendre'* ([Hollis and Guillette, 2011] p. 24). Cela ne signifie cependant pas que tout apprentissage doit nécessairement s'appuyer sur des représentations mentales, et on pourrait valablement arguer que le système nerveux de *C. elegans* est trop rudimentaire pour pouvoir être appelé un cerveau, et aussi pour pouvoir stocker des représentations. L'idée est plutôt que la pression sélective en faveur de l'apprentissage a fait émerger les souvenirs et les concepts, parce que ceux-ci sont un support très efficace pour l'apprentissage.

7 On pourra trouver un schéma général des interconnexions neuronales au sein du cerveau de l'insecte par exemple dans [Strausfeld et al., 1998] p. 30.

binaisons spécifiques de tels stimuli[8]. Par ailleurs, les mushroom bodies sont depuis longtemps connus pour le rôle décisif qu'ils jouent dans la mémoire et l'apprentissage chez les insectes[9], même s'il y a aujourd'hui débat pour savoir s'ils sont eux-mêmes le siège de ces facultés cognitives ou s'ils ne font que relayer l'information vers des centres nerveux supérieurs situés dans le protocerebrum médial et latéral, où ces facultés seraient effectivement réalisées[10].

De même chez les mammifères, l'information sensorielle converge vers le lobe médial temporal (LMT), une région cérébrale connue depuis longtemps pour son implication déterminante dans l'apprentissage et la mémoire. Mais ici il nous faut regarder les choses plus en détail, car la psychologie cognitive distingue généralement plusieurs formes de mémoire dont toutes ne dépendent pas du LMT, et qui ne jouent sans doute pas toutes non plus un rôle dans les inférences automatiques. En effet, bien que la définition et l'étendue exacte de chacune d'entre elles fassent toujours l'objet de discussions, les chercheurs du domaine s'accordent généralement pour distinguer au moins quatre sortes de mémoire.

Une première est appelée la *mémoire de travail*. Elle est définie comme une mémoire à court terme qui stocke l'information nécessaire à l'exécution de taches complexes. C'est par exemple la mémoire qui nous permet de garder à l'esprit les différentes parties d'un problème tandis que nous le résolvons, ou de nous souvenir d'avoir déjà accompli une sous-tache particulière comme partie d'une tache plus complexe. Une deuxième est connue sous le nom de *mémoire procédurale*. Elle est vue comme la mémoire responsable de l'apprentissage des habilités techniques, comme par exemple savoir faire du vélo. Contrairement à la mémoire de travail, c'est une mémoire à long terme. Elle a aussi la caracté-

8 Voir par exemple [Schildberger, 1984] chez le criquet, et [Strausfeld et al., 1998] pp. 27–31 chez la blatte. Pour un exposé général sur l'intégration multi-sensorielle chez les insectes, voir [Wessnitzer and Webb, 2006].

9 Pour une brève revue historique de l'idée de mushroom bodies comme centres de la mémoire et de l'apprentissage, voir [Strausfeld et al., 1998] p. 14.

10 [Strausfeld et al., 1998] pp. 31–32.

ristique notable qu'elle agit généralement en dehors de la conscience du sujet. Une troisième sorte de mémoire est appelée la *mémoire épisodique*. Elle est définie comme la mémoire des événements personnellement vécus, et comme telle elle est vue comme conservant la trace des épisodes marquants de la vie du sujet. C'est une mémoire à long-terme, tout comme la mémoire procédurale. Enfin, le dernier type de mémoire généralement distingué est appelé la *mémoire sémantique*. Elle est définie comme une mémoire à long-terme qui stocke la connaissance que le sujet peut avoir des faits généraux — autrement dit, ce qu'il sait des choses. Les mémoires sémantique et épisodique sont étroitement liées, dans la mesure où la plus grande partie du savoir qu'un sujet peut avoir sur les choses est tiré de sa propre expérience personnelle. Prises ensemble, elles sont connues sous le nom de *mémoire déclarative*. Contrairement à la mémoire procédurale, la mémoire déclarative requiert la conscience du sujet, au sens où on ne peut se souvenir que des faits et des événements que l'on a consciemment vécus [11].

Parmi ces différentes sortes de mémoire, la mémoire sémantique apparait naturellement comme le support très probable des inférences automatiques, et c'est donc à celle-ci que nous nous intéresserons principalement, ainsi qu'à ses relations avec la mémoire épisodique. Contrairement à la mémoire procédurale et à la mémoire de travail, les mémoires épisodique et sémantique dépendent de manière critique du LMT. En témoigne le fait que des lésions du LMT induisent de sévères déficiences au niveau des mémoires sémantique et épisodique, mais pas au niveau de la mémoire procédurale ni de la mémoire de travail. Plus précisément, les lésions du LMT entrainent une profonde

11 Il faut cependant souligner que cela ne signifie pas qu'elle requière la conscience de soi. Il ne faut en effet pas confondre conscience de quelque chose et conscience de soi. Le chat qui épie un oiseau est conscient du fait qu'il y a là un oiseau, mais cela ne signifie pas pour autant qu'il est conscient de sa propre conscience qu'il y a là un oiseau. La conscience de soi n'est qu'une forme particulière de conscience, qui ne peut exister que chez les animaux possédant les canaux appropriés leur permettant de collecter de l'information sur leurs propres états mentaux, et de l'intégrer à celle issue de leurs autres canaux pour la traiter. La plupart des animaux n'ont pas conscience d'eux-mêmes, bien qu'ils soient parfaitement conscients des évènements et des objets auxquels ils réagissent.

amnésie *antérograde* (incapacité à former de nouveaux souvenirs épisodiques et sémantiques), accompagnée d'une amnésie *rétrograde* (perte des souvenirs épisodiques et sémantiques acquis avant la survenue des lésions). Cette dernière est typiquement temporellement gradée, c'est à dire que les souvenirs les plus récents sont perdus tandis que les plus anciens sont épargnés [12]. Chez l'humain par exemple, l'amnésie rétrograde qui suit des lésions du LMT peut selon l'étendue des dommages [13] couvrir une période de temps variant de quelques années à plusieurs décennies avant la survenue des lésions. Cette gradation temporelle a conduit à l'idée largement admise aujourd'hui que le LMT joue un rôle critique dans la formation des souvenirs épisodiques et sémantiques, mais que ces derniers sont ensuite progressivement 'sauvegardés' dans d'autres aires du cerveau, devenant ainsi avec le temps indépendants du LMT [14].

Pour comprendre comment le LMT peut former les souvenirs épisodiques et sémantiques, il peut être utile de jeter un oeil sur son organisation anatomique. Le LMT est une structure composée et hiérarchiquement organisée, qui comprend la formation hippocampale et les cortex entorhinal, perirhinal, et parahippocampal. À la base de la hiérarchie, les cortex perirhinal et parahippocampal reçoivent des signaux directs ou indirects en provenance des différents canaux sensoriels, et notamment des canaux visuel, somato-sensoriel, olfactif, et auditif [15]. Les cortex perirhinal et parahippocampal fournissent ensuite la majeure partie des signaux reçus par le cortex entorhinal, lequel fournit à son tour la majeure partie des signaux reçus par l'hippocampe, situé au plus haut

12 Pour un exposé général sur le rôle joué par le LMT dans la mémoire déclarative, voir [Preston and Wagner, 2007] ; pour une étude détaillée de son rôle dans l'acquisition des souvenirs sémantiques, voir [Bayley and Squire, 2005] et [Levy et al., 2004] ; pour une étude des amnésies antérograde et rétrograde relatives aux connaissances sémantiques, voir [Manns et al., 2003] et [Bayley et al., 2006].

13 [Squire and Bayley, 2007] pp. 185–186.

14 Voir p. 133 ci-dessous.

15 Voir par exemple [Suzuki and Eichenbaum, 2000] p. 176.

niveau de la hiérarchie [16]. Toutes ces connexions sont réciproques, ce qui suggère que l'information circule dans les deux sens. De plus, tous les composants du LMT ont des connexions substantielles réciproques avec l'amygdale, une région adjacente du cerveau connue pour son implication dans le traitement des émotions [17].

Au sommet de la hiérarchie, il a été découvert dans le cortex entorhinal et l'hippocampe des neurones qui répondent de manière spécifique à la présentation de divers stimuli correspondant à un même concept, indépendamment des autres qualités de ces stimuli. Fait notable, ces résultats n'ont pas été obtenus que chez l'humain, mais aussi chez des mammifères non-humains. Par exemple, des neurones répondant de manière spécifique à des situations impliquant la notion de nid ont été trouvés dans l'hippocampe de la souris. Plus précisément, trois types différents de neurones ont été identifiés. Un premier type de neurones répondait exclusivement aux situations dans lesquelles les souris se trouvaient dans un nid, déchargeant de manière robuste et continue chaque fois qu'elles se trouvaient dans un nid et restant presque silencieux sinon. Un second type de neurones faisait exactement le contraire, répondant aux situations dans lesquelles les souris ne se trouvaient pas dans un nid (c'est à dire qu'ils déchargeaient chaque fois que les souris n'étaient pas dans un nid, et cessaient de le faire dès qu'elles entraient dans un nid). Enfin, un dernier type de neurones répondait aux situations dans lesquelles les souris arrivaient à un nid, déchargeant brièvement mais fortement chaque fois qu'elles atteignaient un nid, et seulement à ce moment-là. Ces trois types de neurones étaient insensibles aux nombreux autres objets présentés aux souris, et qui plus est leurs réponses ne dépendaient pas des qualités particulières des nids telles que leur forme géométrique, leur localisation, leur odeur ou le matériau dont ils étaient

16 Pour un schéma synthétique des interconnexions neuronales existant entre les structures qui composent le LMT, voir [Quiroga, 2012] p. 592 ; pour une description plus approfondie, voir [Suzuki and Eichenbaum, 2000] pp. 176–178 et [Preston and Wagner, 2007] pp. 306–308.

17 Voir par exemple [Phelps and LeDoux, 2005].

faits. De fait, le seul critère semblant déterminer la réponse de ces neurones était le fait que l'objet considéré pouvait ou non être utilisé comme un nid par les souris [18].

De même chez l'humain, des neurones répondant de manière spécifique à des personnes ou à des monuments particuliers ont été trouvés dans l'hippocampe et le cortex entorhinal. A nouveau, les neurones identifiés répondaient à n'importe quel stimulus correspondant à leur concept-cible, indépendamment de ses qualités particulières et indépendamment même de sa modalité, et ne répondaient à aucun autre des nombreux stimuli présentés aux sujets. Par exemple, un neurone répondant indifféremment à plusieurs photos d'une actrice célèbre, ainsi qu'à son nom écrit ou prononcé, a été identifié dans une étude. D'autres neurones cependant se montraient moins sélectifs, et répondaient à plusieurs stimuli de l'échantillon utilisé. Mais dans ce cas ces derniers étaient clairement liés, et pouvaient être vus comme différentes instances d'un concept plus général. Par exemple, l'un des neurones étudiés répondait à la fois à la tour Eiffel et à la tour de Pise, tandis qu'un autre répondait à différents acteurs d'une même série télévisée, et un autre encore aux images d'animaux [19].

Il faut préciser ici que le fait qu'un seul neurone répondant à un concept donné ait été identifié à chaque fois ne signifie en aucun cas que ce neurone était le seul à répondre à ce concept dans le cerveau du sujet. Étant des cellules vivantes, les neurones sont intrinsèquement sujets au bruit, aux dommages et à la mort, et la robustesse des processus cérébraux repose très largement sur la redondance [20]. Par conséquent, il est probable que pour chaque concept re-

18 [Lin et al., 2007].

19 [Quiroga, 2012].

20 Il existe cependant des exceptions à cette règle générale. Une exception bien documentée se trouve chez les insectes, où on a pu montrer qu'un unique neurone joue à lui seul le rôle de signal de récompense relativement aux stimuli gustatifs (voir par exemple [Menzel and Giurfa, 2001]p. 63). D'une manière générale, la redondance semble être bien plus faible chez les arthropodes, ce qui est probablement lié au fait que le risque de mort de neurones est moins critique chez ces animaux à la vie courte que chez ceux qui vivent plus longtemps.

présenté dans le cerveau, il y a toute une assemblée de neurones y répondant spécifiquement. Si lors des expériences menées, un seul d'entre eux a été trouvé à chaque fois, c'est parce que pour des raisons techniques et/ou médicales seuls un petit nombre de neurones ont été enregistrés à chaque session.

Parallèlement à ces recherches, un autre groupe d'études a montré que les neurones de l'hippocampe et du cortex entorhinal ne s'activent pas seulement lorsque le sujet perçoit les stimuli qui leur correspondent, mais aussi lorsqu'il imagine ou se remémore ces mêmes stimuli[21]. Autrement dit, ils s'activent chaque fois que le sujet 'pense' à leurs stimuli-cible.

Considérés conjointement, la structure hiérarchique du LMT et la sélectivité des réponses des neurones entorhinaux et hippocampaux suggère fortement que les neurones du LMT intègrent les caractéristiques co-occurrentes issues des différentes modalités sensorielles pour composer des représentations multimodales des objets et des situations rencontrés. L'idée communément admise est que chaque neurone du LMT à son propre niveau hiérarchique reçoit des signaux convergents de la part d'un ensemble défini de neurones du niveau inférieur, et encode ainsi la conjonction des (conjonctions de) caractéristiques encodées par ces derniers[22]. Par exemple, les cortex perirhinal et parahippocampal sont généralement considérés comme encodant des conjonctions de caractéristiques de différentes modalités sensorielles. Pour expliquer leur fonctionnement en parallèle, il a été suggéré que le cortex perirhinal pourrait être plus particulièrement impliqué dans la représentation des objets[23], et le cortex parahippocampal dans celle de leurs contextes[24]. Ces représentations séparées seraient ensuite réunies dans le cortex entorhinal et l'hippocampe, où s'ajouterait aussi de l'information d'ordre émotionnel en provenance

21 [Gelbard-Sagiv et al., 2008].

22 Voir par exemple [Shimamura, 2010] pp. 11206-1209.

23 [Taylor et al., 2006] ; voir aussi [Holdstock et al., 2009].

24 [Aminoff et al., 2013].

de l'amygdale. Les neurones entorhinaux et hippocampaux encoderaient donc ainsi des représentations mentales d'objets et de situations qui seraient à la fois multimodales, contextualisées, et pourvues d'une valeur émotionnelle. Il est cependant difficile de distinguer plus avant les rôles respectifs des neurones entorhinaux et hippocampaux, car ceux-ci font toujours aujourd'hui l'objet d'intenses débats.

Une autre propriété remarquable des neurones du LMT est leur capacité à modifier leur schéma de connexions presque en temps réel. De l'avis général, cette flexibilité repose très vraisemblablement sur un mécanisme neural appelé *potentialisation à long-terme* (PLT), grâce auquel les connexions entre neurones co-actifs peuvent se renforcer rapidement et durablement [25]. Concrètement, un neurone ayant des ramifications à proximité d'un neurone du niveau hiérarchique supérieur mais n'ayant que peu de connexions avec lui pourrait, s'il se trouve être activé en même temps que lui, voir ses connexions avec ce neurone renforcées, et devenir ainsi l'un de ses neurones-inputs. De cette façon, la (conjonction de) caractéristique(s) encodée par le neurone de niveau inférieur pourrait être rapidement intégrée à la conjonction de caractéristiques encodée par le neurone de niveau supérieur [26].

Par ailleurs, la réciprocité des connexions au sein du LMT suggère que l'activation des neurones du LMT pourrait activer en retour les neurones encodant les caractéristiques dans les régions sensorielles supérieures, apportant ainsi aux neurones du LMT un contenu sémantique subjectif [27]. Par exemple, le simple fait de penser au concept de rouge suffit à provoquer en nous une réminiscence de la sensation associée à la couleur rouge. Cela pourrait s'expliquer par le fait que l'activation des neurones du LMT qui répondent au concept de rouge déclenche l'activation de ceux qui encodent la caractéristique 'rouge' dans nos aires visuelles supérieures.

25 Pour un exposé général sur le mécanisme de potentialisation à long terme, voir par exemple [Izquierdo, 1993].

26 Une proposition de ce genre peut se trouver par exemple dans [Shimamura, 2010] p. 1206.

27 [Martin and Chao, 2001], [Eichenbaum, 2000] pp. 47-48, [Squire et al., 2004] pp. 282–283.

Ces considérations ont conduit à la vision dominante selon laquelle l'une des fonctions principales des neurones du LMT serait de former des représentations mentales (c'est à dire, des concepts) des objets et situations rencontrés, en liant rapidement ensemble les caractéristiques simultanément perçues à cette occasion [28]. Cependant, le fait que certains de ces neurones répondent de manière spécifique à des personnes et des objets individuels et d'autres à divers stimuli ayant un lien de similarité, semble indiquer que les neurones du LMT peuvent encoder aussi bien des concepts spécifiques que des concepts plus généraux. Comment l'encodage des représentations plus générales s'articule avec celui des représentations plus spécifiques et notamment avec celui des souvenirs individuels est une question qui reste ouverte, mais il semble évident que le contenu sémantique d'une représentation plus générale — c'est à dire, l'ensemble de caractéristiques qui la compose — devrait dépendre de celui des représentations plus particulières. Une possibilité serait que les représentations plus générales soient issues des plus particulières comme l'ensemble de leurs caractéristiques remarquables communes. En effet, il semble que (par exemple) notre concept général d'oiseau est abstrait des concepts d'oiseaux plus spécifiques que nous pouvons avoir, comme par exemple des concepts d'espèces particulières d'oiseaux, des représentations mentales d'oiseaux individuels, ou encore des souvenirs de situations particulières impliquant des oiseaux. Ces concepts plus spécifiques seraient à leur tout abstraits d'autres encore plus spécifiques, et ainsi de suite jusqu'à nos souvenirs individuels de situations particulières impliquant des oiseaux [29]. Avec le temps, les souvenirs individuels à l'origine d'un concept donné finiraient généralement par s'effacer, mais le concept lui-même resterait comme la trace de l'encodage répété de leurs caractéristiques remarquables communes.

28 [Preston and Wagner, 2007] pp. 312–313.

29 On ne tient pas compte ici du cas particulier et strictement humain des concepts qui peuvent être acquis au moyen de définitions verbales.

Au niveau cérébral, ce processus pourrait résulter du fait que les neurones qui supportent des représentations similaires (c'est à dire, ayant beaucoup de caractéristiques en commun) répondent à des signaux qui sont eux aussi similaires, si bien qu'à cause du bruit neuronal il peut arriver qu'un neurone censé répondre à un certain signal réponde à un moment donné par accident à un autre signal suffisamment semblable. Dans ce cas, l'action combinée de la potentialisation à long terme et de la dépression à long terme (DLT, un mécanisme neuronal complémentaire à la PLT, qui diminue rapidement et durablement l'efficacité des synapses [30]) pourrait modifier les connexions de ce neurone de manière à ce que désormais il réponde à la conjonction des caractéristiques qui sont communes aux deux signaux. Ainsi, une fraction des neurones répondant initialement de manière spécifique à l'un ou l'autre d'une classe donnée de signaux similaires pourrait se mettre progressivement à répondre au concept général qui les subsume toutes. Dans certains cas, les représentations originelles pourraient finir par perdre tout support neuronal, expliquant la perte des souvenirs individuels à l'origine du concept.

Que le processus suggéré ci-dessus soit ou non correct dans ses détails, il est en tous cas très probable que les neurones du LMT sont capables d'extraire des caractéristiques pertinentes — c'est à dire, notamment, invariantes — du flot continu et changeant de la mémoire immédiate, et de les lier entre elles pour composer des représentations plus générales et plus durables, c'est à dire des concepts. Le degré de généralité de ces concepts peut être très variable, certains d'entre eux étant très généraux et d'autres bien plus spécifiques. Dans certains cas, les souvenirs individuels à l'origine d'un concept peuvent se maintenir un certain temps, mais ils finissent généralement par s'effacer au cours du processus d'abstraction.

30 Pour une introduction générale et accessible au phénomène de dépression à long terme, on peut consulter [Bear and Abraham, 1996]. On s'intéresse ici plus particulièrement à la dépression à long terme *hétérosynaptique* ([Bear and Abraham, 1996], pp. 437-442), qui diminue sélectivement l'efficacité des synapses d'entrée inactives d'un neurone actif.

Comme on l'a dit, il est probable que les neurones du LMT renvoient des signaux excitatifs vers les aires sensorielles supérieures du cerveau. Selon la théorie dominante, les neurones de ces zones qui sont régulièrement co-(ré)activés par une même assemblée de neurones du LMT pourraient progressivement développer entre eux des interconnexions directes, et finir ainsi par former un réseau autonome susceptible de supporter par lui-même la représentation correspondante [31]. Les représentations mentales encodées dans le cortex entorhinal et l'hippocampe se verraient ainsi progressivement 'sauvegardées' de manière distribuée entre les différentes aires sensorielles supérieures, ce qui expliquerait le caractère temporellement gradé de l'amnésie rétrograde qui suit les lésions du LMT [32].

Pour autant, cela ne signifie pas que les neurones du cortex entorhinal et de l'hippocampe qui encodent une représentation cessent de le faire après que celle-ci ait été sauvegardée dans les aires sensorielles supérieures. Au contraire, il est probable qu'ils continuent à la supporter aussi longtemps qu'elle fait sens pour le sujet [33]. Ces neurones apparaissent donc comme le premier et le principal support des concepts dans le cerveau des mammifères. C'est pourquoi nous les appellerons *neurones-concept* [34].

Leur caractère de support premier des concepts rend probable que ces neurones sont impliqués dans un grand nombre de fonctions mentales mettant en jeu les concepts et la mémoire. En particulier, il a été suggéré que les transitions entre concepts dans un esprit en train de penser pourraient reflé-

31 [Squire and Alvarez, 1995] pp. 171–174.

32 Voir par exemple [Shimamura, 2002], [Squire et al., 2004] p. 296 et [Preston and Wagner, 2007] pp. 312–315.

33 Cette idée est notamment défendue dans [Quiroga, 2012] pp. 594–595, et [Shimamura, 2002].

34 Le terme est emprunté à [Quiroga, 2012], bien que contrairement à lui nous n'y englobions pas tous les neurones du LMT qui présentent des réponses sélectives, mais seulement ceux du cortex entorhinal et de l'hippocampe. En effet, les neurones des régions inférieures du LMT semblent plutôt ne convoyer que des représentations partielles vers le cortex entorhinal et l'hippocampe, lesquels se chargeraient ensuite de les intégrer pour composer des concepts complets.

ter les transitions dans l'activation des assemblées de neurones-concept qui les supportent [35]. Une autre suggestion intéressante (qui n'est pas incompatible avec la précédente) a été que les souvenirs d'épisodes qui constituent la mémoire épisodique pourraient être encodés dans l'hippocampe sous la forme de séquences ordonnées de représentations mentales discrètes encodées par des assemblées de neurones concepts [36].

Reprenant la première des ces propositions, on suggère ici que les neurones concepts pourraient en particulier être le support d'une forme très particulière de transition entre concepts, à savoir l'opération d'inférences automatiques. Cette hypothèse semble naturelle si on considère que du point de vue évolutif le caractère adaptatif de la mémoire tient essentiellement au fait qu'elle permet de stocker de l'information issue des expériences passées pour l'utiliser dans la prise de décision, et adapter ainsi les décisions prises au milieu dans lequel l'agent considéré vit [37].

Sur le plan neural, un tel processus de transition pourrait reposer sur un mécanisme très simple. Puisqu'une représentation mentale est rappelée à l'esprit dès lors que les neurones qui la supportent sont actifs, on peut supposer que sa prégnance — c'est à dire la facilité avec laquelle elle peut-être rappelée — dépend de la capacité de ces neurones à s'activer en réponse à un signal approprié. Des différences de taille numérique entre assemblées de neurones, couplées à des connexions excitatrices mutuelles entre neurones proches supportant un même concept, pourraient induire des différences de temps de

35 C'est notamment l'idée défendue dans [Quiroga, 2012].

36 [Eichenbaum, 2004].

37 On trouve des idées similaires dans un certain nombre d'articles assez récents ([Buckner, 2010] offre un tour d'horizon rapide de la question). Cependant ces articles s'intéressent généralement à des formes d'inférence qui sont plus complexes que celle que nous étudions ici, et notamment qui reposent de manière décisive sur la mémoire épisodique. Or, il est généralement admis que la plupart des mammifères non-humains ont, au mieux, une mémoire épisodique très limitée, alors même qu'il opèrent indubitablement des inférences, au sens défini p. 119 ci-dessus. Il s'ensuit que les formes d'inférence étudiées dans ces articles ne peuvent représenter à elles seules la totalité des inférences basées sur la mémoire, et aussi que les inférences automatiques ne dépendent pas de la mémoire épisodique.

réponse entre assemblées de neurones supportant des concepts différents. Par ailleurs, des connexions inhibitrices mutuelles entre neurones proches supportant des concepts différents pourraient entraver l'activation des assemblées de neurones dont les temps de réponse sont les plus longs. De ce fait, des assemblées de neurones différentes dont les éléments reçoivent un même signal entreraient concurrence pour l'activation, et seules celles dont les temps de réponse sont les plus courts parviendraient finalement à s'activer, c'est à dire que seules les représentations mentales supportées par ces assemblées seraient finalement rappelées à l'esprit.

Il ne s'agit là bien sûr que d'une hypothèse, qui aurait besoin d'être étayée par des données d'observation. Il conviendrait notamment de vérifier si les schémas de connexion au sein du LMT et les mécanismes synaptiques en jeu sont compatibles avec un tel processus. Mais les connaissances actuelles sur le LMT sont trop fragmentaires et incertaines pour permettre ce genre de vérification, et tout ce que nous pouvons faire pour le moment, c'est de raisonner en termes de plausibilité. À cet égard, il faut souligner que l'activation et l'inhibition mutuelle entre neurones ainsi que la compétition entre assemblées de neurones sont des mécanismes très courants, ce qui rend le processus suggéré d'autant plus plausible.

Il resterait cependant à expliquer comment il se fait que seules les assemblées de neurones qui supportent les représentations mentales les plus spécifiques de l'agent participent à ce processus [38], et aussi comment les conclusions des inférences sont effectivement extraites des représentations finalement réactivées. Différentes hypothèses peuvent être faites relativement à ces deux questions, mais en l'absence de données anatomiques et fonctionnelles plus précises elles seraient purement spéculatives et donc de peu d'intérêt. C'est pourquoi on laissera ici ces questions en suspend.

38 Rappelons que selon notre analyse les inférences automatiques reposent seulement sur les représentations mentales les plus précises du sujet (*c.f.* p. 121 ci-dessus).

Si on quitte à présent le cas particulier des mammifères pour revenir à celui plus général des animaux dotés d'un cerveau, on se rappellera sans doute que dans toutes ces espèces l'information sensorielle est analysée en caractéristiques qui convergent ensuite vers des aires cérébrales centrales où elles sont intégrées, et aussi que dans toutes ces espèces, ces mêmes aires centrales sont impliquées dans l'apprentissage. Ces similitudes rendent très probable que, non seulement chez les mammifères mais plus généralement dans toutes les espèces dotés d'un cerveau, le savoir général sur lequel s'appuie la prise de décision est stocké dans des représentations mentales encodées sous la forme d'ensembles de caractéristiques co-occurrentes. Mais cela n'a peut-être rien de surprenant, si on considère que les caractéristiques et les ensembles de caractéristiques sont sans doute le moyen le plus économique de stocker des représentations d'objets et de situations dans un cerveau [39].

Par ailleurs, l'encodage d'ensembles de caractéristiques co-occurrentes par des neurones dédiés qui additionnent les signaux co-occurrents, comme c'est le cas des neurones du LMT chez les mammifères, est probablement aussi le plus simple et le plus économique qui soit. C'est pourquoi il ne serait pas non plus étonnant qu'une même organisation soit conservée entre les espèces, et que des neurones analogues aux neurones-concept du LMT existent chez toutes les espèces à cerveau, bien qu'ils n'aient pas encore été identifiés chez les espèces appartenant à d'autres groupes taxinomiques.

Enfin, les mécanismes neuronaux impliqués dans le processus suggéré ci-dessus sont eux aussi extrêmement simples et communs, et existent sans doute jusque dans les cerveaux les plus simples. C'est pourquoi il n'y a aucune raison de supposer qu'un tel processus ne puisse exister que chez les mammifères. Au contraire, sa grande simplicité suggère qu'il pourrait se retrouver de façon globalement similaire chez toutes les espèces à cerveaux, ce qui expliquerait l'omniprésence des inférences automatiques à travers le règne animal.

39 Une argumentation lumineuse en ce sens peut être trouvée dans [Srinivasan, 2006]. On peut aussi consulter le commentaire de Chittka et Niven sur Srinivasan ([Chittka and Niven, 2009] p. 1000).

1.2 Présentation informelle du dispositif logique

Dans ce qui précède, on a défendu l'idée selon laquelle les inférences automatiques forment le coeur du raisonnement chez les agents naturels, et on a décrit un processus simple qui pourrait plausiblement les supporter, d'abord au niveau mental de description, puis au niveau neuronal. Au vu de cette analyse, il semble que l'approche la plus commode pour modéliser le raisonnement chez les agents naturels soit de procéder par étapes, et de commencer par modéliser d'abord ce processus central, en laissant la modélisation des facultés cognitives additionnelles pour un travail ultérieur. C'est pourquoi dans ce mémoire on va s'intéresser exclusivement à la modélisation des inférences automatiques et des processus d'apprentissage qui leur sont associés.

Concrètement, nous allons considérer un agent cognitif non-spécifié (désigné dans ce qui suit par 'l'agent') dont on va supposer qu'il opère des inférences automatiques conformément au processus décrit ci-dessus, et nous allons mettre en place un dispositif logique visant à simuler ce processus. Par conséquent, il faudra garder à l'esprit que ce dispositif est conçu pour rendre compte des seules inférences automatiques de l'agent, à l'exclusion de toute autre faculté cognitive dont il pourrait jouir par ailleurs. En particulier, il n'aura pas vocation à modéliser d'autres sortes d'inférences plus sophistiquées, et notamment celles qui reposent de manière décisive sur la conscience, la réflexivité ou la représentation verbale. Mais lorsque les inférences automatiques apparaissent comme un sous-processus de ces dernières, il pourra bien sûr rendre compte de ce sous-processus particulier.

Nous appellerons *système inférentiel* la partie du cerveau de l'agent qui, selon notre hypothèse, réalise les inférences automatiques. C'est ce système inférentiel, ou plus exactement ses opérations, que nous allons tenter de modéliser.

Pour représenter l'information traitée par le système inférentiel de l'agent, nous allons utiliser un langage propositionnel dont les variables vont représenter les caractéristiques que l'agent est physiologiquement capable de percevoir — c'est à dire, les caractéristiques que son système perceptif est capable d'extraire de l'information sensorielle. Ces caractéristiques ne devront donc pas être regardées comme des propriétés des objets du monde extérieur, mais comme de l'information (au sens computationnel du terme) circulant dans le cerveau de l'agent. Nous garderons cependant à l'esprit que, du point de vue de l'agent (c'est à dire, comme il le vit), elles apparaissent comme des propriétés des objets du monde extérieur. C'est pourquoi nous pourrons parfois écrire des choses comme *'selon l'agent, l'objet satisfait la caractéristique f'*. Notons que le nombre fini de neurones que contient le cerveau de l'agent ne peut lui permettre de discriminer, et donc de percevoir, qu'un nombre fini de caractéristiques, et que donc ce langage sera nécessairement fini. Par souci de simplicité, on supposera que l'ensemble des caractéristiques que l'agent est physiologiquement capable de percevoir ne varie pas au cours du temps, si bien que le langage sera fixé une fois pour toutes. Comme il est d'usage, les variables propositionnelles seront écrites en lettres latines minuscules italiques p, q, r, etc., et les formules en lettres grecques minuscules italiques α, β, γ, etc.

On peut remarquer que l'organisation neuronale des canaux sensoriels telle que décrite plus haut ne permet aux agents que de percevoir des caractéristiques 'positives'. Par exemple, on peut percevoir une chose comme étant rouge, mais pas comme n'étant pas noire. Cependant 'rouge' et 'noire' sont des caractéristiques mutuellement exclusives, au sens où reconnaitre une chose comme rouge nous amène automatiquement à rejeter l'idée qu'elle soit noire. Autrement dit, cela nous amène à considérer cette chose comme non-noire. Sur le plan neuronal, cela est sans doute réalisé par inhibition mutuelle entre neurones, mais du point de vue computationnel cela revient à avoir de l'information négative comme 'non-noire' qui circule dans le système inférentiel de l'agent. C'est pourquoi nous allons considérer qu'il existe aussi des caractéris-

tiques négatives, que nous allons naturellement représenter par des littéraux négatifs du langage. Les caractéristiques elles-mêmes seront dénotées par des lettres latines minuscules obliques f, etc. On écrira parfois '*non-f*' pour dénoter la contrepartie négative de la caractéristique *f*. On remarquera qu'il est impossible pour un agent de concevoir un(e) objet/situation comme satisfaisant en même temps une caractéristique f et sa négation *non-f*.

On remarquera aussi que cette même organisation neuronale des canaux sensoriels fait que les agents ne peuvent jamais percevoir autre chose que des caractéristiques et des conjonctions de caractéristiques. En particulier, ils ne peuvent jamais percevoir de disjonctions de caractéristiques. Ceci est vrai même lorsque l'information disponible est ambiguë. Par exemple, un agent ne percevra jamais quelque chose comme étant 'bleu foncé ou noir' : il le percevra comme étant bleu foncé, ou comme étant noir, ou tout simplement comme étant foncé. Dans les cas extrêmes, sa perception pourra varier au cours du temps entre ces possibilités — ce phénomène est connu en psychologie cognitive sous le nom de *'perception multistable'* [40]. Si l'agent est capable de réflexivité, il pourra, par un mouvement réflexif sur lui-même, percevoir sa propre hésitation, et de là inférer que l'objet qu'il considère est 'bleu foncé ou noir'. Mais il s'agit là d'une reconstruction postérieure, et dans tous les cas l'agent ne percevra jamais l'objet comme étant 'bleu foncé ou noir'.

La représentation mentale des éléments d'une classe donnée d'objets ou

40 Ainsi que Sterzer et al. la définissent, *'La perception multistable se produit lorsque l'information sensorielle est ambiguë et consistante avec deux ou plus interprétations mutuellement exclusives. Lorsqu'aucun indice supplémentaire n'est disponible, qui permettrait à la perception de converger vers une interprétation unique, la perception alterne spontanément toutes les quelques secondes entre deux ('bistable') ou plus ('multistable') interprétations du même signal sensoriel'* ([Sterzer et al., 2009] p. 310). Le phénomène de perception multistable est connu depuis longtemps (voir [Schwartz et al., 2012] pour une revue historique), et existe dans chaque modalité sensorielle. Bien que ses bases neurales exactes fassent toujours l'objet d'intenses débats (qui dépassent largement le cadre de ce mémoire) il y a un consensus général autour de l'idée que *'une seule solution perceptuelle ne peut exister à la fois'* ([Leopold and Logothesis, 1999] p. 260).

de situations[41] dans l'esprit de l'agent sera conçue comme l'ensemble des caractéristiques positives et négatives qui, selon lui, sont satisfaites par ces objets/situations. On supposera pour simplifier que toutes les caractéristiques appartenant à cet ensemble contribuent également à la représentation mentale, autrement dit que cet ensemble n'est *pas* structuré. Sous cette hypothèse, une représentation mentale peut simplement être vue comme la conjonction des caractéristiques qui la composent. On supposera d'autre part que les représentations mentales de l'agent sont consistantes, c'est à dire qu'une caractéristique et sa négation ne peuvent pas participer ensemble à une même représentation. Cela nous permettra de figurer les représentations mentales par des mondes partiels. Plus précisément, si r est une représentation mentale dans l'esprit de l'agent, on représentera r par le monde partiel w tel que pour tout littéral λ du langage, w satisfait λ si et seulement si λ représente une caractéristique appartenant à r. On dénotera les représentations mentales à l'aide de lettres latines sans-sherif minuscules obliques a, b, c, r, etc. Étant donnée une représentation mentale r, on notera parfois w_r le monde partiel qui représente r dans le modèle.

On dira qu'une représentation mentale r *satisfait* une caractéristique f si et seulement si f appartient à l'ensemble de caractéristiques qui composent r, c'est à dire si et seulement si, selon l'agent, les objet(s)/situation(s) correspondant(s) satisfont f. Par exemple, la représentation mentale que l'agent a des moineaux (en supposant qu'il en a une) satisfera la caractéristique 'a un bec'

41 Comme on l'a vu, le processus de mémorisation compile généralement plusieurs expériences en une mémoire unique, si bien qu'une représentation mentale est en fait la représentation commune des éléments d'une classe d'objets ou de situations ; lorsqu'une représentation mentale est construite à partir d'une expérience unique, cette classe ne contient qu'un seul élément. Par ailleurs, les représentations mentales se ramenant à des ensembles de caractéristiques, elles ne distinguent pas les objets des situations dans lesquels ceux-ci apparaissent. Ainsi, une représentation d'oiseaux peut contenir de l'information relative au contexte dans lequel ceux-ci ont été observés. C'est pourquoi une représentation mentale peut être vue à la fois comme la représentation d'objet(s) et la représentation de situation(s) impliquant cet/ces objet(s). La capacité à distinguer objets et situations se produit sans doute à un niveau cognitif bien supérieur, sous réserve bien sûr qu'un tel niveau soit réalisé dans le cerveau de l'agent.

si et seulement si dans l'esprit de l'agent, les moineaux ont un bec [42]. Le fait qu'un monde partiel figurant une représentation mentale satisfasse un littéral λ correspondra donc au fait que, selon l'agent, l'/les objet(s) correspondant(s) satisfont la caractéristique représentée par λ. De même, on dira qu'une représentation mentale r satisfait le contenu d'information représenté par la formule α si et seulement si le monde partiel qui représente r satisfait α.

Connaitre une chose n'est rien d'autre qu'avoir une certaine représentation mentale de cette chose. Par conséquent, pour chaque classe d'objets/situations que l'agent connait, il y a une représentation mentale correspondante dans son esprit. Comme on l'a dit, du point de vue computationnel seules les plus spécifiques (précises) d'entre elles jouent un rôle effectif dans l'opération des inférences automatiques, si bien que nous pouvons restreindre notre intérêt à ces dernières. L'ensemble de ces représentations mentales les plus précises peut se représenter directement par l'ensemble $\mathcal{U}$ des mondes partiels w tel que w représente l'une des représentations mentales les plus précises de l'agent. De plus, puisque les représentations mentales sont supportées dans le cerveau de l'agent par des assemblées neurones dédiés, $\mathcal{U}$ peut aussi être vu comme représentant un ensemble d'assemblées de neurones : à savoir, l'ensemble des assemblées neurones-concept dans le cerveau de l'agent qui supportent ses concepts les plus spécifiques.

Le fait que nous parlions spontanément de nos représentations mentales comme étant 'plus' ou 'moins' prégnantes l'une part rapport à l'autre suggère que la différence de prégnance entre représentations mentales est en quelque sorte quantitative. Dans l'idéal, il faudrait vérifier cette intuition en étudiant

42 Afin de faciliter la lecture, dans les exemples on considèrera 'a un bec', 'vole', etc. comme des caractéristiques. Il s'agit là d'une simplification, les caractéristiques étant sans doute en réalité beaucoup plus proches de celles identifiées dans [Tsunoda et al., 2001] et [Tanaka, 2003], tandis que 'a un bec', 'vole', etc., sont sans doute des conjonctions d'un grand nombre de telles caractéristiques. Mais cette simplification est sans conséquence, puisqu'on considère les représentations mentales comme des conjonctions de caractéristiques et que la conjonction est associative.

les mécanismes neuronaux qui gouvernent les interactions entre neurones-concepts. Mais ceux-ci sont encore mal connus, et tout ce que nous pouvons faire pour l'instant c'est de nous fier à l'introspection. Nous allons donc supposer que que la prégnance des représentations mentales est effectivement quantifiable, c'est à dire qu'à chaque représentation mentale dans l'esprit de l'agent correspond un nombre réel qui est la mesure de sa prégnance. Sur cette base, on rendra compte de la différence de prégnance entre représentations mentales en équipant $\mathcal{U}$ d'une relation binaire $<$. Concrètement, si w et w' sont des mondes partiels appartenant à $\mathcal{U}$, le fait que w $<$ w' sera interprété comme le fait que la représentation mentale figurée par w est *plus* prégnante dans l'esprit de l'agent que celle figurée par w'.

On appellera la structure $\mathcal{M} = (\mathcal{U}, <)$ un *modèle à mondes partiels*. Si on interprète $\mathcal{U}$ comme l'ensemble des représentations mentales les plus précises de l'agent, alors $\mathcal{M}$ peut être vu comme figurant la représentation que l'agent se fait du monde — son 'monde propre' (*Umwelt*) comme l'appellent Jackson et Cross [43]. Mais si on interprète $\mathcal{U}$ comme l'ensemble des assemblées de neurones-concept qui supportent les représentations mentales les plus spécifiques de l'agent, alors $\mathcal{M}$ peut être vu comme figurant le système inférentiel de l'agent.

On dira que l'agent est *disposé à inférer* β *de* α si et seulement si il est dans une condition telle que la condition suivante est vérifiée : si il considérait le contenu d'information α (où α représente la totalité de l'information considérée par l'agent à cet instant), alors après quelques instants il en viendrait à considérer (entre autres) le contenu d'information β [44]. Selon le processus inférentiel suggéré dans la section 1.1.2 ci-dessus, l'agent est disposé à inférer β de α si et seulement si toutes ses représentations mentales les plus prégnantes parmi celles qui satisfont α satisfont aussi β. Au niveau neural, une disposi-

43 [Jackson and Cross, 2011], pp. 154–155

44 Cette définition est librement inspirée du livre de Leitgeb [Leitgeb, 2004]

tion à inférer peut être vue comme un arrangement particulier des connexions entre les neurones de son système inférentiel tel que, si une impulsion excitative appropriée était envoyée dans son système inférentiel spécifiquement aux neurones supportant des concepts qui satisfont α, alors après quelques instants seuls des neurones supportant des concepts qui satisfont aussi β seraient actifs [45]. Dans les cas où l'impulsion provient du système perceptif de l'agent, α sera toujours un littéral ou une conjonction de littéraux, puisque les agents naturels ne peuvent jamais percevoir que des caractéristiques ou des conjonctions de caractéristiques. Mais lorsque les inférences automatiques se produisent comme sous-processus d'un processus plus complexe, l'impulsion peut provenir d'autres zones de son cerveau et alors α peut en principe être n'importe quelle formule. C'est pourquoi nous travaillerons sur la totalité du langage, et autoriserons toute formule à être une prémisse ou une conclusion d'inférence automatique.

La disposition de l'agent à inférer β de α est naturellement représentée dans le dispositif logique par le fait que tous les éléments $<$-minimaux de $\mathcal{U}$ parmi ceux qui satisfont α satisfont aussi β. Le lecteur habitué aux logiques préférentielles aura reconnu ici un motif familier : $\mathcal{M}$ induit une relation d'inférence sur le langage exactement de la même manière que les modèles cumulatifs de Kraus, Lehmann and Magidor le font [46]. Dans notre dispositif, la relation d'inférence induite par $\mathcal{M} = (\mathcal{U}, <)$ représentera l'ensemble des dispositions de l'agent à inférer, c'est à dire son savoir général sur les choses. Sur le plan logique, notre tache principale sera de caractériser les relations d'inférences induites par les modèles à mondes partiels, c'est à dire d'identifier les règles logiques qui, d'après le modèle, structurent le savoir général de l'agent.

Il peut être utile de clarifier le statut du langage logique dans le dispositif. Comme on l'a dit, le langage logique est le langage dans lequel nous décri-

45 Voir p. 136 ci-dessus.

46 [Kraus et al., 1990], ci-après KLM.

vons l'information qui circule dans le système inférentiel de l'agent. Ceci n'est pas la même chose que de décrire ses croyances, puisque celles-ci résultent du traitement de l'information par le cerveau de l'agent, et que donc décrire ses croyances suppose d'adopter un point de vue extérieur sur l'agent — c'est ce que fait la logique modale standard. Ce que nous essayons de faire ici c'est au contraire de modéliser le processus computationnel lui-même, qui n'opère pas sur des croyances mais seulement sur de l'information.

Le langage logique ne doit pas non plus être vu comme le langage de l'agent, même au sens d'un éventuel 'langage de la pensée' à la Fodor. Le fait est que le calcul envisagé ne s'opère pas à un niveau linguistique et n'implique aucun langage, fut-il computationnel. Même si les assemblées de neurones-concept peuvent être vues comme des représentants symboliques des représentations mentales qu'elles supportent, il n'existe pas de représentations symboliques des opérateurs logiques : les opérations logiques sont réalisées directement sur les assemblées de neurones, sous la forme d'activation simultanée ou d'absence d'activation. Il s'ensuit que le calcul envisagé n'implique aucun langage chez l'agent, dont langage logique pourrait être réputé tenir lieu. Au contraire, le langage logique est le langage que nous utilisons pour décrire le traitement de l'information dans le cerveau de l'agent, à la manière dont un physicien peut utiliser un langage mathématique adapté pour décrire un phénomène physique.

Par conséquent, les règles logiques dont on montrera qu'elles caractérisent les relations d'inférences induites ne devront pas être vues comme des règles de raisonnement que l'agent applique (ce qui supposerait un langage interne), mais comme des règles auxquelles il obéit, tout comme les phénomènes physiques obéissent aux lois physiques sans rien en savoir.

Chapitre 2

Le dispositif logique

Nous passons maintenant à l'introduction formelle du dispositif logique esquissé ci-dessus. Par souci de généralité et afin de faciliter la comparaison avec les travaux d'autres auteurs, nous allons, dans ce chapitre ainsi que dans celui qui suit, considérer un langage qui n'est *pas* nécessairement fini. Nous ne nous restreindrons aux langages finis qu'à partir du chapitre 4.

Le plan de ce chapitre est le suivant : dans la section 2.1, on définit une notion de monde partiel, ainsi qu'une notion de conséquence monotone et une autre d'équivalence ; dans la section 2.2 on rappelle brièvement le formalisme introduit par Kraus, Lehmann et Magidor, et dans la section 2.3 on s'inspire de leur travail pour définir une notion de modèle à mondes partiels. On montre ensuite que les relations d'inférence induites par les modèles à mondes partiels *'smooth'* forment une sous-classe stricte de celle des relations cumulatives de KLM. On identifie un certain nombre de règles valides supplémentaires, mais celles-ci ne suffisent pas pour obtenir un ensemble de règles complet.

2.1 Mondes partiels, $\mathcal{U}$-conséquence et $\mathcal{U}$-équivalence

Soit $\mathcal{L}$ un langage propositionnel non nécessairement fini, et $Var(\mathcal{L})$ l'ensemble des variables propositionnelles de $\mathcal{L}$. Une *distribution de valeurs de vérité partielle* sur $Var(\mathcal{L})$ est une fonction partielle $\boldsymbol{w} : Var(\mathcal{L}) \longrightarrow \{0, 1\}$. Une *distribution de valeurs de vérité complète* sur $Var(\mathcal{L})$ est une fonction totale

$\boldsymbol{w}$' : $Var(\mathcal{L}) \longrightarrow \{0, 1\}$. En accord avec les définitions usuelles, on considère les fonctions totales comme des fonctions partielles ayant la propriété d'être partout définies, et donc les distributions de valeurs de vérité complètes comme des distributions de valeurs de vérité partielles ayant la propriété d'être définies pour toutes les variables du langage.

Chaque distribution de valeurs de vérité partielle $\boldsymbol{w}$ engendre un *monde partiel* w, et chaque distribution de valeurs de vérité complète $\boldsymbol{w'}$ engendre un *monde complet* w' (à l'exception de la distribution de valeurs de vérité qui est partout non-définie, dont on convient qu'elle n'engendre pas de monde). On note $\mathcal{W}_{\mathcal{L}}$ l'ensemble des mondes complets pour $\mathcal{L}$, et $\mathcal{W}_{\mathcal{L}}^{\text{P}}$ l'ensemble des mondes partiels pour $\mathcal{L}$ ($\mathcal{W}_{\mathcal{L}} \subseteq \mathcal{W}_{\mathcal{L}}^{\text{P}}$). Pour chaque w $\in \mathcal{W}_{\mathcal{L}}^{\text{P}}$, on définit :

. Var^{+}(w) est l'ensemble des variables propositionnelles p de $\mathcal{L}$ telles que $\boldsymbol{w}(p) = 1$;

. Var^{-}(w) est l'ensemble des variables propositionnelles p de $\mathcal{L}$ telles que $\boldsymbol{w}(p) = 0$;

. Lit(w) $= Var^{+}$(w) $\cup \{\neg p_i / p_i \in Var^{-}(\text{w})\}$ est l'ensemble des littéraux de $\mathcal{L}$.

Dans le cas particulier où $\mathcal{L}$ est fini, on utilise la paire (Var^{+}(w), Var^{-}(w)) pour dénoter w. Par exemple, si $Var(\mathcal{L}) = \{p, q, r\}$, la paire ($\{p\}$, $\{q\}$) dénote le monde partiel w tel que $\boldsymbol{w}(p) = 1$, $\boldsymbol{w}(q) = 0$ et $\boldsymbol{w}(r)$ n'est pas définie. De plus (et toujours dans le cas où $\mathcal{L}$ est fini), pour chaque w de $\mathcal{W}_{\mathcal{L}}^{\text{P}}$, on définit la $\mathcal{L}$-formule δ(w) :

$\delta(\text{w}) = \bigwedge \lambda / \lambda \in Lit(\text{w})$ [1]

On appelle δ(w) la *description* du monde partiel w.

La satisfaction des formules de $\mathcal{L}$ par un monde complet w est notée $\models$, et définie de la manière habituelle :

. si α est une variable propositionnelle, alors w $\models \alpha$ ssi $\boldsymbol{w}(\alpha) = 1$;

1 Cette formulation est librement inspirée de [Dubois, 2008] p. 9.

. si $\alpha = \neg\beta$, alors w $\models \alpha$ ssi w $\not\models \beta$;

. si $\alpha = \beta \wedge \gamma$, alors w $\models \alpha$ ssi w $\models \beta$ et w $\models \gamma$,

les autres connecteurs étant définis de la manière habituelle à partir de la négation et de la conjonction. On note $\vdash$ la relation de *conséquence classique* sur $\mathcal{L}$, et $\equiv$ la relation d'*équivalence classique*. Si T est un ensemble de formules de $\mathcal{L}$, on dit que w *satisfait* T et on écrit w $\models T$ si et seulement si pour toute formule $\alpha \in T$, w $\models \alpha$.

La *satisfaction des formules de $\mathcal{L}$ par un monde partiel* w est notée $\Vdash$, et définie à l'aide des supervaluations :

> Pour toute formule α de $\mathcal{L}$ et tout monde partiel w, w $\Vdash \alpha$ si et seulement si pour tout monde complet w' tel que w' $\models Lit$(w), w' $\models \alpha$.

Autrement dit, un monde partiel w satisfait une formule α si et seulement si quelle que soit la manière dont on 'complète' w pour obtenir un monde complet w', celui-ci satisfait α. On vérifie aisément que dans le cas particulier où w est lui-même un monde complet, w $\Vdash \alpha$ ssi w $\models \alpha$. Dans le cas particulier où $\mathcal{L}$ est fini, cette définition se ramène à :

> w $\Vdash \alpha$ ssi pour tout monde complet w' tel que w' $\models \delta$(w), w' $\models \alpha$, *i.e.*
>
> ssi δ(w) $\vdash \alpha$.

On remarque que :

. w $\Vdash \alpha \wedge \beta$ ssi (w $\Vdash \alpha$ and w $\Vdash \beta$) ;

. (w $\Vdash \alpha$ or w $\Vdash \beta$) implique w $\Vdash \alpha \vee \beta$, mais

. w $\Vdash \alpha \vee \beta$ *n'implique pas* (w $\Vdash \alpha$ or w $\Vdash \beta$)

Par exemple, soit w tel que $\boldsymbol{w}(p)$ n'est pas définie : alors w $\Vdash p \vee \neg p$, mais w $\not\Vdash p$ et w $\not\Vdash \neg p$.

Soit $\mathcal{U} \subseteq \mathcal{W}^{\text{P}}_{\mathcal{L}}$ un ensemble de mondes partiels. Pour tout ensemble T de formules de $\mathcal{L}$, on note $\mathcal{W}_u(T)$ l'ensemble (possiblement vide) de mondes partiels w $\in \mathcal{U}$ tels que pour toute formule $\alpha \in T$, w $\Vdash \alpha$. Si $T = \{\alpha\}$, on omet les parenthèses et on écrit simplement $\mathcal{W}_u(\alpha)$.

On définit sur $\mathcal{L}$ une relation de $\mathcal{U}$-*conséquence* $\|\!\sim_{\mathcal{U},\mathcal{L}}$:

$\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$ ssi $\mathcal{W}_u(\alpha) \subseteq \mathcal{W}_u(\beta)$.

Il est immédiat que $\|\!\sim_{\mathcal{U},\mathcal{L}}$ est transitive. Elle est aussi supra-classique (c'est à dire que $\alpha \vdash \beta$ implique $\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$) : supposons en effet que $\alpha \vdash \beta$, et soit w $\in \mathcal{W}_u(\alpha)$: tout monde complet w' qui satisfait Lit(w) satisfait aussi α, et donc satisfait aussi β, donc w $\|\!= \beta$, c'est à dire que w $\in \mathcal{W}_u(\beta)$. Par contre, $\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$ n'implique pas $\alpha \vdash \beta$: par exemple, si w $= (\{p,\ q\}, \varnothing)$ et $\mathcal{U} = \{\text{w}\}$, on a $p \|\!\sim_{\mathcal{U},\mathcal{L}} q$ bien que $p \nvdash q$. $\|\!\sim_{\mathcal{U},\mathcal{L}}$ ne satisfait pas non plus la contraposition (c'est à dire que $\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$ n'implique pas $\neg\beta \|\!\sim_{\mathcal{U},\mathcal{L}} \neg\alpha$) : par exemple, si w $= (\{p,\ q\}, \varnothing)$, w' $= (\varnothing, \{q\})$ et $\mathcal{U} = \{\text{w, w'}\}$, on a $p \|\!\sim_{\mathcal{U},\mathcal{L}} q$, mais $\neg q \not\|\!\sim_{\mathcal{U},\mathcal{L}} \neg p$, puisque w' $\|\!= \neg q$ et w' $\|\!\neq \neg p$. Enfin, pour toute formule α de $\mathcal{L}$, on a $\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \bot$ ssi $\mathcal{W}_u(\alpha) = \varnothing$.

On définit aussi sur $\mathcal{L}$ une relation de $\mathcal{U}$-*équivalence* $\cong_{\mathcal{U},\mathcal{L}}$:

$\alpha \cong_{\mathcal{U},\mathcal{L}} \beta$ ssi $\alpha \|\!\sim_{\mathcal{U},\mathcal{L}} \beta$ et $\beta \|\!\sim_{\mathcal{U},\mathcal{L}} \alpha$, c'est à dire

ssi $\mathcal{W}_u(\alpha) = \mathcal{W}_u(\beta)$

Par supra-classicalité de $\|\!\sim_{\mathcal{U},\mathcal{L}}$, $\alpha \equiv \beta$ implique $\alpha \cong_{\mathcal{U},\mathcal{L}} \beta$.

2.2 Rappel du formalisme de Kraus, Lehmann et Magidor

On rappelle les définitions des modèles cumulatifs et préférentiels et des relations d'inférences induites par ces modèles, telles que données par Kraus, Lehmann et Magidor dans [Kraus et al., 1990]. Comme il est d'usage, si $E' \subseteq E$ sont des ensembles et si $<$ est une relation binaire sur E, on dit que x est $<$-*minimal dans* E' ssi $x \in E'$ et pour tout y de E', $y \not< x$. On note $\mathcal{P}(E)$ l'ensemble des parties de E.

Étant donné un langage propositionnel $\mathcal{L}$, un *modèle cumulatif* est une structure $\mathcal{M} = < S, l, \prec >$, où $\mathcal{S}$ est un ensemble (intuitivement interprété comme un ensemble d'états), $l : \mathcal{S} \longrightarrow \mathcal{P}(\mathcal{W}_\mathcal{L}) - \{\varnothing\}$ est une fonction qui éti-

quette chaque état avec un ensemble (non-vide)[2] de mondes complets, et $\prec$ est une relation binaire sur $\mathcal{S}$ (dite 'relation de préférence'), qui satisfait la propriété *'smoothness'* (définie plus bas).

La *satisfaction d'une formule* α *de* $\mathcal{L}$ *par un état* s est notée $s \models\!\!\models \alpha$ et est définie par :

$s \models\!\!\models \alpha$ ssi pour tout w de $l(s)$, w $\models \alpha$.

La propriété *smoothness* se ramène à la condition suivante : pour tout s de $\mathcal{S}$ et toute formule α de $\mathcal{L}$, si $s \models\!\!\models \alpha$ et s n'est pas $\prec$-minimal dans $\{s \in \mathcal{S} / s \models\!\!\models \alpha\}$, alors il existe s' de $\mathcal{S}$ tel que $s' \prec s$ et s' est $\prec$-minimal dans $\{s \in \mathcal{S} / s \models\!\!\models \alpha\}$. Par souci de brièveté, lorsque s est $\prec$-minimal dans $\{s \in \mathcal{S} / s \models\!\!\models \alpha\}$, on dira que s est $\prec$-*minimal pour* α.

La *relation d'inférence* $\mid\!\sim_{\mathcal{M}}$ induite par $\mathcal{M}$ sur $\mathcal{L}$ est définie par :

$\alpha \mid\!\sim_{\mathcal{M}} \beta$ ssi tout s $\prec$-minimal pour α satisfait β.

Les auteurs montrent que toute relation d'inférence $\mid\!\sim_{\mathcal{M}}$ induite par un modèle cumulatif $\mathcal{M}$ satisfait les règles :

Reflexivity	$\alpha \mid\!\sim_{\mathcal{M}} \alpha$
Left Equivalence	si $\alpha \equiv \beta$ et $\alpha \mid\!\sim_{\mathcal{M}} \gamma$, alors $\beta \mid\!\sim_{\mathcal{M}} \gamma$
Right Weakening	si $\alpha \vdash \beta$ et $\gamma \mid\!\sim_{\mathcal{M}} \alpha$, alors $\gamma \mid\!\sim_{\mathcal{M}} \beta$
Cut	si $\alpha \wedge \beta \mid\!\sim_{\mathcal{M}} \gamma$ et $\alpha \mid\!\sim_{\mathcal{M}} \beta$, alors $\alpha \mid\!\sim_{\mathcal{M}} \gamma$
Cautious Monotony	si $\alpha \mid\!\sim_{\mathcal{M}} \beta$ et $\alpha \mid\!\sim_{\mathcal{M}} \gamma$, alors $\alpha \wedge \beta \mid\!\sim_{\mathcal{M}} \gamma$

2 La définition des modèles cumulatifs dans [Kraus et al., 1990] (déf. 5 p. 16) exclut explicitement la possibilité d' 'étiquettes vides', c'est à dire d'états s tels que l(s) = $\varnothing$. Cependant la preuve de complétude (p. 19, entre les lemmes 10 and 11) utilise un état étiqueté par l'ensemble vide (cet état est la classe de $\bot$). Il apparaît que la restriction à des étiquettes non-vides est en fait inutile, tout modèle cumulatif comportant des états étiquetés par le vide pouvant se ramener à un modèle équivalent qui n'en comporte pas (on n'en donnera pas la preuve ici). On peut donc corriger la définition 5 de KLM et autoriser les étiquettes vides. Mais on peut aussi garder la définition 5 telle qu'elle est et corriger plutôt la preuve de complétude donnée p. 19, en prenant $\mathcal{S} = (\mathcal{L}/\sim)$ – la classe d'équivalence de $\bot$ au lieu de $\mathcal{S} = (\mathcal{L}/\sim)$. Comme la construction indiquée par KLM implique que pour tout état s $\in (\mathcal{L}/\sim)$ – la classe d'équivalence de $\bot$', s $\prec$ la classe d'équivalence de $\bot$, on vérifie aisément que supprimer la classe de $\bot$ du modèle n'affecte en rien la relation d'inférence induite. Pour des raisons de commodité, c'est cette deuxième solution qu'on adopte ici.

et que réciproquement, toute relation d'inférence $\mid\!\sim$ satisfaisant cet ensemble de règles admet un modèle cumulatif. Cet ensemble de règles est nommé *système C*, et les relations définies sur le langage qui satisfont toutes les règles de C sont appelées des *relations d'inférence cumulatives*. Les auteurs donnent aussi un certain nombre de règles dérivées de C, parmi lesquelles :

And	si $\alpha \mid\!\sim_{\mathcal{M}} \beta$ et $\alpha \mid\!\sim_{\mathcal{M}} \gamma$, alors $\alpha \mid\!\sim_{\mathcal{M}} \beta \wedge \gamma$
Equivalence	si $\alpha \mid\!\sim_{\mathcal{M}} \beta$, $\beta \mid\!\sim_{\mathcal{M}} \alpha$ et $\alpha \mid\!\sim_{\mathcal{M}} \gamma$, alors $\beta \mid\!\sim_{\mathcal{M}} \gamma$

Par ailleurs on obtient immédiatement

Supra-classicality	si $\alpha \vdash \beta$, alors $\alpha \mid\!\sim_{\mathcal{M}} \beta$

à partir de *Reflexivity* et *Right Weakening*.

Un *modèle préférentiel* est un modèle cumulatif $\mathcal{M} = < S, l, \prec >$ dans lequel les états sont étiquetés par des singletons et $\prec$ est un ordre partiel strict. Les auteurs montrent qu'en plus des règles de C, la règle *Or* suivante est valide dans les modèles préférentiels :

Or	si $\alpha \mid\!\sim_{\mathcal{M}} \gamma$ et $\beta \mid\!\sim_{\mathcal{M}} \gamma$, alors $\alpha \vee \beta \mid\!\sim_{\mathcal{M}} \gamma$,

et que réciproquement, toute relation d'inférence sur $\mathcal{L}$ qui satisfait les règles de C $\cup \{Or\}$ admet un modèle préférentiel. L'ensemble de règles C $\cup \{Or\}$ est appelé *système P*, et les relations d'inférence qui satisfont toutes les règles de P sont appelées des *relations préférentielles*.

2.3 Modèles à modes partiels et relations d'inférences induites

Soit $\mathcal{L}$ comme défini ci-dessus, $\mathcal{U} \subseteq \mathcal{W}^{\mathrm{P}}_{\mathcal{L}}$ et $<$ une relation binaire sur $\mathcal{U}$. On appelle $\mathcal{M} = (\mathcal{U}, <)$ un *modèle à mondes partiels*.

Dans le cas particulier où $\mathcal{L}$ est fini, $\mathcal{W}^{\mathrm{P}}_{\mathcal{L}}$ est fini et les mondes partiels qui le composent le sont aussi. Par conséquent si $\mathcal{L}$ est fini, $\mathcal{U}$ est un ensemble fini de

mondes partiels finis. Les modèles à mondes partiels tels que $\mathcal{U}$ satisfait cette condition sont appelés *modèles à mondes partiels finis*.

On dit qu'un monde partiel w $\in \mathcal{U}$ est $<$-*minimal* pour α si w est $<$-minimal dans $\mathcal{W}_u(\alpha)$. On définit la relation d'inférence $\mid\!\sim_{\mathcal{M}}$ induite par $\mathcal{M}$ sur $\mathcal{L}$ par :

$\alpha \mid\!\sim_{\mathcal{M}} \beta$ ssi tout w $<$-minimal pour α satisfait β.

On dit que $<$ est $\mathcal{L}$-*smooth* si et seulement si pour tout w $\in \mathcal{U}$ et pour toute formule α de $\mathcal{L}$, la condition suivante est satisfaite :

si w $\|\!\models \alpha$ et w n'est pas $<$-minimal pour α, alors il existe w' $\in \mathcal{U}$ tel que w' $<$ w et w' est $<$-minimal pour α.

$\mathcal{L}$-smoothness est la transposition de la propriété *smoothness* habituelle dans le contexte des mondes partiels. Si $\mathcal{M} = (\mathcal{U}, <)$ est un modèle à mondes partiels tel que $<$ est $\mathcal{L}$-smooth, on dit que $\mathcal{M}$ est un *modèle à mondes partiels* $\mathcal{L}$-*smooth*.

Il s'avère que tout modèle à mondes partiels $\mathcal{L}$-smooth peut être vu comme un modèle cumulatif. En effet, soit $\mathcal{M} = (\mathcal{U}, <)$ un modèle à mondes partiels $\mathcal{L}$-smooth, et $\mid\!\sim_{\mathcal{M}}$ la relation d'inférence induite par $\mathcal{M}$. On peut construire un modèle cumulatif $\mathcal{M}' = < S, l, \prec >$ tel que $\mathcal{S} = \mathcal{U}$, $\prec = <$, et $\mid\!\sim_{\mathcal{M}'} = \mid\!\sim_{\mathcal{M}}$. Pour cela il suffit de définir l par : pour tout w $\in \mathcal{S}$, l(w) = {w' $\in \mathcal{W}_{\mathcal{L}}$ / w' $\models$ *Lit*(w)}. On note que l ainsi définie est injective, puisque $\mathcal{U}$ peut contenir au plus une copie de chaque mode partiel. On vérifie que pour tout w $\in \mathcal{S}$ et toute formule α de $\mathcal{L}$, on a w $\|\!\models \alpha$ ssi w $\models\!\!\models \alpha$:

w $\|\!\models \alpha$ ssi (par déf. de $\|\!\models$) pour tout w' $\in \mathcal{W}_{\mathcal{L}}$ t.q. w' $\models$ *Lit*(w), w' $\models \alpha$, *i.e.*

ssi pour tout w' $\in l$(w), w' $\models \alpha$, *i.e.*

ssi w $\models\!\!\models \alpha$.

Par conséquent, puisque $\prec = <$, w est $<$-minimal pour α ssi il est $\prec$-minimal minimal α. Donc, puisque $<$ est $\mathcal{L}$-smooth, $\prec$ satisfait la propriété *smoothness*, et donc $\mathcal{M}'$ est un modèle cumulatif. De plus, puisque pour toute formule α, {w $\in \mathcal{U}$/w est $<$-minimal pour α} = {w $\in \mathcal{S}$/w est $\prec$- minimal pour α}, on a $\mid\!\sim_{\mathcal{M}'} = \mid\!\sim_{\mathcal{M}}$.

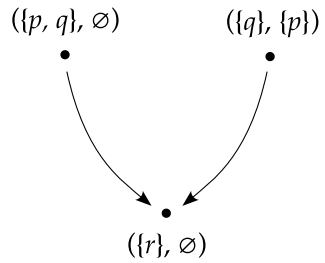

Figure 1 : $\mathcal{M}_1 = (\mathcal{U}_1, <_1)$

Il s'ensuit que si $\mathcal{M} = (\mathcal{U}, <)$ est un modèle à mondes partiels $\mathcal{L}$-smooth, alors $\mid\!\sim_{\mathcal{M}}$ est une relation d'inférence cumulative, c'est à dire que le système C de KLM est adéquat pour les modèles à mondes partiels $\mathcal{L}$-smooth.

Cependant C n'est pas complet pour ces mêmes modèles. Par exemple la règle ***(*)*** ci-dessous est valide dans les modèles à mondes partiels $\mathcal{L}$-smooth mais n'est pas dérivable de C :

(*) Pour toutes p et $q \in Var(\mathcal{L})$, ($p \mid\!\sim_{\mathcal{M}} \bot$ et $q \mid\!\sim_{\mathcal{M}} \bot$) ssi $p \vee q \mid\!\sim_{\mathcal{M}} \bot$

En effet, pour tout modèle à mondes partiels $\mathcal{L}$-smooth $\mathcal{M} = (\mathcal{U}, <)$ et toutes variables propositionnelles p et q de $\mathcal{L}$, on a $p \mid\!\sim_{\mathcal{M}} \bot$ et $q \mid\!\sim_{\mathcal{M}} \bot$ ssi $\mathcal{W}_u(p) = \mathcal{W}_u(q) = \varnothing$, ssi $\mathcal{W}_u(p \vee q) = \varnothing$, ssi $p \vee q \mid\!\sim_{\mathcal{M}} \bot$. Que ***(*)*** ne peut être dérivée de C suit de ce qu'elle n'est pas valide dans les modèles cumulatifs. Par exemple le modèle cumulatif $\mathcal{M}' = < S, l, \prec >$ tel que $\mathcal{S} = \{\text{s}\}$, $l(\text{s}) = \{ (\{p\}, \{q\}), (\{q\}, \{p\}) \}$, et $\prec = \varnothing$ induit une relation d'inférence $\mid\!\sim_{\mathcal{M}'}$ telle que $p \mid\!\sim_{\mathcal{M}'} \bot$, $q \mid\!\sim_{\mathcal{M}'} \bot$ et $p \vee q \not\mid\!\sim_{\mathcal{M}'} \bot$ (puisque $\text{s} \not\models p$ et $\text{s} \not\models q$, mais $\text{s} \models p \vee q$).

Pour autant la règle *Or* n'est pas valide dans les modèles à mondes partiels $\mathcal{L}$-smooths. Par exemple, soit $\mathcal{M}_1$ le modèle $(\mathcal{U}_1, <_1)$, tel que $\mathcal{U}_1 = \{(\{p, q\}, \varnothing), (\{q\}, \{p\}), (\{r\}, \varnothing)\}$, et $<_1 = \{((\{r\}, \varnothing), (\{p, q\}, \varnothing)), ((\{r\}, \varnothing), (\{q\}, \{p\}))\}$, ce qui peut être représenté graphiquement comme dans la Figure 1 : on a $p \mid\!\sim_{\mathcal{M}} q$ et $\neg p \mid\!\sim_{\mathcal{M}} q$, mais $p \vee \neg p \not\mid\!\sim_{\mathcal{M}} q$, car $(\{r\}, \varnothing)$ est $<$-minimal pour $p \vee \neg p$.

La question se pose des règles qu'il faudrait ajouter à C pour obtenir un ensemble de règles adéquat et complet pour les modèles à mondes partiels

$\mathcal{L}$-smooths. Malheureusement, aucun ensemble complet de règles n'a pu être trouvé jusqu'ici, et il y a des raisons de penser qu'il est impossible de parvenir à la complétude, du moins tant qu'on travaille dans un langage propositionnel standard. Ces raisons seront examinées dans le prochain chapitre. Auparavant on mentionne quelques règles qui, même si elles ne nous permettent pas d'atteindre la complétude, sont utiles dans ce contexte. Les règles suivantes sont valides dans les modèles à mondes partiels $\mathcal{L}$-smooths :

$\mathcal{U}$-Consistence	si $\alpha \mid\!\sim_{\mathcal{M}} \bot$, alors $\alpha \parallel\!\!\sim_{\mathcal{U},\mathcal{L}} \bot$
$\mathcal{U}$-Left Equivalence	si $\alpha \cong_{\mathcal{U},\mathcal{L}} \beta$ et $\alpha \mid\!\sim_{\mathcal{M}} \gamma$, alors $\beta \mid\!\sim_{\mathcal{M}} \gamma$
$\mathcal{U}$-Right Weakening	si $\alpha \parallel\!\!\sim_{\mathcal{U},\mathcal{L}} \beta$ et $\gamma \mid\!\sim_{\mathcal{M}} \alpha$, alors $\gamma \mid\!\sim_{\mathcal{M}} \beta$

La preuve de $\mathcal{U}$-*Consistence* est immédiate en utilisant $\mathcal{L}$-smoothness et les définitions de $\mid\!\sim_{\mathcal{M}}$ et $\parallel\!\!\sim_{\mathcal{U},\mathcal{L}}$; $\mathcal{U}$-*Left Equivalence* découle du fait que $\alpha \cong_{\mathcal{U},\mathcal{L}} \beta$ ssi $\mathcal{W}_u(\alpha) = \mathcal{W}_u(\beta)$, et $\mathcal{U}$-*Right Weakening* découle du fait que $\alpha \parallel\!\!\sim_{\mathcal{U},\mathcal{L}} \beta$ ssi $\mathcal{W}_u(\alpha) \subseteq \mathcal{W}_u(\beta)$. Puisque $\parallel\!\!\sim_{\mathcal{U},\mathcal{L}}$ est supra-classique, $\mathcal{U}$-*Left Equivalence* et $\mathcal{U}$-*Right Weakening* sont légèrement plus fortes que les règles originales *Left Equivalence* et *Right Weakening*, qu'elles peuvent remplacer. De même, la version de *Supra-classicality* pour les ensembles de mondes partiels,

Supra-$\mathcal{U}$-Consequence	si $\alpha \parallel\!\!\sim_{\mathcal{U},\mathcal{L}} \beta$, alors $\alpha \mid\!\sim_{\mathcal{M}} \beta$

qui peut être dérivée de *Reflexivity* et $\mathcal{U}$-*Right Weakening*, est légèrement plus forte que la règle *Supra-classicality* originale.

Chapitre 3

Le problème de la complétude dans le contexte des modèles à mondes partiels

Pour étudier le problème de la complétude dans le contexte des modèles à mondes partiels, nous allons nous inspirer de deux articles de D. M. Gabbay et K. Schlechta, à savoir [Gabbay and Schlechta, 2008] et [Gabbay and Schlechta, 2009]. Bien que la discussion qui suit ne constitue pas à proprement parler une preuve, elle donne néanmoins de bonnes raisons de penser qu'aucun ensemble de règles adéquat et complet pour les relations d'inférence induites sur $\mathcal{L}$ par les modèles à mondes partiels ne peut être trouvé.

Dans ces articles, les auteurs utilisent un langage propositionnel (non-nécessairement fini) $\mathcal{L}$. $\mathcal{M}_\mathcal{L}$ est l'ensemble des modèles classiques de $\mathcal{L}$, c'est à dire, selon notre terminologie, des mondes complets pour $\mathcal{L}$ (c'est à dire que l'ensemble $\mathcal{M}_\mathcal{L}$ de Gabbay et Schlechta est notre ensemble $\mathcal{W}_\mathcal{L}$). Pour tout ensemble T de formules de $\mathcal{L}$, $M_{(T)} \subseteq \mathcal{M}_\mathcal{L}$ est l'ensemble de modèles de T. $D_\mathcal{L}$ est l'ensemble des sous-ensembles définissables de $\mathcal{M}_\mathcal{L}$, c'est à dire que $D_\mathcal{L} = \{M_{(T)} / T$ est un ensemble de $\mathcal{L}$-formules$\}$. Une *structure préférentielle* est une paire $\mathcal{M} = < U, \prec >$, où U est une ensemble et $\prec$ est une relation binaire sur U. Étant donné un ensemble $\mathcal{Y}$ d'ensembles, une fonction μ de domaine $\mathcal{Y}$ est définie. Dans le cas général où U peut contenir plusieurs 'copies' de certains éléments (c'est à dire que U est en fait un en-

semble de paires $< x\,,\ i >$, avec i un index), μ est définie par : pour tout $X \in \mathcal{Y}$, $\mu(X) = \{x \in X \,/\, \exists < x\,,\ i > \in U$ et $\neg\exists < x\text{'}, j > \in U$ t.q. ($x\text{'} \in X$ et $< x\text{'}, j > \prec < x\,,\ i >)\}$. Dans le cas particulier où U contient au plus une copie de chaque élément et où on peut donc abandonner les index, $\mu(X)$ est simplement définie par $\{x/x$ is $\prec$-minimal in $X \cap U\}$. L'interprétation visée de μ est celle d'une fonction de minimalisation (remarquons cependant que $\mu(X)$ peut être vide, notamment dans le cas où $X \cap \{x \,/\exists\ i$ t.q.$< x\,,\ i > \in U\}$ l'est). Lorsque $\mathcal{Y} \subseteq \mathcal{P}(\mathcal{M}_\mathcal{L})$, μ engendre une relation d'inférence $\mid\!\sim$ de la manière habituelle : pour tout ensemble de $\mathcal{L}$-formules T et toute $\mathcal{L}$-formule α, $T \mid\!\sim \alpha$ si et seulement si $\mu(M_{(T)}) \subseteq M_{(\{\alpha\})}$. Les structures préférentielles peuvent donc être vues comme une généralisation des modèles préférentiels de KLM. Les auteurs étudient les propriétés algébriques de ces structures, et, dans les cas où $\mathcal{Y} \subseteq \mathcal{P}(\mathcal{M}_\mathcal{L})$, la correspondance entre ces propriétés et les règles d'inférences habituelles (par exemple, la règle *Reflexivity* correspond au fait que pour tout $X \in (\mathcal{Y} \cap D_\mathcal{L})$, $\mu(X) \subseteq X$). Parmi ces propriétés, les propriétés de clôture du domaine de μ se révèlent particulièrement importantes. En particulier, les auteurs montrent dans [Gabbay and Schlechta, 2008] que lorsque la propriété *smoothness* n'est pas satisfaite et que le domaine de μ n'est pas clos sous les unions finies, la contrepartie algébrique de la cumulativité se brise en une infinité de propriétés non-équivalentes, ce qui compromet les chances de parvenir à un théorème de représentation.

On ne peut bien sûr pas utiliser directement ces résultats pour nos besoins propres, puisque les modèles que nous considérons satisfont *smoothness*, et sont de plus construits à partir de mondes partiels. Mais on peut s'inspirer de leur approche, et notamment regarder de près les problèmes de définissabilité et de propriétés de clôture. Pour cela il nous faut d'abord accommoder leurs outils au contexte des mondes partiels, et en premier lieu nous donner une notion de définissabilité appropriée, ce qu'on va faire comme suit :

Si $\mathcal{L}$ est un langage propositionnel (non nécessairement fini), $\mathcal{U} \subseteq \mathcal{W}^{\text{P}}_\mathcal{L}$ et $A \subseteq \mathcal{U}$, on dira que A est *$\mathcal{L}$-définissable dans $\mathcal{U}$* si et seulement si il existe une

$\mathcal{L}$-formule α telle que $A = \mathcal{W}_u(\alpha)$. Par exemple, { ($\{p, q\}, \varnothing$), ($\{p\}, \{q\}$) } est $\mathcal{L}$-définissable dans $\mathcal{U}$ = { ($\{p, q\}, \varnothing$), ($\{p\}, \{q\}$), ($\{r\}, \varnothing$) }, mais pas dans $\mathcal{U}'$ = { ($\{p, q\}, \varnothing$), ($\{p\}, \{q\}$), ($\{p, r\}, \varnothing$) }, puisque pour toute $\mathcal{L}$-formule α telle que ($\{p, q\}, \varnothing$) $\|\!=\alpha$ et ($\{p\}, \{q\}$) $\|\!=\alpha$, on a aussi ($\{p, r\}, \varnothing$) $\|\!=\alpha$.

On notera $D_{\mathcal{U},\mathcal{L}}$ l'ensemble des sous-ensembles de $\mathcal{U}$ qui sont $\mathcal{L}$-définissables dans $\mathcal{U}$. Autrement dit, $D_{\mathcal{U},\mathcal{L}} = \{A \subseteq \mathcal{U} \,/\, \exists\, \alpha \in \mathcal{L}$ tel que $A = \mathcal{W}_u(\alpha)\}$. Comme on ne s'intéressera toujours qu'à la définissabilité d'un sous-ensemble de $\mathcal{U}$ dans $\mathcal{U}$ lui-même, par souci de brièveté on parlera de *sous-ensembles $\mathcal{L}$-définissables de $\mathcal{U}$*. $D_{\mathcal{U},\mathcal{L}}$ est clos sous les intersections finies, puisque pour toutes $\mathcal{L}$-formules α et β, $\mathcal{W}_u(\alpha) \cap \mathcal{W}_u(\beta) = \mathcal{W}_u(\alpha \wedge \beta)$. Mais $D_{\mathcal{U},\mathcal{L}}$ n'est généralement *pas* clos sous les unions finies. Par exemple, si $\mathcal{U} = \{(\{p, q\}, \varnothing), (\{p\}, \{q\}), (\{p, r\}, \varnothing)\}$, $\mathcal{W}_u(p \wedge q) \cup \mathcal{W}_u(p \wedge \neg q) = \{(\{p, q\}, \varnothing), (\{p\}, \{q\})\} \notin D_{\mathcal{U},\mathcal{L}}$.

Il est bien connu qu'un modèle préférentiel $< S, l, \prec >$ peut être vu comme une paire $(U, \prec')$, où $U \subseteq (\mathcal{W}_{\mathcal{L}} \times I)$ (avec I un ensemble d'index) et $\prec'$ est une relation de préférence sur U : il suffit de remplacer chaque état s de S par une copie judicieusement indexée w_i du monde complet w tel que $\mathrm{w} = l(s)$, et de poser que $\prec' = \prec$. De même, un modèle a cumulatif $\mathcal{M} = < S, l, \prec >$ peut être vu comme une paire $(V, \prec'')$, où $V \subseteq \mathcal{P}(\mathcal{W}_{\mathcal{L}})$ et $\prec''$ est une relation de préférence sur V. Il suffit pour cela de considérer la relation cumulative $\mid\!\sim_{\mathcal{M}}$ induite par $\mathcal{M}$ et de suivre les instructions données par KLM dans leur théorème de représentation [1] pour construire le modèle cumulatif $\mathcal{M}' = < S', l', \prec' >$ de $\mid\!\sim_{\mathcal{M}}$ correspondant. Ensuite, puisque par construction l' est injective, il suffit de remplacer chaque état s de S' par $l'(s)$, et de poser que $\prec'' = \prec'$ pour obtenir $(V, \prec'')$.

Il faut aussi rappeler que la demande de KLM que $\prec$ soit un ordre partiel strict dans les modèles préférentiels ne vise qu'à la praticité mais n'ajoute rien du point de vue logique, si bien que dans les modèles préférentiels comme dans les modèles cumulatifs, tout ce qui est finalement demandé pour $\prec$ c'est qu'elle

1 [Kraus et al., 1990], section 3.5.

satisfasse *smoothness*. Il s'ensuit que ce qui différentie essentiellement modèles préférentiels et cumulatifs, c'est que dans les premiers la relation de préférence peut être vue comme étant définie sur un ensemble de mondes complets (indexés), tandis que dans les seconds elle peut être vue comme étant définie sur un ensemble *d'ensembles de* mondes complets.

Enfin, on sait que du côté logique cette différence correspond à l'addition de la règle *Or* à C. *Or* peut donc être vue comme permettant l'existence de modèles dans lesquels $\prec$ est définie sur un ensemble de mondes.

Or, une caractéristique essentielle des modèles à mondes partiels est précisément que la relation $<$ y est définie entre mondes (partiels) et non pas entre ensembles de mondes (partiels). Un ensemble de règles complet pour les relations d'inférences induites par ces modèles devrait donc contenir une ou plusieurs règles correspondant à cette propriété. Mais comme on l'a vu, *Or* n'est pas valide dans les modèles à mondes partiels.

L'approche algébrique de Gabbay et Schlechta permet de comprendre comment la règle *Or* produit son effet dans le contexte des relations préférentielles. Dans [Gabbay and Schlechta, 2009], ils montrent que la contrepartie algébrique de la règle *Or* est la propriété (μOr) :

(μOr) Pour tous X et Y du domaine, $\mu(X \cup Y) \subseteq \mu(X) \cup \mu(Y)$

où μ est la fonction de minimalisation. Plus précisément, ils montrent que pour toute relation $\mid\sim$ sur $\mathcal{L}$ satisfaisant *Left Equivalence* et *Right Weakening*, la fonction μ de domaine $D_{\mathcal{L}}$ définie par : $\mu(M_{(T)}) = M_{(\bar{\bar{T}})}$ (où $\bar{\bar{T}} = \{\mathcal{L}$-formules α / $T \mid\sim \alpha\}$) satisfait (μOr) si et seulement si $\mid\sim$ satisfait *Or* [2].

Par ailleurs, on montre aisément que pour tout ensemble Z de paires $< x , i >$, toute relation binaire $\prec$ sur Z et tout ensemble $\mathcal{Y}$ d'ensembles, si μ est une fonction de domaine $\mathcal{Y}$ définie par : pour tout $X \in \mathcal{Y}$, $\mu(X) = \{x \in X$ / $\exists < x , i > \in Z$ et $\neg\exists < x', j > \in Z$ t.q. ($x' \in X$ et $< x', j > \prec < x , i >$)$\}$, alors μ satisfait (μOr) pourvu que $\mathcal{Y}$ soit clos sous unions finies. En effet,

2 [Gabbay and Schlechta, 2009], proposition 3.8.

soit X et Y des éléments de $\mathcal{Y}$; supposons qu'il existe $x \in \mu(X \cup Y)$ tel que $x \notin \mu(X) \cup \mu(Y)$. Par la définition de μ, $\mu(X \cup Y) \subseteq X \cup Y$, donc $x \in X$ ou $x \in Y$. Si $x \in X$, alors puisque $x \notin \mu(X)$, par la définition de μ encore il existe une paire $<x', j>$ appartenant à Z telle que $x' \in X$ et $<x', j> \prec <x, i>$. $x' \in X \cup Y$, donc $x \notin \mu(X \cup Y)$, ce qui est absurde. Et de même si $x \in Y$. Il s'ensuit que pour tout ensemble de paires Z, tout ensemble $\mathcal{Y}$ d'ensembles et toute fonction μ de domaine $\mathcal{Y}$, si μ *falsifie* (μOr), alors il ne peut exister de relation binaire $\prec$ sur Z telle que pour tout $X \in \mathcal{Y}$, $\mu(X) = \{x \in X / \exists <x, i> \in Z$ et $\neg\exists <x', j> \in Z$ t.q. $(x' \in X$ et $<x', j> \prec <x, i>)\}$. Autrement dit, cela signifie que si μ falsifie (μOr), alors il ne peut pas exister de relation $\prec$ sur Z telle que μ est la fonction de minimalisation induite par $\prec$. Remarquons que dans tous les cas la clôture de $\mathcal{Y}$ sous les unions finies est requise, car sinon $\mu(X \cup Y)$ pourrait ne pas être définie, auquel cas la définition de (μOr) est dénuée de sens. Ces résultats peuvent être facilement transposés au cas particulier où Z contient au plus une copie de chaque élément, et dans aussi au cas encore plus particulier où Z est un ensemble de mondes partiels.

Dans le cas particulier des modèles à mondes partiels, Z est $\mathcal{U}$ et $\prec$ est $<$, et donc par le résultat ci-dessus on a bien que pour toutes $\mathcal{L}$-formules α et β, {w/w est $<$-minimal dans $\mathcal{W}_u(\alpha) \cup \mathcal{W}_u(\beta)$} $\subseteq$ {w/w est $<$-minimal dans $\mathcal{W}_u(\alpha)$} $\cup$ {w/w est $<$-minimal dans $\mathcal{W}_u(\beta)$}. Mais parce que $\mid\!\sim_M$ ne peut saisir que les sous-ensembles $\mathcal{L}$-définissables de $\mathcal{U}$, il semble que la relation $\mid\!\sim_M$ est incapable d'exprimer cette propriété. En effet, toute fonction μ définie au moyen de $\mid\!\sim_M$ (en prenant $\mu(\mathcal{W}_u(\alpha)) = \mathcal{W}_u(\overline{\overline{\{\alpha\}}})$ avec $\overline{\overline{\{\alpha\}}} = \{\beta / \alpha \mid\!\sim_M \beta\}$, ou tout autre définition qu'on pourrait vouloir) aura toujours $D_{\mathcal{U},\mathcal{L}}$ pour domaine. Mais $D_{\mathcal{U},\mathcal{L}}$ n'est pas nécessairement clos sous unions finies, et dans les cas où il ne l'est pas (μOr) ne sera pas définie sur la totalité de $D_{\mathcal{U},\mathcal{L}}$ mais seulement sur les sous-ensembles de $D_{\mathcal{U},\mathcal{L}}$ qui sont clos sous unions finies. On peut supposer qu'une fonction μ judicieusement choisie va bien satisfaire (μOr) à l'intérieur de chacun de ces sous-ensembles, mais il y aura toujours plusieurs manières d'étendre μ de manière à clore $D_{\mathcal{U},\mathcal{L}}$ sous unions finies, dont certaines qui ne

satisfont pas (μOr). En d'autres termes, μ ne va pas 'en dire suffisamment' pour exclure la possibilité qu'il n'existe aucune relation $<$ sur $\mathcal{U}$ telle que μ serait un fragment de la fonction de minimalisation induite par $<$. Il semble donc que quelques soient les règles valides qu'on pourrait trouver pour $\mid\!\sim_{\mathcal{M}}$, celles-ci ne seront pas suffisantes, dans le cas général, pour garantir l'existence d'une relation $<$ adéquate sur $\mathcal{U}$, et donc pour obtenir un théorème de représentation. Dans ces conditions, l'existence d'un ensemble de règles complet pour les relations d'inférence induites sur $\mathcal{L}$ par les modèles à mondes partiels semble extrêmement douteuse.

Chapitre 4

Extension du langage

L'absence de clôture de $D_{\mathcal{U},\mathcal{L}}$ sous les unions finies peut être vue comme la conséquence d'un manque d'expressivité de $\mathcal{L}$ relativement aux mondes partiels. En effet, $\mathcal{L}$ permet d'exprimer le fait qu'un monde partiel w satisfait $\alpha \vee \beta$, mais pas que : il satisfait α ou il satisfait β. Dans le contexte des mondes complets les deux sont équivalents, mais pas dans celui des mondes partiels. Une façon simple de clore $D_{\mathcal{U},\mathcal{L}}$ sous les unions finies serait donc d'ajouter un connecteur au langage, qui lui permettrait d'exprimer ce fait. C'est ce qu'on va faire dans ce qui suit. Pour des raisons de simplicité et aussi parce que seuls les langages finis nous seront utiles pour ce que nous avons à faire, nous allons à partir de maintenant nous restreindre aux langages finis.

Prenant pour base un langage propositionnel fini $\boldsymbol{L}$ (qui est donc un cas particulier du langage $\mathcal{L}$ utilisé jusqu'ici), dans la section 4.1 nous y ajoutons deux connecteurs binaires[1] $\|$ et $\curlywedge$ (on montrera plus loin que $\curlywedge$ n'apporte en fait rien à l'expressivité du nouveau langage, mais n'est ajouté que pour des raisons de praticité). Dans la section 4.2 on transpose les notions précédemment définies dans ce nouveau langage $\boldsymbol{L}^{\|}$, et dans la section 4.3 on donne

1 Un symbole $\|$ est aussi utilisé par Gabbay et Schlechta, avec un sens algébrique. L'usage que nous allons en faire ici est différent, et ne devra pas être confondu avec celui fait par ces auteurs (dans tous les cas le contexte d'usage suffit à marquer la différence, puisqu'ici nous utiliserons ce symbole comme connecteur logique).

un certain nombre de règles d'inférences valides pour les relations d'inférences induites sur $\boldsymbol{L}^{\shortparallel}$ par les modèles à mondes partiels $\boldsymbol{L}^{\shortparallel}$-smooth. Enfin, dans la section 4.4 on introduit quelques définitions supplémentaires qui nous seront utiles pour la suite.

4.1 Le langage $\boldsymbol{L}^{\shortparallel}$

Soit $\boldsymbol{L}$ un langage propositionnel *fini*. On définit le langage $\boldsymbol{L}^{\shortparallel}$ par :

. Si α est une $\boldsymbol{L}$-formule, alors α est une $\boldsymbol{L}^{\shortparallel}$-formule (*i.e.*, $\boldsymbol{L} \subseteq \boldsymbol{L}^{\shortparallel}$) ;

. Si α et β sont des $\boldsymbol{L}^{\shortparallel}$-formules, alors $\alpha \curlywedge \beta$ et $\alpha \parallel \beta$ sont des $\boldsymbol{L}^{\shortparallel}$-formules ;

. Rien d'autre n'est une $\boldsymbol{L}^{\shortparallel}$-formule.

La satisfaction d'une $\boldsymbol{L}^{\shortparallel}$-formule α par un monde partiel w est notée w $\Vdash \alpha$ comme auparavant. Elle est définie par les clauses suivantes :

. Si α est une $\boldsymbol{L}$-formule, alors w $\Vdash \alpha$ ssi $\delta(\mathrm{w}) \vdash \alpha$ (comme auparavant).

. Si $\alpha = \beta \parallel \gamma$, alors w $\Vdash \alpha$ ssi (w $\Vdash \beta$ ou w $\Vdash \gamma$) ;

. Si $\alpha = \beta \curlywedge \gamma$, alors w $\Vdash \alpha$ ssi (w $\Vdash \beta$ et w $\Vdash \gamma$).

Comme auparavant, on note $\mathcal{W}_u(\alpha)$ l'ensemble des mondes partiels w $\in \mathcal{U}$ tels que w $\Vdash \alpha$. Dans le cas particulier où w est un monde complet, on vérifie aisément que w $\Vdash \beta \parallel \gamma$ ssi w $\models \beta \vee \gamma$ ssi w $\Vdash \beta \vee \gamma$. Dans le cas général, w $\Vdash \beta \parallel \gamma$ implique w $\Vdash \beta \vee \gamma$ mais non l'inverse. Par ailleurs, $\parallel$ est associatif : pour toutes $\boldsymbol{L}^{\shortparallel}$-formules α, β et γ et tout monde partiel w, w $\Vdash \alpha \parallel (\beta \parallel \gamma)$ ssi w $\Vdash (\alpha \parallel \beta) \parallel \gamma$. Par conséquent dans ce qui suit on pourra omettre les parenthèses, et écrire simplement $\alpha \parallel \beta \parallel \gamma$. Si $X = \{\alpha_1, \alpha_2, \ldots \alpha_n\}$ est un ensemble de $\boldsymbol{L}^{\shortparallel}$-formules, on écrira parfois $\Big\Vert \alpha_i/(\alpha_i \in X)$ comme une abréviation de la $\boldsymbol{L}^{\shortparallel}$-formule $\alpha_1 \parallel \alpha_2 \ldots \parallel \alpha_n$. Enfin, une conséquence directe de la définition de $\parallel$ est que $\mathcal{W}_u(\alpha \parallel \beta) = \mathcal{W}_u(\alpha) \cup \mathcal{W}_u(\beta)$.

Pour ce qui est du nouveau connecteur $\curlywedge$, pour l'instant on se contentera de remarquer qu'il opère sur les $\boldsymbol{L}^{\shortparallel}$-formules exactement de la même manière que la conjonction habituelle $\wedge$ le fait sur les $\boldsymbol{L}$-formules, puisque pour toutes

$\boldsymbol{L}$-formules α et β et tout monde partiel w, w $\|\!\models \alpha \wedge \beta$ ssi w $\|\!\models \alpha$ et w $\|\!\models \beta$. $\curlywedge$ peut donc être vu comme une extension de la conjonction habituelle $\wedge$ aux formules qui contiennent $\|$.

On peut se demander si l'introduction du connecteur $\|$ se justifie par des raisons autres que purement techniques, c'est à dire si $\|$ possède une interprétation naturelle dans le modèle. Il s'avère que $\|$ est la contrepartie syntaxique d'une opération mentale couramment réalisée sur l'information par les agents naturels. Plus précisément, la disjonction-$\|$ est sans doute la forme première sous laquelle la disjonction apparaît chez les êtres vivants.

Le fait est que chaque fois qu'un agent considère différents cas ou options (comme par exemple, pour décider quel cas est le plus plausible, ou quelle action doit être entreprise), il forme une disjonction de ce genre. Être incertain de si un objet ou une situation donné(e) satisfait la conjonction de caractéristiques α ou la conjonction de caractéristiques β n'est rien d'autre que croire qu'il/elle satisfait au moins l'une des deux, c'est à dire, qu'il/elle satisfait $\alpha \parallel \beta$. Pour décider laquelle de ces deux possibilités est la plus probable, l'agent entre l'information $\gamma \curlywedge (\alpha \parallel \beta)$ dans son système inférentiel (où γ représente la conjonction des caractéristiques qu'il croit par ailleurs l'objet satisfaire), et vérifie le résultat. Il est clair que cette opération mentale ne peut être représentée au moyen de la disjonction habituelle $\vee$, puisque $\gamma \wedge (p \vee\neg\, p) \equiv \gamma \wedge (q \vee\neg\, q) \equiv \gamma$, tandis qu'être incertain de si un objet donné satisfait la caractéristique représentée par p plutôt que sa négation $\neg\, p$ n'est pas la même chose que d'être incertain de si il satisfait la caractéristique représentée par q plutôt que sa négation $\neg\, q$, et n'est pas non plus la même chose que de ne pas avoir de telles hésitations.

Il faut souligner que réaliser ce genre de disjonctions ne requiert aucune capacité cognitive particulière de la part de l'agent, puisque avoir à l'esprit l'idée qu'un objet donné satisfait $\alpha \parallel \beta$ est simplement ne pas avoir choisi entre deux représentations mentales concurrentes de cet objet, la première selon laquelle il satisfait α, et la seconde selon laquelle il satisfait β. Selon le modèle neuronal

proposé dans la section 1.1.2, les disjonctions-‖ sont directement représentées dans le système inférentiel de l'agent par des unions d'ensembles (d'assemblées) de neurones. Concrètement, si A est l'ensemble des neurones qui représentent information α dans le système inférentiel de l'agent (c'est à dire, si A est l'ensemble de ses neurones-concept qui supportent des représentations mentales d'objets qui, selon lui, satisfont α) et si de même B est l'ensemble de ses neurones-concept qui représentent l'information β, alors $\alpha \parallel \beta$ est simplement représenté par $A \cup B$. Au contraire, les disjonctions-$\vee$ ne peuvent pas se représenter par des opérations ensemblistes sur les assemblées de neurones-concepts, ce qui rend très probable qu'elles ne peuvent être implémentées qu'au niveau verbal de représentation, à condition bien sûr que l'agent considéré possède un tel niveau de représentation. De ce fait, les disjonctions-‖ semblent bien être la forme première de la disjonction chez les agents naturels, tandis que les disjonctions-$\vee$ apparaissent plutôt comme une construction secondaire, reposant sur des capacités cognitives supérieures. Cela rend l'usage du connecteur ‖ très naturel dans notre contexte.

Fait intéressant, les clauses de satisfaction définies ci-dessus sont similaires à celles données par Kripke pour ses modèles intuitionnistes[2]. Plus précisément, un monde partiel w $\in \mathcal{W}_L^p$ peut être simulé par un modèle de Kripke $\Phi : (Var(\boldsymbol{L}) \times \mathbf{K}) \longrightarrow \{\mathrm{T}, \mathrm{F}\}$ défini sur une structure ($\mathbf{G}$, $\mathbf{K}$, $\boldsymbol{R}$) telle que $\mathbf{G}$ = w, $\mathbf{K} = \{\mathrm{w}\} \cup \{\mathrm{w'} \in \mathcal{W}_L \ / \ \delta(\mathrm{w'}) \vdash \delta(\mathrm{w})\}$, $\boldsymbol{R}$ est la clôture réflexive de $\{(\mathrm{w}, \mathrm{w'}) \ / \ \mathrm{w'} \in \mathbf{K}\}$ et pour toute $p \in Var(\boldsymbol{L})$ et tout w' $\in \mathbf{K}$, $\Phi(p, \mathrm{w'}) = \mathrm{T}$ ssi $\boldsymbol{w}'(p) = 1$, et $\Phi(p, \mathrm{w'}) = \mathrm{F}$ sinon. Il est en effet immédiat que pour toute $p \in Var(\boldsymbol{L})$, w $\Vdash p$ ssi $\Phi(p, \mathrm{w}) = \mathrm{T}$, et en utilisant les clauses de Kripke pour étendre Φ aux formules, on obtient que si α est une $\boldsymbol{L}$-formule ne contenant pas d'autres connecteurs que $\neg$ et $\wedge$, alors w $\Vdash \alpha$ ssi $\Phi(\alpha, \mathrm{w}) = \mathrm{T}$ (la preuve se fait par récurrence sur la construction des formules). En utilisant ensuite la clause

2 Voir [Kripke, 1965] pp. 94 – 100. Cette similarité m'a été suggérée par K. Schlechta dans une communication personnelle.

de satisfaction de $\|$ d'une part, et la clause de Kripke pour $\vee$ d'autre part, on obtient que si β est une $\boldsymbol{L}$-formule ne contenant pas d'autres connecteurs que $\neg$ et $\wedge$, alors w $\|\!\models \alpha \parallel \beta$ ssi $\Phi(\alpha \vee \beta, \mathrm{w}) = \mathrm{T}$. La disjonction-$\|$ peut donc être vue, en un certain sens, comme 'intuitionniste'. Cependant, et contrairement à ce qui est le cas général dans les modèles de Kripke, ici on a bien $\Phi(\alpha, \mathrm{w}) = \mathrm{T}$ ssi $\Phi(\neg\neg\alpha, \mathrm{w}) = \mathrm{T}$, c'est à dire que w $\|\!\models \alpha$ ssi w $\|\!\models \neg\neg\alpha$.

4.2 Transposition des définitions précédentes dans le contexte du langage $\boldsymbol{L}^{\|}$

Nous allons à présent transposer les notions précédemment définies — $\mathcal{U}$-conséquence, définissabilité, *smoothness*, relations d'inférence etc. — dans le contexte du langage $\boldsymbol{L}^{\|}$.

Soit $\mathcal{U} \subseteq \mathcal{W}^{\mathrm{p}}_{\boldsymbol{L}}$. Puisque $\boldsymbol{L}$ est fini, $\mathcal{U}$ est un ensemble fini de mondes partiels finis. On étend la relation de $\mathcal{U}$-*conséquence* à $\boldsymbol{L}^{\|}$:

Pour toutes $\boldsymbol{L}^{\|}$-formules α et β, $\alpha \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \beta$ ssi $\mathcal{W}_u(\alpha) \subseteq \mathcal{W}_u(\beta)$

Il est immédiat que $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$ est transitive, et que $\|\!\sim_{\mathcal{U},\boldsymbol{L}} \subseteq \|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$. On remarque que :

- $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$ est supra-classique, puisque $\|\!\sim_{\mathcal{U},\boldsymbol{L}}$ l'est.
- Pour toute $\boldsymbol{L}^{\|}$ formule α, $\alpha \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \bot$ ssi $\mathcal{W}_u(\alpha) = \varnothing$.
- Pour toutes $\boldsymbol{L}^{\|}$ formules α, β et γ, $\alpha \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \gamma$ et $\beta \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \gamma$ ssi $\alpha \parallel \beta \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \gamma$.
- Pour toutes $\boldsymbol{L}^{\|}$ formules α, β et γ, $\alpha \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \beta$ et $\alpha \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \gamma$ ssi $\alpha \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \beta \curlywedge \gamma$.
- Pour tout w $\in \mathcal{U}$ et toute $\boldsymbol{L}^{\|}$ formule α, $\delta(\mathrm{w}) \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \alpha$ ssi w $\|\!\models \alpha$ (voir preuve en Annexe A). Par conséquent, dans le cas particulier où α est une $\boldsymbol{L}$-formule, on a $\delta(\mathrm{w}) \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \alpha$ ssi w $\|\!\models \alpha$ ssi $\delta(\mathrm{w}) \vdash \alpha$.

De même, on étend la relation de $\mathcal{U}$-*équivalence* à $\boldsymbol{L}^{\|}$:

Pour toutes $\boldsymbol{L}^{\|}$-formules α et β, $\alpha \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \beta$ ssi $\alpha \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \beta$ et $\beta \,\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}\, \alpha$, c'est à dire, ssi $\mathcal{W}_u(\alpha) = \mathcal{W}_u(\beta)$.

Comme c'était déjà le cas pour $\cong_{\mathcal{U},\boldsymbol{L}}$, la supra-classicalité de $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$ fait que $\alpha \equiv \beta$ implique $\alpha \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \beta$. Par ailleurs, on vérifie aisément que :

. $(\alpha \parallel \beta) \parallel (\beta \parallel \gamma) \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \alpha \parallel \beta \parallel \gamma$.

. $(\alpha \parallel \beta \parallel \gamma) \curlywedge (\beta \parallel \gamma) \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \beta \parallel \gamma$.

. $\bot \parallel \alpha \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \alpha$.

. Si $\alpha \not\parallel_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \bot$, alors $\alpha \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \big\| \delta(\mathrm{w})/(\mathrm{w} \in \mathcal{W}_u(\alpha))$.

. Si $\alpha \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \alpha$', alors $\alpha \parallel \beta \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \alpha' \parallel \beta$ ('substitution des disjoints $\mathcal{U}$-équivalents').

. Si α et β sont des $\boldsymbol{L}$-formules, alors $\alpha \curlywedge \beta \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \alpha \wedge \beta$.

. $(\alpha_1 \parallel \alpha_2) \curlywedge \beta \cong_{\mathcal{U}, \boldsymbol{L}^{\parallel}} (\alpha_1 \curlywedge \beta) \parallel (\alpha_2 \curlywedge \beta)$.

On remarque que ces trois derniers faits pris ensemble entraînent que toute $\boldsymbol{L}^{\parallel}$-formule α qui contient $\curlywedge$ est $\mathcal{U}$-équivalente à une $\boldsymbol{L}^{\parallel}$-formule α' qui ne contient pas $\curlywedge$, mais contient la conjonction habituelle $\wedge$.

Étant donné $A \subseteq \mathcal{U}$, on dira que A est $\boldsymbol{L}^{\parallel}$-*définissable* (*dans* $\mathcal{U}$) ssi il existe une $\boldsymbol{L}^{\parallel}$-formule α telle que $A = \mathcal{W}_u(\alpha)$. On notera $D_{\mathcal{U}, \boldsymbol{L}^{\parallel}}$ l'ensemble des sous-ensembles de $\mathcal{U}$ qui sont $\boldsymbol{L}^{\parallel}$-définissables dans $\mathcal{U}$. Comme on ne s'intéressera toujours qu'à la définissabilité d'un sous-ensemble de $\mathcal{U}$ dans $\mathcal{U}$ lui-même, par souci de brièveté on parlera de sous-ensembles $\boldsymbol{L}^{\parallel}$-*définissables de* $\mathcal{U}$. Puisque $\boldsymbol{L} \subseteq \boldsymbol{L}^{\parallel}$, on a $D_{\mathcal{U}, \boldsymbol{L}} \subseteq D_{\mathcal{U}, \boldsymbol{L}^{\parallel}}$. Remarquons que bien que $\mathcal{U}$ soit fini, généralement $D_{\mathcal{U}, \boldsymbol{L}^{\parallel}} \neq \mathcal{P}(\mathcal{U})$. Par exemple, soit $\mathcal{U} = \{(\{p, q\}, \varnothing), (\{p\}, \varnothing)\}$; $\{(\{p\}, \varnothing)\} \notin D_{\mathcal{U}, \boldsymbol{L}^{\parallel}}$, puisque toute $\boldsymbol{L}^{\parallel}$ formule satisfaite par $(\{p\}, \varnothing)$ l'est aussi par $(\{p, q\}, \varnothing)$. Mais contrairement à $D_{\mathcal{U}, \boldsymbol{L}}$, $D_{\mathcal{U}, \boldsymbol{L}^{\parallel}}$ est clos sous les unions finies : si A et $B \in D_{\mathcal{U}, \boldsymbol{L}^{\parallel}}$, c'est à dire si il existe des $\boldsymbol{L}^{\parallel}$-formules α et β telles que $A = \mathcal{W}_u(\alpha)$ et $B = \mathcal{W}_u(\beta)$, alors $A \cup B = \mathcal{W}_u(\alpha) \cup \mathcal{W}_u(\beta) = \mathcal{W}_u(\alpha \parallel \beta)$, donc $A \cup B \in D_{\mathcal{U}, \boldsymbol{L}^{\parallel}}$.

De plus, puisque toute $\boldsymbol{L}^{\parallel}$-formule α qui contient $\curlywedge$ est $\mathcal{U}$-équivalente à une $\boldsymbol{L}^{\parallel}$-formule α' qui ne contient pas $\curlywedge$, tout sous-ensemble de $\mathcal{U}$ définissable par une $\boldsymbol{L}^{\parallel}$-formule contenant $\curlywedge$ l'est aussi par une $\boldsymbol{L}^{\parallel}$-formule qui ne contient pas $\curlywedge$. Par conséquent $\curlywedge$ n'ajoute rien à l'expressivité de $\boldsymbol{L}^{\parallel}$, et du point de vue logique il aurait été équivalent de définir $\boldsymbol{L}^{\parallel}$ par la seule addition de $\parallel$. On ne l'a pas fait pour des raisons de praticité, dans la mesure où cela va nous permettre

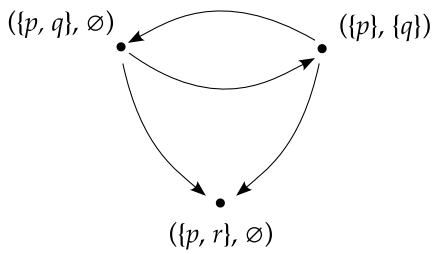

Figure 2 : $\mathcal{M}_2 = (\mathcal{U}_2, <_2)$

d'adapter à $\boldsymbol{L}^{\shortparallel}$ les règles habituelles pour les relations d'inférences. Cependant, puisque $\curlywedge$ n'est rien d'autre que l'extension de la conjonction usuelle $\wedge$ aux formules contenant $\|$, dans la suite on abandonnera l'usage du signe $\curlywedge$ pour dénoter les conjonctions de formules contenant $\|$, et on utilisera indifféremment $\wedge$ pour toutes les conjonctions de $\boldsymbol{L}^{\shortparallel}$ formules.

Soit $<$ une relation binaire sur $\mathcal{U}$. On notera $\|\!\sim_{\mathcal{M}}$ la relation d'inférence induite de la manière habituelle sur $\boldsymbol{L}^{\shortparallel}$ par $\mathcal{M} = (\mathcal{U}, <)$, c'est à dire :

> Pour toutes $\boldsymbol{L}^{\shortparallel}$-formules α et β, $\alpha \|\!\sim_{\mathcal{M}} \beta$ ssi tout w $<$-minimal dans $\mathcal{W}_u(\alpha)$ satisfait β.

On dira que $<$ est $\boldsymbol{L}^{\shortparallel}$-*smooth* si et seulement si pour tout w de $\mathcal{U}$ et toute $\boldsymbol{L}^{\shortparallel}$-formule α, la condition suivante est vérifiée :

> Si w $\|\!\models \alpha$ et w n'est pas $<$-minimal dans $\mathcal{W}_u(\alpha)$, alors il existe w' appartenant à $\mathcal{U}$ tel que w' $<$ w et w' est $<$-minimal dans $\mathcal{W}_u(\alpha)$.

Puisque $\boldsymbol{L} \subseteq \boldsymbol{L}^{\shortparallel}$, toute relation $\boldsymbol{L}^{\shortparallel}$-smooth est aussi $\boldsymbol{L}$-smooth, mais non l'inverse. Considérons par exemple le modèle $\mathcal{M}_2 = (\mathcal{U}_2, <_2)$ de la figure 2 : $<$ est $\boldsymbol{L}$-smooth, puisque toute $\boldsymbol{L}$-formule satisfaite à la fois par $(\{p,\ q\},\ \varnothing)$ et $(\{p\},\ \{q\})$ l'est aussi par $(\{p,\ r\},\ \varnothing)$. Mais $<$ n'est pas $\boldsymbol{L}^{\shortparallel}$-smooth, puisqu'il n'y a pas de monde $<$-minimal pour $(p \wedge q) \,\|\, (p \wedge \neg q)$.

Si $\mathcal{M} = (\mathcal{U}, <)$ est un modèle à monde partiels tel que $<$ est $\boldsymbol{L}^{\shortparallel}$-smooth, on appellera $\mathcal{M}$ un *modèle à mondes partiels* $\boldsymbol{L}^{\shortparallel}$*-smooth*. Tout modèle à monde partiels $\boldsymbol{L}^{\shortparallel}$-smooth est un modèle $\boldsymbol{L}$-smooth, mais non l'inverse.

4.3 Relations d'inférence induites sur $\boldsymbol{L}^{\shortparallel}$ par les modèles à mondes partiels $\boldsymbol{L}^{\shortparallel}$-smooth

On dira qu'une $\boldsymbol{L}^{\shortparallel}$-formule α est *maximale-consistante dans* $\mathcal{U}$ si

i) $\alpha \in \boldsymbol{L}$,

ii) $\alpha \not\shortparallel_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \bot$, et

iii) Pour toute $\boldsymbol{L}$-formule β telle que $\alpha \nvdash \beta$, $\alpha \wedge \beta \shortparallel_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \bot$.

Pour toute formule α maximale-consistante dans $\mathcal{U}$, il existe un monde partiel $\mathrm{w} \in \mathcal{U}$ tel que $\delta(\mathrm{w}) \equiv \alpha$. En effet, si α est maximale-consistante dans $\mathcal{U}$, alors par la clause ii) il existe $\mathrm{w} \in \mathcal{U}$ tel que $\mathrm{w} \Vdash \alpha$, ssi (par la première clause) $\delta(\mathrm{w}) \vdash \alpha$. Donc $\alpha \wedge \delta(\mathrm{w}) \equiv \delta(\mathrm{w}) \not\shortparallel_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \bot$, ce qui par la clause iii) implique que $\alpha \vdash \delta(\mathrm{w})$.

Soit $\mathcal{M} = (\mathcal{U}, <)$ un modèle à mondes partiels $\boldsymbol{L}^{\shortparallel}$-smooth. $\shortparallel\!\!\sim_{\mathcal{M}}$ satisfait les règles :

Reflexivity	$\alpha \shortparallel\!\!\sim_{\mathcal{M}} \alpha$
($\boldsymbol{L}^{\shortparallel}$) ***$\mathcal{U}$-Left Equivalence***	Si $\alpha \cong_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \beta$ et $\alpha \shortparallel\!\!\sim_{\mathcal{M}} \gamma$, alors $\beta \shortparallel\!\!\sim_{\mathcal{M}} \gamma$
($\boldsymbol{L}^{\shortparallel}$) ***$\mathcal{U}$-Right Weakening***	Si $\alpha \shortparallel_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \beta$ et $\gamma \shortparallel\!\!\sim_{\mathcal{M}} \alpha$, alors $\gamma \shortparallel\!\!\sim_{\mathcal{M}} \beta$
Cut	Si $\alpha \wedge \beta \shortparallel\!\!\sim_{\mathcal{M}} \gamma$ et $\alpha \shortparallel\!\!\sim_{\mathcal{M}} \beta$, alors $\alpha \shortparallel\!\!\sim_{\mathcal{M}} \gamma$
Cautious Monotony	Si $\alpha \shortparallel\!\!\sim_{\mathcal{M}} \beta$ et $\alpha \shortparallel\!\!\sim_{\mathcal{M}} \gamma$, alors $\alpha \wedge \beta \shortparallel\!\!\sim_{\mathcal{M}} \gamma$
($\boldsymbol{L}^{\shortparallel}$) ***$\mathcal{U}$-Consistence***	Si $\alpha \shortparallel\!\!\sim_{\mathcal{M}} \bot$, et $\alpha \shortparallel_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \bot$
$\shortparallel$-Or	Si $\alpha \shortparallel\!\!\sim_{\mathcal{M}} \gamma$ et $\beta \shortparallel\!\!\sim_{\mathcal{M}} \gamma$, alors $\alpha \shortparallel \beta \shortparallel\!\!\sim_{\mathcal{M}} \gamma$
Injectivity	Pour toute α maximale-consistante dans $\mathcal{U}$, si $\alpha \shortparallel \beta \shortparallel \gamma \shortparallel\!\!\sim_{\mathcal{M}} \beta \shortparallel \gamma$ et $\alpha \shortparallel \beta \not\shortparallel\!\!\sim_{\mathcal{M}} \beta$, alors $\alpha \shortparallel \gamma \shortparallel\!\!\sim_{\mathcal{M}} \gamma$

La preuve de *Reflexivity* est triviale. Celles pour *$\mathcal{U}$-Left Equivalence* et *$\mathcal{U}$-Right Weakening* sont similaires à celles données dans la section 2.3 ci-dessus pour $\shortparallel\!\!\sim_{\mathcal{M}}$ (il suffit de remplacer $\mathcal{L}$ par $\boldsymbol{L}^{\shortparallel}$, et plus généralement les notions définies relativement à $\mathcal{L}$ par les notions correspondantes définies relativement à $\boldsymbol{L}^{\shortparallel}$). Celles pour *Cut* et *Cautious Monotony* sont similaires à celles données dans [Kraus

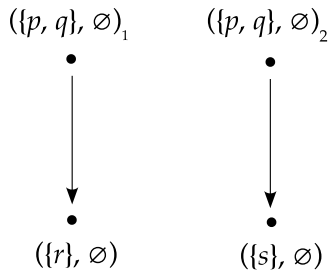

Figure 3 : $\mathcal{S} = (\mathcal{V}, <)$

et al., 1990][3]. La preuve de $\mathcal{U}$-*Consistence* est immédiate en faisant appel à $\boldsymbol{L}^{\|}$-*smoothness* et aux définitions de $\|\!\sim_{\mathcal{M}}$ et $\|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}$. Pour $\|$-*Or*, soit w <-minimal pour $\alpha \parallel \beta$: ou bien w $\|\!\models \alpha$, ou bien w $\|\!\models \beta$. Si w $\|\!\models \alpha$, alors il est <-minimal pour α (supposons que non, alors il existe w' < w tel que w' $\|\!\models \alpha$, donc w n'est pas <-minimal pour $\alpha \parallel \beta$, ce qui est absurde). Puisque par hypothèse $\alpha \|\!\sim_{\mathcal{M}} \gamma$, w $\|\!\models \gamma$. De même, si w $\|\!\models \beta$, alors w $\|\!\models \gamma$. La preuve pour *Injectivity* est donnée en Annexe B.

Injectivity est ainsi nommée parce qu'elle n'est pas valide dans les structures $\mathcal{S} = (\mathcal{V}, <)$ telles que $\mathcal{V}$ contient plusieurs copies d'un même monde partiel, ce qui dans le formalisme de KLM correspond au fait que la fonction d'étiquetage n'est pas injective. Par exemple, soit $\mathcal{V} = \{(\{p, q\}, \varnothing)_1, (\{p, q\}, \varnothing)_2, (\{r\}, \varnothing), (\{s\}, \varnothing)\}$. Considérons la structure $\mathcal{S} = (\mathcal{V}, <)$ de la Figure 3, et définissons $\|\!\sim_{\overline{\mathcal{V},\boldsymbol{L}}}$, $\|\!\sim_{\mathcal{S}}$, ... etc. comme attendu. $p \wedge q$ est maximal-consistant dans $\mathcal{V}$, cependant on a $(p \wedge q) \parallel r \parallel s \|\!\sim_{\mathcal{S}} r \parallel s$ et $(p \wedge q) \parallel r \not\|\!\sim_{\mathcal{S}} r$, mais $(p \wedge q) \parallel s \not\|\!\sim_{\mathcal{S}} s$.

Le sens de la règle *Injectivity* est plus facile à discerner si on considère la formulation suivante, qui est équivalente :

> Pour toute formule α maximale-consistante dans $\mathcal{U}$,
> si $\alpha \parallel \beta \not\|\!\sim_{\mathcal{M}} \beta$ et $\alpha \parallel \gamma \not\|\!\sim_{\mathcal{M}} \gamma$, alors $\alpha \parallel \beta \parallel \gamma \not\|\!\sim_{\mathcal{M}} \beta \parallel \gamma$

Puisque α est maximale-consistante dans $\mathcal{U}$, il existe un monde partiel w

3 [Kraus et al., 1990], p. 18

appartenant à $\mathcal{U}$ tel que $\alpha \equiv \delta(\mathrm{w})$. Selon l'interprétation visée de $\mathcal{M}$, w représente une représentation mentale $\boldsymbol{a}$ dans l'esprit de l'agent, donc w peut être aussi noté $\mathrm{w}_{\boldsymbol{a}}$. $\delta(\mathrm{w}_{\boldsymbol{a}}) = \delta(\mathrm{w})$ représente la conjonction des caractéristiques qui, d'après l'agent, sont satisfaites par les objets/situations desquels $\boldsymbol{a}$ est la représentation mentale. De même, soient $\boldsymbol{b}_1$, ... , $\boldsymbol{b}_n$ les représentations mentales de l'agent telles que $\mathcal{W}_u(\beta) = \{\mathrm{w}_{\boldsymbol{b1}}, \ldots, \mathrm{w}_{\boldsymbol{bn}}\}$, et $\boldsymbol{c}_1$, ... , $\boldsymbol{c}_m$ ses représentations mentales telles que $\mathcal{W}_u(\gamma) = \{\mathrm{w}_{\boldsymbol{c1}}, \ldots, \mathrm{w}_{\boldsymbol{cm}}\}$. Si $\beta \ncong_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$ et $\gamma \ncong_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$, alors $\beta \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \big\|\delta(\mathrm{w}_{\boldsymbol{bi}})/(\mathrm{w}_{\boldsymbol{bi}} \in \mathcal{W}_u(\beta))$ et $\gamma \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \big\|\delta(\mathrm{w}_{\boldsymbol{cj}})/(\mathrm{w}_{\boldsymbol{cj}} \in \mathcal{W}_u(\gamma))$, donc par substitution des $\|$-disjoints $\mathcal{U}$-équivalents (voir p.168), $\mathcal{U}$-*Left Equivalence* et $\mathcal{U}$-*Right Weakening*, on obtient :

si $\delta(\mathrm{w}_{\boldsymbol{a}})\| \delta(\mathrm{w}_{\boldsymbol{b1}})\| \ldots \| \delta(\mathrm{w}_{\boldsymbol{bn}}) \;|\!\!\sim_{\mathcal{M}}\; \delta(\mathrm{w}_{\boldsymbol{b1}})\| \ldots \| \delta(\mathrm{w}_{\boldsymbol{bn}})$
et $\delta(\mathrm{w}_{\boldsymbol{a}})\| \delta(\mathrm{w}_{\boldsymbol{c1}})\| \ldots \| \delta(\mathrm{w}_{\boldsymbol{cm}}) \;|\!\!\sim_{\mathcal{M}}\; \delta(\mathrm{w}_{\boldsymbol{c1}})\| \ldots \| \delta(\mathrm{w}_{\boldsymbol{cm}})$,
alors $\delta(\mathrm{w}_{\boldsymbol{a}})\| \delta(\mathrm{w}_{\boldsymbol{b1}})\| \ldots \| \delta(\mathrm{w}_{\boldsymbol{bn}}) \| \delta(\mathrm{w}_{\boldsymbol{c1}})\| \ldots \| \delta(\mathrm{w}_{\boldsymbol{cm}}) \;|\!\!\sim_{\mathcal{M}}\; \| \delta(\mathrm{w}_{\boldsymbol{b1}})\| \ldots$
$\ldots \| \delta(\mathrm{w}_{\boldsymbol{bn}}) \| \delta(\mathrm{w}_{\boldsymbol{c1}})\| \ldots \| \delta(\mathrm{w}_{\boldsymbol{cm}})$

ce qui peut s'interpréter ainsi : si lorsque l'information qu'il considère est compatible seulement avec ses représentations mentales $\boldsymbol{a}$, $\boldsymbol{b}_1$, ... , $\boldsymbol{b}_n$, l'agent ne néglige pas la possibilité $\boldsymbol{a}$, et si lorsque l'information qu'il considère est compatible seulement avec ses représentations mentales $\boldsymbol{a}$, $\boldsymbol{c}_1$, ... , $\boldsymbol{c}_m$ il ne néglige pas non plus la possibilité $\boldsymbol{a}$, alors, lorsque l'information qu'il considère est compatible seulement avec $\boldsymbol{a}$, $\boldsymbol{b}_1$, ..., $\boldsymbol{b}_n$, $\boldsymbol{c}_1$, ..., $\boldsymbol{c}_m$, il ne néglige pas non plus la possibilité $\boldsymbol{a}$. Si $\beta \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$, alors $\{\mathrm{w}_{\boldsymbol{b1}}, \ldots, \mathrm{w}_{\boldsymbol{bn}}\} = \varnothing$ et par substitution des $\|$-disjoints $\mathcal{U}$-équivalents, $\mathcal{U}$-*Left Equivalence* et $\mathcal{U}$-*Right Weakening*, le sens de *Injectivity* devient trivial. De même si $\gamma \cong_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$.

On appellera PW (pour 'partial worlds') l'ensemble de règles ci-dessus. Les règles suivantes peuvent être utilement dérivées de PW :

($\boldsymbol{L}^{\|}$)***Supra-$\mathcal{U}$-Consequence***
Si $\alpha \;|\!\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}\; \beta$, alors $\alpha \;|\!\!\sim_{\mathcal{M}}\; \beta$
(de *Reflexivity* et $\mathcal{U}$-*Right Weakening*).

Equivalence

Si $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$, $\beta \mid\!\!\sim_{\mathcal{M}} \alpha$ et $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, alors $\beta \mid\!\!\sim_{\mathcal{M}} \gamma$

(de *Cautious Monotony*, $\mathcal{U}$*-Left Equivalence* et *Cut*).

And

Si $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$ et $\alpha \mid\!\!\sim_{\mathcal{M}} \gamma$, alors $\alpha \mid\!\!\sim_{\mathcal{M}} \beta \wedge \gamma$

(de *Reflexivity*, $\mathcal{U}$*-Right Weakening*, *Cautious Monotony*et *Cut*).

$\|$***-Disjunct Equivalence***

Si $\alpha_1 \mid\!\!\sim_{\mathcal{M}} \alpha_2$, $\alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_1$ et $\alpha_1 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \gamma$, alors $\alpha_2 \parallel \beta \mid\!\!\sim_{\mathcal{M}} \gamma$

(de *Reflexivity*, $\mathcal{U}$*-Right Weakening*, *Cautious Monotony*,

$\mathcal{U}$*-Left Equivalence* et *Equivalence*).

$\|$***-Transitivity***

Si $\alpha_1 \mid\!\!\sim_{\mathcal{M}} \alpha_2$, ..., $\alpha_{n-1} \mid\!\!\sim_{\mathcal{M}} \alpha_n$, alors $\alpha_1 \parallel \alpha_n \mid\!\!\sim_{\mathcal{M}} \alpha_n$

(de *Reflexivity*, $\|$*-Or*, $\mathcal{U}$*-Right Weakening*,*Cautious Monotony*,

$\mathcal{U}$*-Left Equivalence* et $\|$*-Disjunct Equivalence*).

La preuve de *Supra-*$\mathcal{U}$*-Consequence* est immédiate. Celles de *Equivalence* et *And* sont similaires à celles données dans [Kraus et al., 1990][4]. Celles de $\|$*-Disjunct Equivalence* et $\|$*-Transitivity* sont données dans les Annexes C et D.

4.4 Quelques définitions supplémentaires

Pour finir, on donne deux définitions supplémentaires, qui nous seront utiles au chapitre suivant.

4.4.1 La formule $C_{\|\sim}(\alpha)$

Soit $\boldsymbol{L}^{\|}/_{\cong_{\mathcal{U},\boldsymbol{L}^{\|}}}$ l'ensemble des classes d'équivalence de $\boldsymbol{L}^{\|}$ sous $\cong_{\mathcal{U},\boldsymbol{L}^{\|}}$, et g une fonction de choix sur $\boldsymbol{L}^{\|}/_{\cong_{\mathcal{U},\boldsymbol{L}^{\|}}}$ qui sélectionne un représentant pour chaque

4 [Kraus et al., 1990], (p. 14)

classe d'équivalence. Puisque $\boldsymbol{L}^{\|}$ est fini, $\boldsymbol{L}^{\|}/_{\cong_{\mathcal{U},\boldsymbol{L}^{\|}}}$ l'est aussi. Pour toute $\boldsymbol{L}^{\|}$-formule α, on note $C_{\|\sim}(\alpha)$ la conjonction des $\boldsymbol{L}^{\|}$-formules β telles que : $\alpha \,\|\!\sim \beta$ et β est le représentant de sa classe d'équivalence sous $\cong_{\mathcal{U},\boldsymbol{L}^{\|}}$. Par une série d'applications de la règle dérivée *And*, on montre aisément que pour toute $\boldsymbol{L}^{\|}$-formule α, $\alpha \,\|\!\sim C_{\|\sim}(\alpha)$. De plus, pour toutes $\boldsymbol{L}^{\|}$-formules α et β, $\alpha \,\|\!\sim \beta$ ssi $C_{\|\sim}(\alpha) \,\|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}\, \beta$: en effet, si $\alpha \,\|\!\sim \beta$, alors β est $\mathcal{U}$-équivalent à un conjoint de $C_{\|\sim}(\alpha)$, et donc tout monde partiel de $\mathcal{U}$ qui satisfait $C_{\|\sim}(\alpha)$ satisfait aussi β, c'est à dire que $C_{\|\sim}(\alpha) \,\|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}\, \beta$. Réciproquement, supposons que $C_{\|\sim}(\alpha) \,\|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}\, \beta$. Puisque $\alpha \,\|\!\sim C_{\|\sim}(\alpha)$, par $\mathcal{U}$-*Right Weakening* on obtient $\alpha \,\|\!\sim \beta$.

4.4.2 Précisifications et modèles à mondes partiels sans précisifications

Étant donné w $\in \mathcal{W}_{\boldsymbol{L}}^{\mathrm{P}}$, une *précisification* de w est un monde partiel w' $\in \mathcal{W}_{\boldsymbol{L}}^{\mathrm{P}}$ tel que $\boldsymbol{w} \subset \boldsymbol{w}$' (où $\boldsymbol{w}$ et $\boldsymbol{w}$' sont les distributions de valeurs de vérité qui engendrent respectivement w et w', et $\subset$ est l'inclusion stricte). En d'autres termes, w' est une précisification de w si et seulement si δ(w') $\vdash \delta$(w) et δ(w) $\nvdash \delta$(w'), ou, de manière équivalente, si et seulement si δ(w') $\|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}\, \delta$(w) et δ(w) $\|\!\!\!/_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}\, \delta$(w') (puisque δ(w) et δ(w) sont des $\boldsymbol{L}$-formules, voir p. 167 ci-dessus).

On dira qu'un ensemble de mondes partiels $\mathcal{U}$ est *sans précisifications* si et seulement si pour tous w et w' de $\mathcal{U}$, w' *n'est pas* une précisification de w.

Si $\mathcal{U}$ est sans précisifications, alors il est immédiat que pour tout w appartenant à $\mathcal{U}$, δ(w) est maximal-consistante dans $\mathcal{U}$. Comme (*modulo* l'équivalence classique) $\{\alpha/\alpha$ est maximal-consistante dans $\mathcal{U}\} \subseteq \{\delta(\mathrm{w})/\,\mathrm{w} \in \mathcal{U}\}$ (cf. p. 170), il s'ensuit que, si $\mathcal{U}$ est sans précisifications, alors (*modulo* l'équivalence classique) $\{\alpha/\alpha$ est maximal-consistante dans $\mathcal{U}\} = \{\delta(\mathrm{w})/\mathrm{w} \in \mathcal{U}\}$.

De plus, si $\mathcal{U}$ est sans précisifications, alors $D_{\mathcal{U},\boldsymbol{L}^{\|}} = \mathcal{P}(\mathcal{U})$. En effet, supposons qu'il existe un ensemble $A = \{\mathrm{w}_1\ ,\ \ldots\ ,\ \mathrm{w}_n\} \subseteq \mathcal{U}$ qui n'est pas $\boldsymbol{L}^{\|}$-définissable. C'est à dire que pour toute $\boldsymbol{L}^{\|}$-formule α telle que tout w_i de A

satisfait α, il existe w' appartenant à $\mathcal{U}$ tel que w' $\notin A$ et w' $\Vdash \alpha$. En particulier, il existe w' tel que w' $\Vdash \big\|\delta(\mathrm{w}_i)/(\mathrm{w}_i \in A)$ et w' $\notin A$. Ceci implique qu'il existe un $\mathrm{w}_i \in A$ tel que w' $\Vdash \delta(\mathrm{w}_i)$, c'est à dire, tel que $\delta(\mathrm{w'}) \vdash \delta(\mathrm{w}_i)$. Puisque w' $\notin A$, w' $\neq \mathrm{w}_i$, donc w' est une précisification de w_i.

On dira qu'un modèle à mondes partiels $\mathcal{M} = (\mathcal{U}, <)$ est *sans précisifications* si et seulement si $\mathcal{U}$ est sans précisifications.

Chapitre 5

Deux théorèmes de représentation

On donne dans ce chapitre deux théorèmes de représentation. Le premier établit que PW est adéquat et complet pour les relations d'inférence induites sur $\boldsymbol{L}^{\shortparallel}$ par les modèles à mondes partiels $\boldsymbol{L}^{\shortparallel}$-smooth finis sans précisifications. Pour le second, on introduit d'abord une notion de modèles à mondes partiels *rangés*, inspirée de la notion de *modèles rangés* de Lehmann et Magidor [1], et une règle supplémentaire *Rankedness*. Puis on montre que PW $\cup$ $\{$*Rankedness*$\}$ est adéquat et complet pour les relations d'inférence induites sur $\boldsymbol{L}^{\shortparallel}$ par les modèles à mondes partiels rangés finis sans précisifications [2].

1 [Lehmann and Magidor, 1992]

2 Il a été remarqué que ces deux résultats seraient conservés si on définissait $\mathcal{U}$ plus généralement comme un sous-ensemble de $\mathcal{P}(\mathcal{W}_{\boldsymbol{L}})$ — il nous faudrait alors définir une précisification d'un élément s de $\mathcal{U}$ comme un sous ensemble strict de s, $\delta(s)$ comme $\bigvee\delta(\mathrm{w})/\mathrm{w} \in s$, et les autres notions comme attendues. PW et PW $\cup$ $\{$*Rankedness*$\}$ seraient en effet respectivement adéquats et complets pour les relations d'inférence induites sur $\boldsymbol{L}^{\shortparallel}$ par les modèles finis sans précisification $\boldsymbol{L}^{\shortparallel}$-smooth et rangés résultants (ce sont les propriétés de $\|\!\sim_{\mathcal{U},\boldsymbol{L}}$ qui changeraient). Cependant de tels modèles ne pourraient pas être utilisés pour les fins qui sont les nôtres, puisque les éléments de $\mathcal{U}$ ont vocation à représenter des représentations mentales d'objets/situations, lesquelles ne peuvent être figurées que par des mondes partiels.

5.1 Théorème de représentation pour les modèles à mondes partiels $\boldsymbol{L}^{\|}$-smooth finis sans précisifications

Dans la section 4.3, on a montré que PW est adéquat pour les relations d'inférence induites sur $\boldsymbol{L}^{\|}$ par les modèles à mondes partiels finis $\boldsymbol{L}^{\|}$-smooth. Il est donc *a fortiori* adéquat pour celles induites par les modèles à mondes partiels finis $\boldsymbol{L}^{\|}$-smooth sans précisifications. On va montrer ici que PW est aussi complet pour ces dernières. Pour ce faire, on va montrer que toute relation $\|\!\sim$ qui satisfait PW et telle que $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$ est générée par un sous-ensemble $\mathcal{U}$ de $\mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$ ne contenant pas de précisifications admet un modèle à mondes partiels $(\mathcal{U}, <)$ tel que $<$ est $\boldsymbol{L}^{\|}$-smooth.

Soit $\boldsymbol{L}$, $\boldsymbol{L}^{\|}$ et $\mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$ comme précédemment, $\mathcal{U}$ un sous-ensemble de $\mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$ sans précisifications, $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$ la relation de $\mathcal{U}$-conséquence induite sur $\boldsymbol{L}^{\|}$ par $\mathcal{U}$, et $\|\!\sim$ une relation binaire sur $\boldsymbol{L}^{\|}$ satisfaisant PW. Puisque $\boldsymbol{L}$ est fini, $\mathcal{U}$ est un ensemble fini de mondes partiels finis. On définit une relation $<$ sur $\mathcal{U}$ par : pour tous w et w' de $\mathcal{U}$, w $<$ w' ssi w $\neq$ w' et δ(w')$\|$ δ(w) $\|\!\sim \delta$(w). Il est immédiat que $\mathcal{M} = (\mathcal{U}, <)$ est un modèle à mondes partiels fini sans précisifications. On montre que $<$ est $\boldsymbol{L}^{\|}$-smooth :

Lemme 1 *Pour tout* w $\in \mathcal{U}$, w $\|\!\models C_{\|\sim}(\delta(\mathrm{w}))$

Démonstration :

Soit w $\in \mathcal{U}$:

i) δ(w) $\|\!\not\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$, donc par $\mathcal{U}$-*Consistence*,
δ(w) $\|\!\not\sim \bot$, ssi
$C_{\|\sim}(\delta(\mathrm{w}))$ $\|\!\not\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \bot$, ssi
$\exists$ w' $\in \mathcal{U}$ tel que w' $\|\!\models C_{\|\sim}(\delta(\mathrm{w}))$.

ii) Par *Reflexivity*, δ(w) $\|\!\sim \delta$(w), ssi
$C_{\|\sim}(\delta(\mathrm{w}))$ $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}} \delta$(w), donc par i),
w' $\|\!\models \delta$(w).
Donc w' $\|\!\models \delta(\mathrm{w}) \wedge \delta(\mathrm{w}')$, donc

$\delta(\mathrm{w}) \wedge \delta(\mathrm{w'}) \not\|\!\sim_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \bot$.

iii) Par hypothèse $\mathcal{U}$ est sans précisifications, donc $\delta(\mathrm{w})$ et $\delta(\mathrm{w'})$ sont toutes deux maximales-consistantes. Par définition de maximal-consistance, ii) implique que $\delta(\mathrm{w}) \equiv \delta(\mathrm{w'})$, ssi $\mathrm{w} = \mathrm{w'}$. Par i), ceci implique que $\mathrm{w} \|\!= C_{\|\sim}(\delta(\mathrm{w}))$.

□

Lemme 2 $<$ *est* $\boldsymbol{L}^{\|}$*-smooth.*

Démonstration :

i) De par sa définition, $<$ est irréflexive.

ii) De plus, $<$ est transitive : en effet, supposons qu'il existe w_1, w_2 and w_3 appartenant à $\mathcal{U}$ et tels que $\mathrm{w}_3 < \mathrm{w}_2 < \mathrm{w}_1$. Alors :

1. Par définition de $<$, $\delta(\mathrm{w}_1) \| \delta(\mathrm{w}_2) \|\!\sim \delta(\mathrm{w}_2)$ et $\delta(\mathrm{w}_2) \| \delta(\mathrm{w}_3) \|\!\sim \delta(\mathrm{w}_3)$.

2. Par *Reflexivity*, $\delta(\mathrm{w}_1) \|\!\sim \delta(\mathrm{w}_1)$, donc par $\mathcal{U}$*-Right Weakening* $\delta(\mathrm{w}_1) \|\!\sim \delta(\mathrm{w}_1) \| \delta(\mathrm{w}_2)$.
De même, $\delta(\mathrm{w}_2) \|\!\sim \delta(\mathrm{w}_2) \| \delta(\mathrm{w}_3)$.

3. Par $\|$-*Transitivity* sur 1 et 2, on obtient $\delta(\mathrm{w}_1) \| \delta(\mathrm{w}_3) \|\!\sim \delta(\mathrm{w}_3)$.

4. Par ailleurs, $\mathrm{w}_1 \neq \mathrm{w}_3$. Supposons en effet que $\mathrm{w}_1 = \mathrm{w}_3$:

a) Alors $\delta(\mathrm{w}_1) = \delta(\mathrm{w}_3)$, donc par 1, $\delta(\mathrm{w}_3) \| \delta(\mathrm{w}_2) \|\!\sim \delta(\mathrm{w}_2)$.
Donc par 2 + *Equivalence*, $C_{\|\sim}(\delta(\mathrm{w}_2)) = C_{\|\sim}(\delta(\mathrm{w}_2) \| \delta(\mathrm{w}_3))$.

b) Par *Reflexivity*, $\delta(\mathrm{w}_3) \|\!\sim \delta(\mathrm{w}_3)$, donc par $\mathcal{U}$*-Right Weakening*, $\delta(\mathrm{w}_3) \|\!\sim \delta(\mathrm{w}_2) \| \delta(\mathrm{w}_3)$.
Par 1 + *Equivalence*, $C_{\|\sim}(\delta(\mathrm{w}_3)) = C_{\|\sim}(\delta(\mathrm{w}_2) \| \delta(\mathrm{w}_3))$.
Donc par a), $C_{\|\sim}(\delta(\mathrm{w}_2)) = C_{\|\sim}(\delta(\mathrm{w}_3))$.

c) Par *Reflexivity*, $C_{\|\sim}(\delta(\mathrm{w}_2)) \|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \delta(\mathrm{w}_2)$ et $C_{\|\sim}(\delta(\mathrm{w}_3)) \|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \delta(\mathrm{w}_3)$.

d) Par le Lemme 1, $\delta(\mathrm{w}_2) \|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} C_{\|\sim}(\delta(\mathrm{w}_2))$, donc par b) + c) + transitivité de $\|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}}$, $\delta(\mathrm{w}_2) \|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \delta(\mathrm{w}_3)$, ssi (puisque $\delta(\mathrm{w}_3)$ est une $\boldsymbol{L}$-formule) $\delta(\mathrm{w}_2) \vdash \delta(\mathrm{w}_3)$.

e) De même, $\delta(\mathrm{w}_3) \|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} C_{\|\sim}(\delta(\mathrm{w}_3)) = C_{\|\sim}(\delta(\mathrm{w}_2)) \|_{\overline{\mathcal{U},\boldsymbol{L}^{\|}}} \delta(\mathrm{w}_2)$, donc $\delta(\mathrm{w}_3) \vdash \delta(\mathrm{w}_2)$.
Par d), $\delta(\mathrm{w}_2) \equiv \delta(\mathrm{w}_3)$, ssi $\mathrm{w}_2 = \mathrm{w}_3$

f) Par hypothèse $w_3 < w_2$, donc par définition de $<$, $w_2 \neq w_3$, ce qui contredit e).

5. par 3 et 4, $w_3 < w_1$.

iii) Par i) et ii), $<$ est acyclique.

iv) Puisque $\mathcal{U}$ est fini, il résulte de iii) que pour tout sous-ensemble $A \subseteq \mathcal{U}$, il existe au moins un $w \in A$ tel que w is $<$-minimal dans A. Puisque $<$ est transitive, pour tout w' $\in A$ qui n'est pas $<$-minimal dans A, il existe un $w \in A$ tel que w est $<$-minimal dans A est $w < w'$. Ceci vaut notamment pour les sous-ensembles $A \subseteq \mathcal{U}$ tels que A est $\boldsymbol{L}^{\shortparallel}$-définissable.

□

Il résulte du Lemme 2 que $\mathcal{M}$ est un modèle à mondes partiels $\boldsymbol{L}^{\shortparallel}$-smooth fini sans précisifications. On montre maintenant que $\mathcal{M}$ est un modèle de $\shortparallel\!\sim$:

Soit $\shortparallel\!\sim_{\mathcal{M}}$ la relation d'inférence induite par $\mathcal{M}$ sur $\boldsymbol{L}^{\shortparallel}$. On va montrer que $\shortparallel\!\sim_{\mathcal{M}} = \shortparallel\!\sim$. Comme on l'a vu dans la section 4.3, $\shortparallel\!\sim_{\mathcal{M}}$ satisfait PW et ses règles dérivées. Soit α une $\boldsymbol{L}^{\shortparallel}$-formule.

Lemme 3 *Si $\mathcal{W}_u(\alpha) = \varnothing$, alors, pour toute $\boldsymbol{L}^{\shortparallel}$-formule β, on a à la fois $\alpha \shortparallel\!\sim_{\mathcal{M}} \beta$ et $\alpha \shortparallel\!\sim \beta$.*

Démonstration :
Immédiate, en utilisant le fait que $\mathcal{W}_u(\alpha) = \varnothing$ ssi $\alpha \shortparallel\!\sim_{\mathcal{U},\boldsymbol{L}^{\shortparallel}} \bot$, *Supra-$\mathcal{U}$-Consequence* et *$\mathcal{U}$-Right Weakening*.

□

Pour le reste de la démonstration, on supposera que $\mathcal{W}_u(\alpha) \neq \varnothing$.

Soit $X = \{w \in \mathcal{W}_u(\alpha) \,/\, \forall\, w' \in \mathcal{W}_u(\alpha)$ t.q. $w' \neq w$, $\delta(w) \,\|\, \delta(w') \shortparallel\!\not\sim \delta(w')\}$, et $Y = \mathcal{W}_u(\alpha) - X$. Par la définition de $<$, X est l'ensemble des éléments $<$-minimaux de $\mathcal{W}_u(\alpha)$. Par le Lemme 2, $X \neq \varnothing$. Disons que $X = \{w_1, \ldots, w_n\}$.

Pour chaque $w_i \in X$, soit $y_i = \{w \in Y / \ \delta(w) \| \ \delta(w_i) \shortparallel\!\sim \delta(w_i)\}$. Par la définition de $<$, y_i est l'ensemble des éléments de $\mathcal{W}_u(\alpha)$ qui sont minimisés par w_i (y_i peut être vide). Par le Lemme 2, $\bigcup y_i/(w_i \in X) = Y$.

Lemme 4 $C_{\Vdash\!\sim}(\alpha) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\parallel}}} \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$

Démonstration :

i) Pour tout $\mathrm{w}_i \in X$, par des applications successives de $\|$-*Or* + $\mathcal{U}$*-Left Equivalence*, on obtient $(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \mathrm{y}_i)) \,\|\delta(\mathrm{w}_i) \Vdash\!\sim \delta(\mathrm{w}_i)$.

Donc par $\mathcal{U}$*-Right Weakening*, pour tout $\mathrm{w}_i \in X$, on a :

$(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \mathrm{y}_i)) \,\|\, \delta(\mathrm{w}_i) \Vdash\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$.

ii) Par une série d'applications de $\|$-*Or* + $\mathcal{U}$*-Left Equivalence* sur i), on obtient $(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in Y)) \,\|\, (\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)) \Vdash\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$, c'est à dire $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \mathcal{W}_u(\alpha)) \Vdash\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$.

iii) Par hypothèse $\mathcal{W}_u(\alpha) \neq \varnothing$, donc $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \mathcal{W}_u(\alpha)) \cong_{\mathcal{U},\boldsymbol{L}^{\parallel}} \alpha$.

donc par $\mathcal{U}$*-Left Equivalence* sur ii), $\alpha \Vdash\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$, ssi

$C_{\Vdash\!\sim}(\alpha) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\parallel}}} \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$.

□

Lemme 5 $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\parallel}}} C_{\Vdash\!\sim}(\alpha)$

Démonstration :

Supposons le contraire, c'est à dire qu'il existe $\mathrm{w}_k \in X$ tel que $\delta(\mathrm{w}_k) \not\Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\parallel}}} C_{\Vdash\!\sim}(\alpha)$. Par simplicité, disons que w_k est w_1.

i) Par la règle dérivée *And*, $\alpha \Vdash\!\sim C_{\Vdash\!\sim}(\alpha)$.

ii) Pour tout $\mathrm{w}_k \in X$, $\mathrm{w}_k \Vdash\!= \alpha$, ssi $\delta(\mathrm{w}_k) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\parallel}}} \alpha$.

Donc $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\parallel}}} \alpha$, donc par *Supra-*$\mathcal{U}$*-Consequence*, $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash\!\sim \alpha$.

iii) Dans la démonstration du Lemme 4 (au point iii) on a montré que $\alpha \Vdash\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$. Donc par la règle dérivée *Equivalence* sur i) et ii), on obtient $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash\!\sim C_{\Vdash\!\sim}(\alpha)$.

iv) Par *Reflexivity*, $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash\!\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$, donc par *And* sur iii), $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash\!\sim (\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)) \wedge C_{\Vdash\!\sim}(\alpha)$.

v) $(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)) \wedge C_{\Vdash\!\sim}(\alpha) \cong_{\mathcal{U},\boldsymbol{L}^{\parallel}} (\delta(\mathrm{w}_1) \wedge C_{\Vdash\!\sim}(\alpha)) \,\|\, (\delta(\mathrm{w}_2) \wedge C_{\Vdash\!\sim}(\alpha))\|$ ... $\|\, (\delta(\mathrm{w}_n) \wedge C_{\Vdash\!\sim}(\alpha))$.

vi) Par hypothèse, $\delta(\mathrm{w}_1) \not\Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\parallel}}} C_{\Vdash\!\sim}(\alpha)$, ssi $\mathrm{w}_1 \Vdash\!\neq C_{\Vdash\!\sim}(\alpha)$.

Par hypothèse, $\mathcal{U}$ est sans précisifications, ce qui implique qu'il n'existe

aucun w appartenant à $\mathcal{U}$ et tel que w $\Vdash \delta(\mathrm{w}_1) \wedge C_{\Vdash\sim}(\alpha)$.

Donc $\delta(\mathrm{w}_1) \wedge C_{\Vdash\sim}(\alpha) \cong_{\mathcal{U},\boldsymbol{L}^{\Vdash}} \bot$.

vii) Nécessairement, il existe $\mathrm{w}_2 \in X$ tel que $\mathrm{w}_2 \neq \mathrm{w}_1$: en effet, supposons le contraire, alors $\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) = \delta(\mathrm{w}_1)$, donc par $\mathcal{U}$-*Right Weakening* sur iv) et vi), $\delta(\mathrm{w}_1) \Vdash\sim \bot$, donc par $\mathcal{U}$-*Consistence*, $\delta(\mathrm{w}_1) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\Vdash}}} \bot$, ssi $\mathrm{w}_1 \notin \mathcal{U}$, ce qui est absurde.

viii) Par v) et vi), $(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \mathrm{X})) \wedge C_{\Vdash\sim}(\alpha) \cong_{\mathcal{U},\boldsymbol{L}^{\Vdash}}$
$(\delta(\mathrm{w}_2) \wedge C_{\Vdash\sim}(\alpha))\| \ldots \| (\delta(\mathrm{w}_n) \wedge C_{\Vdash\sim}(\alpha)) \cong_{\mathcal{U},\boldsymbol{L}^{\Vdash}}$
$(\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_2, \ldots, \mathrm{w}_n\})) \wedge C_{\Vdash\sim}(\alpha)$.

ix) Par $\mathcal{U}$-*Right Weakening* sur iv) et viii),
$\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash\sim (\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_2, \ldots, \mathrm{w}_n\})) \wedge C_{\Vdash\sim}(\alpha)$,
donc par $\mathcal{U}$-*Right Weakening*,
$\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_2, \ldots, \mathrm{w}_n\})$.

x) $\mathrm{w}_1 \in X$, donc par définition de X, $\delta(\mathrm{w}_1) \| \delta(\mathrm{w}_2) \Vdash\not\sim \delta(\mathrm{w}_2)$.

Donc par ix), il existe $\mathrm{w}_3 \in X$ tel que $\mathrm{w}_1 \neq \mathrm{w}_3 \neq \mathrm{w}_2$.

xi) Par hypothèse $\mathcal{U}$ est sans précisifications, donc pour tout $\mathrm{w}_i \in X$, $\delta(\mathrm{w}_i)$ est maximale-consistante.

Donc, puisque $\delta(\mathrm{w}_1)\| \delta(\mathrm{w}_2) \Vdash\not\sim \delta(\mathrm{w}_2)$, par *Injectivity* sur ix) et x),
$\delta(\mathrm{w}_1) \| (\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_3, \ldots, \mathrm{w}_n\})) \Vdash\sim \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in \{\mathrm{w}_3, \ldots, \mathrm{w}_n\})$.
Puisque pour tout i $(3 \leq i \leq \mathrm{n})$, $\delta(\mathrm{w}_1) \| \delta(\mathrm{w}_i) \Vdash\not\sim \delta(\mathrm{w}_i)$, des applications successives de la règle *Injectivity* mènent à une contradiction.
Donc $\delta(\mathrm{w}_1) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}}} C_{\Vdash\sim}(\alpha)$.

Un raisonnement similaire peut être fait pour tout $\mathrm{w}_k \in X$ (il suffit de changer l'indexation). Donc pour tout w_k de X, $\delta(\mathrm{w}_k) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\Vdash}}} C_{\Vdash\sim}(\alpha)$, ssi
$\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\Vdash}}} C_{\Vdash\sim}(\alpha)$.

□

Lemme 6 *Si $\mathcal{W}_u(\alpha) \neq \varnothing$, alors, pour toute $\boldsymbol{L}^{\shortparallel}$-formule β, on a $\alpha \mid\!\!\sim \beta$ ssi $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$.*

Démonstration :

Supposons que $\mathcal{W}_u(\alpha) \neq \varnothing$. Par les Lemmes 4 et 5, $C_{\shortparallel\sim}(\alpha) \cong_{\mathcal{U}, \boldsymbol{L}^{\shortparallel}} \big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X)$.

Pour toute $\boldsymbol{L}^{\shortparallel}$-formule β, on a :

$\alpha \mid\!\!\sim \beta$ ssi

$C_{\shortparallel\sim}(\alpha) \mid\!\!\sim_{\overline{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}} \beta$, ssi

$\big\|\delta(\mathrm{w}_k)/(\mathrm{w}_k \in X) \mid\!\!\sim_{\overline{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}} \beta$, ssi

pour tout w_k tel que $\mathrm{w}_k \in X$, $\mathrm{w}_k \mid\!\!\sim_{\overline{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}} \beta$, ssi

pour tout w_k tel que w_k est $<$-minimal pour α, $\mathrm{w}_k \models \beta$, ssi

$\alpha \mid\!\!\sim_{\mathcal{M}} \beta$

□

Lemme 7 *$\mathcal{M} = (\mathcal{U}, <)$ est un modèle à mondes partiels $\boldsymbol{L}^{\shortparallel}$-smooth fini sans précisifications tel que $\mid\!\!\sim_{\mathcal{M}} = \mid\!\!\sim$.*

Démonstration :

Immédiate à partir des Lemmes 2 à 6.

□

Théorème 1 *Si $\mid\!\!\sim$ est une relation sur $\boldsymbol{L}^{\shortparallel}$ qui satisfait PW et telle que $\mid\!\!\sim_{\overline{\mathcal{U}, \boldsymbol{L}^{\shortparallel}}}$ est engendrée par un sous-ensemble $\mathcal{U}$ de $\mathcal{W}_{\boldsymbol{L}}^{\mathrm{p}}$ qui ne contient pas de précisifications, alors $\mid\!\!\sim$ admet un modèle à mondes partiels $\boldsymbol{L}^{\shortparallel}$-smooth fini sans précisifications.*

Démonstration :

Immédiate, à partir du Lemme 6.

□

Corollaire 1 *PW est adéquat et complet pour les relations d'inférences induites sur $\boldsymbol{L}^{\shortparallel}$ par les modèles à mondes partiels $\boldsymbol{L}^{\shortparallel}$-smooth finis sans précisifications.*

Remarque : $\|$-*Transitivity* est valide dans tout modèle à mondes partiels $\boldsymbol{L}^{\|}$-smooth, que $<$ soit transitive ou non. Cependant $\|$-*Transitivity* force la construction ci-dessus à être transitive. De ce fait, pour tout modèle à mondes partiels $\boldsymbol{L}^{\|}$-smooth fini sans précisifications $\mathcal{M}$ il existe un modèle à mondes partiels fini sans précisifications transitif (et irréflexif) $\mathcal{M}'$ tel que $\|\sim_{\mathcal{M}} = \|\sim_{\mathcal{M}'}$. D'autre part, tout modèle strictement ordonné est $\boldsymbol{L}^{\|}$-smooth pourvu que $\mathcal{U}$ soit fini. Par conséquent, bien que la classe des modèles à mondes partiels finis strictement ordonnés sans précisifications soit une sous-classe stricte de celle des modèles $\boldsymbol{L}^{\|}$-smooth finis sans précisifications, toutes deux engendrent le même ensemble de relations d'inférence induites (contrairement à ce qui est le cas pour les modèles KLM, pour lesquels *smoothness* et ordre strict engendrent des ensembles différents de relations d'inférence induites, même dans le cas fini).

5.2 Théorème de représentation pour les modèles à mondes partiels rangés finis sans précisifications

Une relation binaire $<$ sur un ensemble E est un *ordre modulaire* ssi $<$ est un ordre partiel strict qui satisfait en outre la propriété [3]

Modularité *(1ère version)* :
Pour tous x, y et z appartenant à E, si $x < y$, alors $z < y$ ou $x < z$.

Si on écrit $y \leq z$ pour signifier que $z \not< y$, alors on peut de manière équivalente formuler la propriété *Modularité* comme suit :

Modularité *(2ème version)* :
Pour tous x, y et z appartenant à E, si $x < y$ et $y \leq z$, alors $x < z$.

En pratique, *Modularité* ordonne les éléments de E en 'rangs'.

3 Définition reprise de [Lehmann and Magidor, 1992] p. 19.

Si $\mathcal{M} = (\mathcal{U}, <)$ est un modèle à monde partiels tels que $<$ est un ordre modulaire, on dira que $\mathcal{M}$ est un modèle à mondes partiels *rangé*.

Soit $\boldsymbol{L}$ un langage propositionnel fini, $\mathcal{U} \subseteq \mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$, et $\mathcal{M} = (\mathcal{U}, <)$ un modèle à mondes partiels rangé. Puisque $\mathcal{U}$ est un ensemble fini de mondes partiels finis et $<$ est un ordre partiel strict, $\mathcal{M}$ est un modèle à mondes partiels $\boldsymbol{L}^{\|}$-smooth fini. Par conséquent $\|\!\sim_{\mathcal{M}}$ satisfait les règles de PW. Par surcroit, $\|\!\sim_{\mathcal{M}}$ satisfait la règle *Rankedness* :

> ***Rankedness*** Pour toutes $\boldsymbol{L}^{\|}$-formules α_1, α_2 et α_3 (toutes $\not\equiv$) maximales-consistantes dans $\mathcal{U}$,
> si $\alpha_1 \parallel \alpha_2 \; \|\!\sim_{\mathcal{M}} \alpha_2$ et $\alpha_1 \parallel \alpha_3 \; \|\!\not\sim_{\mathcal{M}} \alpha_3$, alors $\alpha_2 \parallel \alpha_3 \; \|\!\sim_{\mathcal{M}} \alpha_2$

(voir la démonstration en Annexe E). *A fortiori*, si $\mathcal{M} = (\mathcal{U}, <)$ est un modèle à mondes partiels fini rangé sans précisifications, alors $\|\!\sim_{\mathcal{M}}$ satisfait PW $\cup$ $\{$*Rankedness*$\}$.

Réciproquement, si $\boldsymbol{L}$ est un langage propositionnel fini, $\mathcal{U}$ un sous-ensemble sans précisifications de $\mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$, $\|\!\sim_{\mathcal{U},\boldsymbol{L}^{\|}}$ la relation de $\mathcal{U}$-conséquence engendrée sur $\boldsymbol{L}^{\|}$ par $\mathcal{U}$ et $\|\!\sim$ une relation binaire sur $\boldsymbol{L}^{\|}$ qui satisfait PW $\cup$ $\{$*Rankedness*$\}$, alors, en utilisant le fait que lorsque $\mathcal{U}$ est sans précisifications (et *modulo* l'équivalence classique) $\{\alpha \in \boldsymbol{L}^{\|} / \alpha \text{ est maximale-consistante dans } \mathcal{U}\} = \{\delta(\mathrm{w})/\mathrm{w} \in \mathcal{U}\}$, on vérifie aisément que la construction proposée section 5.1 ci-dessus produit un modèle à mondes partiels rangé fini sans précisifications. Par suite, PW $\cup$ $\{$*Rankedness*$\}$ est adéquat et complet pour les relations d'inférence induites sur $\boldsymbol{L}^{\|}$ par les modèles à mondes partiels rangés finis sans précisifications.

Chapitre 6

Modélisation des inférences automatiques et de l'apprentissage

Pour finir, on donne des précisions sur la manière dont les modèles à monde partiels peuvent être utilisés pour modéliser les inférences automatiques et l'apprentissage. Tout d'abord (section 6.1), on formalise la modélisation des inférences automatiques qu'on avait esquissée dans la section 1.2. Après quoi (section 6.2) on examine les processus d'apprentissage associés aux inférences automatiques, et on montre comment ceux-ci peuvent être modélisés dans le dispositif logique proposé.

6.1 Modélisation des inférences automatiques

Soit $\mathcal{A}$ l'agent cognitif considéré. On montre comment choisir un langage propositionnel $\boldsymbol{L}$ et construire un modèle à mondes partiels $\mathcal{M} = (\mathcal{U}, <)$ tel que pour toute $\boldsymbol{L}^{\shortparallel}$-formules α et β, $\alpha \mid\!\!\sim_{\mathcal{M}} \beta$ si et seulement si $\mathcal{A}$ est disposé à inférer β de α, au sens défini p. 144 ci-dessus.

Soit $\mathcal{F}^+$ l'ensemble des caractéristiques que $\mathcal{A}$ est physiologiquement capable de percevoir. On choisit un langage propositionnel $\boldsymbol{L}$ tel qu'il existe une bijection $\sigma : \mathcal{F}^+ \longrightarrow Var(\boldsymbol{L})$. Puisque $\mathcal{F}^+$ est fini[1], $\boldsymbol{L}$ est fini.

1 Voir p. 140 ci-dessus.

Comme on l'a remarqué dans la section 1.2 ci-dessus (p. 141), les agents sont capables de concevoir la négation des caractéristiques qu'ils peuvent percevoir, et de raisonner sur ces caractéristiques niées comme s'il s'agissait de caractéristiques (négatives). Soit η la fonction qui à chaque caractéristique f de $\mathcal{F}^+$ associe sa négation *non-f*, et $\mathcal{F}^-$ l'image de $\mathcal{F}^+$ par η. $\mathcal{F} = \mathcal{F}^+ \cup \mathcal{F}^-$ est l'ensemble des caractéristiques sur lesquelles $\mathcal{A}$ est capable raisonner.

On définit une fonction $\rho : \mathcal{F} \longrightarrow \{\lambda \in \boldsymbol{L} \ / \ \lambda \text{ est un littéral}\}$ par :

Si $f \in \mathcal{F}^+$, alors $\rho(f) = \sigma(f)$, sinon, $\rho(f) = \neg\sigma(\eta^{-1}(f))$

Il est immédiat que ρ est une bijection. Elle définit une relation de représentation entre les littéraux de $\boldsymbol{L}$ et les caractéristiques appartenant à $\mathcal{F}$, au sens où λ représentera f dans le modèle si et seulement si $\lambda = \rho(f)$. On vérifie aisément que pour toute $p \in Var(\boldsymbol{L})$ et toute $f \in \mathcal{F}^+$, $p = \rho(f)$ si et seulement si $\neg p = \rho(\textit{non-}f)$.

Soit $\mathcal{R}$ l'ensemble des représentations mentales les plus précises de $\mathcal{A}$. Comme spécifié sans la section 1.2, on regarde les représentations mentales de $\mathcal{A}$ comme des ensembles de caractéristiques positives et négatives, c'est à dire que pour toute $r \in \mathcal{R}$, $r \subseteq \mathcal{F}$. De plus, puisque $\mathcal{R}$ ne contient que les représentations mentales les plus précises de $\mathcal{A}$, on a que pour toutes r et r' de $\mathcal{R}$, $r \not\subset r'$ (où $\subset$ dénote l'inclusion stricte). Par ailleurs, on suppose que les représentations mentales de $\mathcal{A}$ sont consistantes, c'est à dire que pour toute $r \in \mathcal{R}$ et toute $f \in \mathcal{F}^+$, $\{f, \textit{non-}f\} \not\subseteq r$. Enfin, il est clair que $\mathcal{R}$ est fini, puisque $\mathcal{F}^+$, et donc aussi $\mathcal{F}$, le sont.

Soit $\rho' : \mathcal{R} \longrightarrow \mathcal{W}_{\boldsymbol{L}}^{\mathrm{p}}$ la fonction qui à chaque $r \in \mathcal{R}$ associe le monde partiel w tel que $\delta(\mathrm{w}) = \bigwedge \rho(f)/f \in r$. ρ' définit une relation de représentation entre les mondes partiels pour $\boldsymbol{L}$ et les représentations mentales les plus précises de $\mathcal{A}$. Plus précisément, un monde partiel $\mathrm{w} \in \mathcal{W}_{\boldsymbol{L}}^{\mathrm{p}}$ représentera r dans le modèle si et seulement si $\mathrm{w} = \rho'(r)$.

Pour toute r de $\mathcal{R}$, on suppose qu'une mesure de la prégnance de r dans l'es-

prit de $\mathcal{A}$ est disponible, et qu'elle peut être exprimée au moyen d'un nombre réel appartenant à l'intervalle $]\,0,\,1\,[$ (voir p. 144 ci-dessus). On définit une fonction $v : \mathcal{R} \longrightarrow]\,0,\,1\,[$ qui à chaque $r \in \mathcal{R}$ associe la mesure de sa prégnance. On utilise l'intervalle réel ouvert car la prégnance d'une représentation mentale ne peut pas être arbitrairement élevée, bien qu'aucune valeur maximale ne puisse être fixée.

On définit $\mathcal{U} = \{\mathrm{w} \in \mathcal{W}_L^p / \exists\ r \in \mathcal{R}$ t.q. $\mathrm{w} = \rho'(r)\}$. Il est immédiat que $\mathcal{U}$ est fini et sans précisifications. $\mathcal{U}$ représentera $\mathcal{R}$ dans le modèle. Étant donné un monde partiel w, le fait que $\mathrm{w} \in \mathcal{U}$ signifie que $\mathcal{A}$ connait une classe d'objets/situations tels que $r = \rho'^{-1}(\mathrm{w})$ est la représentation mentale de cet/ces objet(s)/situation(s) dans son esprit. Étant donné un littéral λ de $\boldsymbol{L}$, le fait que $\mathrm{w} \Vdash \lambda$ signifie que, selon cette représentation, (c'est à dire, selon $\mathcal{A}$), les objet(s)/situation(s) correspondant(s) satisfait/ont la caractéristique $f = \rho^{-1}(\lambda)$.

Comme on l'a dit précédemment, les réseaux de neurones biologiques sont sujets au bruit, et il est probable que les petites différences de prégnance entre représentations sont masquées par ce bruit. De ce fait, une petite différence de prégnance ne suffira sans doute pas à permettre à une représentation d'en supplanter une autre dans la compétition qui les oppose pour la réactivation. Il s'ensuit que des représentations de prégnance similaire seront également rappelées — ou au contraire, également inhibées par une autre représentation plus prégnante qu'elles. Pour rendre compte de ce phénomène, on va utiliser une fonction d'arrondi $rd :]\,0,\,1\,[\longrightarrow]\,0,\,1\,[$ afin de gommer les petites différences de prégnance. Disons que rd est la fonction qui arrondi à la n-ième décimale (mais d'autres fonctions d'arrondi pourraient tout aussi bien convenir).

On définit ensuite une relation binaire $<$ sur $\mathcal{U}$ par :

Pour tous w et w' de $\mathcal{U}$, $\mathrm{w} < \mathrm{w'}$ ssi $rd(v(\rho'^{-1}(\mathrm{w}))) >_{\mathbb{R}} rd(v(\rho'^{-1}(\mathrm{w'})))$

où $>_{\mathbb{R}}$ est l'ordre habituel sur $]\,0,\,1\,[$. Autrement dit, $\mathrm{w} < \mathrm{w'}$ si et seulement si la représentation mentale figurée par w est 'suffisamment plus prégnante'

que celle figurée par w' . Il est immédiat que $<$ est un ordre modulaire, donc $\mathcal{M} = (\mathcal{U}, <)$ est un modèle à monde partiels fini rangé sans précisifications. $\mathcal{M}$ peut être vu comme représentant la vision du monde de l'agent, ou, selon selon le point de vue qu'on prend, comme représentant son système inférentiel.

Prenant $\boldsymbol{L}$ pour base, on définit $\boldsymbol{L}^{\|}$ comme indiqué dans la section 4.1, le sens du connecteur $\|$ étant celui suggéré dans cette même section. La relation d'inférence $\|\!\sim_{\mathcal{M}}$ induite par $\mathcal{M}$ sur $\boldsymbol{L}^{\|}$ représente l'ensemble des dispositions à inférer de $\mathcal{A}$, c'est à dire son savoir général sur les choses. Plus précisément, si α et β sont des $\boldsymbol{L}^{\|}$-formules, le fait que $\alpha \|\!\sim_{\mathcal{M}} \beta$ représente le fait que $\mathcal{A}$ est disposé à inférer β de α, au sens défini p. 144 ci-dessus. Comme on l'a montré dans la section 5, $\|\!\sim_{\mathcal{M}}$ est close sous les règles de PW $\cup$ $\{$*Rankedness*$\}$. Ces règles sont des propriétés de la relation d'inférence $\|\!\sim_{\mathcal{M}}$, et doivent donc être interprétées comme des règles qui, selon le modèle, structurent le savoir général de l'agent sur les choses. Elles sont une conséquence du processus supposé, et s'imposeront donc de fait dans tout cerveau exécutant un processus de ce type.

6.2 Modélisation de l'apprentissage

Passons maintenant à la modélisation de l'apprentissage. D'une manière générale, l'apprentissage peut être défini comme une révision du savoir d'un agent, visant à intégrer à ce savoir de l'information nouvellement acquise. Dans le contexte des inférences automatiques, l'apprentissage peut se ramener à une révision des dispositions de l'agent à opérer des inférences automatiques, c'est à dire, à une révision de la relation $\|\!\sim_{\mathcal{M}}$. En pratique, cela signifie que l'information entrant dans le système inférentiel de l'agent déclenche à la fois l'opération d'inférences automatiques conformément à ses dispositions actuelles, et une modification subséquente de ces mêmes dispositions. Cette dernière peut prendre différentes formes, selon la manière dont l'information entrante s'in-

tègre aux connaissances actuelles de l'agent. On va commencer ici par examiner les différentes formes d'apprentissage, et essayer pour chacune d'elles d'identifier le processus qui la supporte, d'abord au niveau mental de description puis, quoique de façon plus sommaire, au niveau neural. On proposera ensuite une modélisation formelle des processus envisagés.

6.2.1 Les types différents d'apprentissage

Relativement aux inférences automatiques, l'apprentissage peut se diviser en deux grandes catégories, selon que le contenu de l'information entrante s'inscrit ou non dans les représentations mentales actuelles de l'agent. S'il s'y inscrit, alors ce contenu apparait à un certain degré comme familier à l'agent, et donne lieu à un type d'apprentissage qu'on qualifiera d'*induit par la répétition*. Cet apprentissage consiste essentiellement en un renforcement du souvenir par répétition de l'expérience. Si au contraire ce contenu ne s'inscrit pas dans les représentations mentales actuelles de l'agent, alors il lui apparait comme nouveau, ce qui provoque son étonnement et déclenche un type d'apprentissage qu'on appellera *induit par la nouveauté*. On examine successivement ces deux cas.

Apprentissage induit par la répétition

L'apprentissage induit par la répétition se produit lorsque l'expérience répétée d'un même objet ou d'une même situation par un agent entraîne le renforcement de la prégnance de la représentation correspondante dans son esprit. Par exemple, supposons que l'agent sait qu'il existe des cygnes blancs et des cygnes noirs, mais que dans son environnement les cygnes blancs sont nettement plus nombreux que les noirs, si bien qu'il rencontre des cygnes blancs bien plus souvent que des noirs. Il est probable que lorsqu'il pense à des cygnes, ce sont les cygnes blancs qui lui viennent à l'esprit en premier, et qu'il va par conséquent être disposé à inférer 'blanc' de 'cygne'. Mais supposons que le nombre de cygnes noirs augmente peu à peu, et qu'avec le temps les cygnes noirs fi-

nissent par devenir plus nombreux que les blancs. On peut s'attendre à ce que la représentation que l'agent a des cygnes noirs devienne de plus en plus prégnante dans son esprit, jusqu'à devenir plus prégnante que sa représentation des cygnes blancs. De ce fait, sa disposition à inférer 'blanc' de 'cygne' devrait progressivement disparaître, et être finalement remplacée par une disposition à inférer 'noir' de 'cygne'.

Ce type d'apprentissage repose principalement sur le fait plus général que la réactivation d'une représentation (c'est à dire son rappel) dans l'esprit d'un agent en renforce la prégnance, tandis que les représentations non réactivées tendent au contraire à perdre progressivement de leur force. Au niveau neuronal, ceci résulte sans doute des phénomènes de potentialisation à long terme (PLT) et de dépression à long terme (DLT) décrits plus haut [2], et dont il est généralement admis qu'ils sous-tendent un certain nombre de fonctions liées à la mémoire [3]. Plus précisément, on pourrait imaginer que la réactivation d'une assemblée de neurones puisse, par le biais de la PLT, renforcer les connexions entre les neurones qui la composent, rendant ainsi l'assemblée dans son ensemble plus résistante au bruit neuronal et à l'inhibition. Elle pourrait aussi amener l'assemblée à s'agrandir, en lui permettant d'agréger par PLT et DLT un certain nombre de neurones accidentellement actifs, d'une manière assez similaire à celle suggérée p. 134 ci-dessus (ce qui, à nouveau, rendrait l'assemblée plus résistante au bruit et à l'inhibition). Enfin, la réactivation pourrait renforcer les connexions entre les neurones émetteurs du signal et ceux qui font partie de l'assemblée, rendant celle-ci plus sensible au signal. La diminution de la prégnance des représentations non-réactivées pourrait quant à elle découler de la perte, par les assemblées neuronales correspondantes, des neurones qui sont capturés par les assemblées réactivées, et/ou de l'affaiblissement des connexions entre les neurones émetteurs du signal et ceux qui composent l'assemblée.

2 Voir pp. 132 et 134 Ci-dessus.

3 Voir par exemple [Izquierdo, 1993].

Il faut cependant souligner que la fréquence avec laquelle un agent rencontre un objet ou une situation donnée n'est pas le seul critère qui détermine la prégnance de sa représentation dans son esprit. L'intérêt que porte l'agent à cette objet ou à cette situation, c'est à dire le poids émotionnel du contenu informationnel correspondant, est aussi un facteur déterminant. Il est bien connu en effet que les contenus à forte valeur émotionnelle (que celle-ci soit positive ou négative) sont mieux mémorisés et moins sensibles à l'oubli que ceux à plus faible valeur émotionnelle [4]. Chez les mammifères, cette observation psychologique est corroborée au niveau neuronal par le fait bien attesté que l'amygdale, qui est connue pour être un centre de traitement des émotions [5], joue un rôle prépondérant dans les processus de mémorisation en modulant l'action de la PLT, et ceci notamment dans le lobe médial temporal [6]. Il est probable que des mécanismes similaires existent dans les autres groupes taxinomiques, étant donné qu'on retrouve des systèmes d'assignation de valeur dans tout le règne animal, et jusque chez les arthropodes [7]. Du point de vue de l'évolution, ceci fait sens si on considère que les erreurs de prédiction peuvent avoir des coûts très variables pour les agents naturels, et que l'agent le mieux adapté à son environnement n'est pas celui qui fait le moins d'erreurs (c'est à dire, dont les prédictions suivent au mieux les probabilités mathématiques), mais celui qui évite le mieux les erreurs les plus coûteuses, même si cela implique pour lui de faire globalement davantage d'erreurs de prédiction. Par exemple, considérons un oiseau qui picore des graines au sol tandis qu'un chat se repose à proximité. Chaque fois que le chat bouge, l'oiseau doit deviner s'il va ou non bondir vers lui pour essayer de l'attraper. Pour l'oiseau, le coût représenté par le fait de s'imaginer à tort que le chat va bondir est assez faible, puisque cela le pousse

4 Voir par exemple [Cahill and McGaugh, 1995].

5 Voir pp. 129 et 132 ci-dessus.

6 Voir par exemple [McGaugh, 2000] et [Phelps and LeDoux, 2005] p. 177.

7 Voir [Giurfa, 2007] pp. 811-812 pour l'existence de neurones encodant des valeurs dans le cerveau de l'abeille domestique, ainsi que leur rôle déterminant dans l'apprentissage.

juste à s'envoler inutilement. Par contre, le coût représenté par le fait de croire à tort que le chat ne va pas bondir est extrêmement élevé, puisque cela pourrait lui être fatal. Par conséquent, même si le chat ne bondit en réalité qu'une fois sur cent, les chances de survie de l'oiseau sont meilleures s'il s'attend à ce que le chat bondisse que s'il s'attend à ce qu'il ne bondisse pas. Mais puisque la situation dans laquelle le chat bondit est critique pour la survie de l'oiseau, on peut supposer qu'elle a dans son esprit une valeur émotionnelle très élevée, et en fait bien plus élevée que la situation dans laquelle le chat se repose, qui n'est pas aussi critique. Puisque les contenus à plus forte valeur émotionnelle sont mieux mémorisés, il est probable que, même si en réalité l'oiseau observe le chat se reposer bien plus souvent qu'il ne l'observe bondir, sa représentation mentale de la situation dans laquelle le chat bondit sera plus prégnante que sa représentation de la situation dans laquelle le chat se repose, amenant l'oiseau à s'attendre à ce que le chat bondisse et augmentant ainsi ses chances de survie. La modulation de la mémorisation en fonction de la valeur (ou 'intérêt') assignée par l'agent aux objets et situations rencontrés semble donc apporter un avantage adaptatif significatif[8]. Comme il est probable que des systèmes d'attribution de valeur existent chez toutes les espèces pourvues d'un cerveau, on peut supposer qu'une telle modulation de la mémorisation existe, au moins dans une certaine mesure, chez la plupart de ces espèces sinon toutes.

Apprentissage induit par la nouveauté

L'apprentissage induit par la nouveauté se présente sous plusieurs formes. La première est celle où l'agent apprend l'existence d'une certaine classe d'objets ou de situations. Par exemple, supposons que 'cygne', 'blanc' et 'noir' sont des caractéristiques[9] que l'agent est physiologiquement capable de percevoir, et

8 On peut trouver des idées similaires chez [McGaugh, 2000] p. 248 et [Phelps and LeDoux, 2005] p. 177.

9 À nouveau pour simplifier, on regarde ici 'cygne' comme une caractéristique, bien qu'en réalité ce soit plus certainement une conjonction de nombreuses caractéristiques (voir la note n°42 page 143 ci-dessus).

supposons qu'il sait qu'il existe des cygnes blancs, mais ignore qu'il en existe aussi des noirs. Selon notre analyse, cela signifie qu'il possède une représentation mentale de cygne blanc, mais pas de représentation mentale de cygne noir. En d'autres termes, cela signifie que l'une (au moins) de ses représentations mentales satisfait à la fois les caractéristiques 'cygne' et 'blanc', mais aucune ne satisfait à la fois 'cygne' et 'noir'. Supposons qu'à présent l'agent aperçoit un cygne noir, c'est à dire que son système perceptif détecte les caractéristiques co-occurrentes 'cygne' et 'noir' et transmet cette information à son système inférentiel. Il est clair que cette information ne s'inscrit dans aucune de ses représentations mentales, au sens où aucune d'entre elles ne satisfait ces deux caractéristiques. Cette discordance entre son observation présente et sa représentation du monde est ce qu'en langage courant on appelle 'l'étonnement'. Mais bientôt l'agent revient de sa surprise, et admet que les cygnes noirs existent. C'est à dire qu'il corrige sa vision du monde en y ajoutant une représentation de cygne noir, afin de la rendre consistante avec l'information collectée. L'étonnement joue ici le rôle d'un signal d'erreur qui déclenche le processus de révision. La prégnance de cette nouvelle représentation dans l'esprit de l'agent dépendra principalement de son intérêt pour les cygnes noirs. Plus celui-ci sera fort, et plus la prégnance de la nouvelle représentation dans son esprit sera élevée. Cet intérêt quant à lui dépendra vraisemblablement de son intérêt préalable pour les (conjonctions de) caractéristiques qui composent cette nouvelle représentation, ici les cygnes et les choses noires. Mais nous n'allons pas étudier davantage ici la question l'intérêt chez les agents naturels, car cela nous entraînerait trop loin de nos préoccupations présentes.

Une deuxième forme d'apprentissage induit par la nouveauté est celle où l'agent complète sa connaissance d'une classe donnée d'objets ou de situations avec des informations supplémentaires. Par exemple, supposons que l'agent connait les cygnes blancs mais n'a jamais eu l'occasion d'observer leurs pattes, et n'a par conséquent aucune idée de la couleur de celles-ci. De ce fait, sa représentation mentale des cygnes blancs ne dit rien quant à la couleur de leurs

pattes, c'est à dire qu'elle ne satisfait ni ne falsifie aucune des caractéristiques relatives à la couleur des pattes que l'agent est physiologiquement capable de percevoir. Supposons qu'à présent il aperçoit un cygne blanc marchant au bord de la rivière, et observe que ses pattes sont en fait noires. À nouveau, le contenu de cette observation ne s'inscrit dans aucune de ses représentations mentales actuelles, puisqu'aucune d'entre elles ne satisfait à la fois les caractéristiques 'cygne', 'blanc' et 'pattes noires'. Ceci provoque donc à nouveau son étonnement, quoique peut-être un étonnement d'une qualité un peu différente, et peut-être aussi moins fort que dans le cas précédent. Et à nouveau, cet étonnement déclenche la révision de sa vision du monde, afin de la mettre en conformité avec l'information collectée. Mais cette fois, la révision consiste à compléter sa connaissance des cygnes blancs en ajoutant la caractéristique 'pattes noires' à la représentation qu'il s'en fait. Dans ce cas, la prégnance de la représentation ainsi complétée ne dépendra pas seulement de l'intérêt que l'agent porte à l'information entrante, mais aussi de la prégnance préalable de sa représentation des cygnes blancs. En effet, puisque l'observation d'un cygne blanc implique la reconnaissance de l'objet considéré comme étant un cygne blanc, et donc la réactivation de la représentation mentale des cygnes blancs de l'agent, il semble que si l'agent est peu intéressé par le contenu de l'observation, la prégnance de la représentation supplémentée devrait être celle de la représentation non-supplémentée, augmentée du gain de prégnance dû à sa réactivation, comme dans le cas de l'apprentissage induit par répétition. Mais si l'intérêt de l'agent pour le contenu observé est suffisamment fort, il semble qu'il va influer sur la prégnance de la représentation supplémentée et que celle-ci dépendra alors principalement de la force de cet intérêt, comme dans le premier cas d'apprentissage induit par la nouveauté.

Une variante de ce cas de complémentation est la fusion de plusieurs représentations en une seule. Par exemple, supposons que l'agent a déjà rencontré des merles, mais qu'il ignore quel est leur chant ; et supposons que par ailleurs qu'il a déjà entendu un certain chant d'oiseau, mais qu'il ignore quel est l'oi-

seau qui chante ainsi. Il a donc deux représentations distinctes, l'une des merles et l'autre de ce type de chant particulier. Supposons qu'à présent il observe un merle chantant de cette façon. À nouveau, ceci provoque son étonnement, puisqu'aucune de ses représentations mentales ne satisfait à la fois les caractéristiques qui composent sa représentation des merles et celles qui composent sa représentation du chant en question ; et à nouveau cet étonnement déclenche un processus d'apprentissage. Mais cette fois, l'agent fusionne ses précédentes représentations en une seule. Il s'agit bien d'un cas de complémentation, puisque ce faisant l'agent ajoute en même temps les caractéristiques qui composent sa représentation du chant à sa représentation des merles, et les caractéristiques qui composent sa représentation des merles à sa représentation du chant. Dans ce cas, on peut supposer que si l'intérêt de l'agent pour l'information entrante n'est pas assez fort pour influer sur la prégnance de la nouvelle représentation, alors celle-ci dépendra principalement de la prégnance de la plus prégnante des représentations d'origine, c'est à dire que la prégnance de la nouvelle représentation sera égale à celle de la plus prégnante des représentations d'origine, augmentée du gain de prégnance résultant de sa réactivation. Si au contraire l'intérêt de l'agent pour l'information entrante est suffisamment fort, alors la prégnance de la nouvelle représentation devrait dépendre comme précédemment de la force de l'intérêt de l'agent pour cette information.

Dans tous les cas, la prégnance des autres représentations mentales de l'agent (c'est à dire de celles qui n'ont pas été réactivées par l'observation) devrait diminuer légèrement, à cause de l'oubli progressif qui affecte les représentations non-réactivées. Il faut cependant noter que toutes les représentations non-réactivées ne devraient pas être également affectées, et que celles dont le contenu est d'un intérêt plus grand pour l'agent devraient mieux résister à l'oubli que celles dont le contenu est d'un intérêt plus faible.

Du point de vue neuronal, la complémentation d'une représentation mentale pourrait simplement résulter du mécanisme de conjonction de caractéristiques décrit pages 132–132 ci-dessus. Plus précisément, les neurones qui

supportent la représentation à complémenter pourraient simplement ajouter les nouvelles caractéristiques à celles dont ils supportaient déjà la conjonction. La création d'une nouvelle représentation suppose quant à elle la formation d'une nouvelle assemblée de neurones-concept. Ceci pourrait résulter d'une réassignation d'un certain nombre de neurones issus d'autres assemblées, dont les connexions pourraient être rapidement modifiées par PLT et DLT. Plus précisément, on pourrait imaginer que certains neurones supportant des concepts plus généraux qui subsument la nouvelle représentation à créer pourraient prendre en charge la conjonction des caractéristiques actuellement observées, exactement comme dans le cas de la supplémentation. Une autre possibilité, qui n'est pas incompatible avec la précédente, est qu'un certain nombre de neurones supportant des représentations assez similaires à celle qui est en formation pourraient être capturés par la nouvelle assemblée. En effet, puisque le signal auquel ils sont câblés pour répondre est assez similaire à celui qui est à l'origine de la nouvelle représentation, il se pourrait qu'un certain nombre d'entre eux s'activent accidentellement sous l'effet du bruit neuronal. L'action combinée de la PLT et de la DLT pourrait alors, par un mécanisme comparable à celui suggéré page 134, modifier les connexions de ces neurones de manière à qu'ils répondent désormais au nouveau signal. La différence entre le cas présent et celui considéré page 134 tiendrait dans l'intérêt de l'agent pour le contenu représenté par le signal. Si cet intérêt est suffisamment fort, alors la nouvelle représentation pourrait parvenir à se consolider, mais sinon elle se désagrégerait rapidement comme évoqué page 134.

Il faut cependant préciser que toutes les représentations d'un agent ne sont pas nécessairement acquises par l'apprentissage. Il est probable au contraire que les agents naturels naissent avec un certain nombre de représentations innées, qui sont ensuite ou bien utilisées telles quelles, ou bien complémentées. Chez les espèces dans lesquelles les jeunes naissent avec des cerveaux immatures comme chez l'humain, ces représentations innées sont certainement très rudimentaires et très peu nombreuses, et largement remaniées par la suite par

l'apprentissage. Mais il devrait néanmoins en exister, comme par exemple celle qui permet à un mammifère nouveau-né de reconnaitre une mamelle à laquelle se nourrir. Chez les espèces dans lesquelles les jeunes naissent plus autonomes, celles-ci pourraient à l'inverse être plus nombreuses et plus élaborées, et former le substrat de ce qu'on appelle en langage courant 'l'instinct'.

6.2.2 Modélisation de l'apprentissage

Les processus d'apprentissage décrits ci-dessus peuvent tous se ramener à des modifications de l'ensemble $\mathcal{R}$ des représentations mentales les plus précises de l'agent, et/ou à des modifications de la prégnance $v(\mathsf{r})$ de ces mêmes représentations. Par conséquent ils peuvent tous être représentés dans le dispositif logique proposé par les modifications correspondantes du modèle à mondes partiels $\mathcal{M} = (\mathcal{U}, <)$ qui représente la vision du monde de l'agent. Dans ce qui suit on donne une version formelle de ces processus d'apprentissages, et on montre comment ils induisent une révision du modèle $\mathcal{M}$.

Soient $\mathcal{A}$, $\mathcal{F}^+$, $\mathcal{F}^-$, $\mathcal{F}$, σ, $\boldsymbol{L}$, ρ et rd tels que définis dans la section 6.1, et soit $T = \langle t_1, ..., t_n \rangle$ une suite d'instants. Étant donné un objet/situation $\mathfrak{s}$ et un instant $t_i \in T$ tel que $\mathcal{A}$ observe $\mathfrak{s}$ à l'instant t_i, $\{f \in \mathcal{F} /\ \mathcal{A}$ observe f à l'instant $t_i\}$ est appelé l'*observation totale (de $\mathfrak{s}$) par $\mathcal{A}$ à l'instant t_i*. Il faut souligner qu'une observation totale ne contient pas nécessairement toutes les caractéristiques que $\mathcal{A}$ pourrait en principe observer étant donnée $\mathfrak{s}$, mais seulement celles qu'il remarque effectivement à cet instant précis [10].

Soit $\mathcal{O}$ l'ensemble des $\boldsymbol{L}$-formules o telles que :

i) o est un littéral ou une conjonction de littéraux,

ii) si f et f' sont des caractéristiques mutuellement exclusives (au sens indiqué p. 140 ci-dessus) et $\lambda_1 = \rho(f)$ figure dans o, alors $\lambda_2 = \neg\rho(f')$ figure aussi dans o,

iii) $o \not\equiv \bot$.

10 Cette notion est librement inspirée du travail de Leitgeb [Leitgeb, 2004].

$\mathcal{O}$ est l'ensemble des $\boldsymbol{L}$-formules qui représentent le contenu informationnel des observations totales théoriquement possibles de $\mathcal{A}$.

Étant donnée une situation $\mathfrak{s}$ considérée par $\mathcal{A}$ à l'instant t_i, on note o_i le contenu de l'observation totale correspondante. o_i est à interpréter comme l'information envoyée par les aires sensorielles du cerveau de l'agent à son système inférentiel à l'instant t_i. On note $\mathcal{F}_{o_i}$ l'observation totale correspondante, c'est à dire que $\mathcal{F}_{o_i} = \{f \in \mathcal{F} / \rho(f)$ figure dans $o_i\} = \{f \in \mathcal{F} /$ $\mathcal{A}$ observe f à l'instant $t_i\}$. Remarquons qu'il n'y a pas besoin de supposer que pour chaque instant t_i de T il existe une observation $\mathcal{F}_{o_i}$ correspondante, car un agent n'est pas forcément toujours attentif à ses perceptions.

Pour chaque $o \in \mathcal{O}$, on suppose qu'une mesure est disponible de combien $\mathcal{A}$ serait intéressé par o s'il venait à l'observer, et que cette mesure est donnée par un nombre réel appartenant à l'intervalle $]\,0,\,1\,[$. Pour des raisons de simplicité, on suppose aussi que cette mesure ne varie pas au cours du temps. Soit $\mathcal{I} : \mathcal{O} \longrightarrow]\,0,\,1\,[$ la fonction qui à chaque $o \in \mathcal{O}$, associe la mesure de l'intérêt de $\mathcal{A}$ pour o.

Pour chaque instant $t_i \in T$, on notera $\mathcal{R}_i$ l'ensemble des représentations mentales les plus précises de $\mathcal{A}$ à l'instant t_i, et v_i la fonction qui à chaque $r \in \mathcal{R}_i$ associe la mesure de sa prégnance dans l'esprit de $\mathcal{A}$ à ce même instant, comme spécifié p. 189 ci-dessus.

Supposons que pour quelque $i \in \{\,1,\,...,\,n\,\}$ nous connaissions $\mathcal{R}_i$ et v_i. Suivant les indications fournies dans la section 6.1, on peut définir :

- $\rho'_i : \mathcal{R}_i \longrightarrow \mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}}$, la fonction qui à chaque $r \in \mathcal{R}_i$ associe le monde partiel w tel que $\delta(\mathrm{w}) = \bigwedge \rho(f)/f \in r$,
- $\mathcal{U}_i = \{\mathrm{w} \in \mathcal{W}^{\mathrm{P}}_{\boldsymbol{L}} / \exists\ r \in \mathcal{R}_i$ t.q. $\mathrm{w} = \rho'_i(r)\}$,
- $<_i$, la relation binaire sur $\mathcal{U}_i$ telle que pour tous w et w' de $\mathcal{U}_i$, w $<_i$ w' ssi $rd(v_i({\rho'_i}^{-1}(\mathrm{w}))) >_{\mathbb{R}} rd(v_i({\rho'_i}^{-1}(\mathrm{w}')))$,
- $\mathcal{M}_i = (\mathcal{U}_i, <_i)$.

$\mathcal{M}_i$ est un modèle à mondes partiels fini rangé sans précisifications, et le modèle de la vision du monde de $\mathcal{A}$ à l'instant t_i, au sens de la section 6.1.

Supposons maintenant qu'à l'instant t_i, $\mathcal{A}$ considère une situation $\mathfrak{s}$ et fait l'observation totale $\mathcal{F}_{o_i}$. D'après l'analyse proposée dans la section 6.2.1, il peut en résulter deux types d'apprentissage différents, selon qu'il existe ou non une représentation mentale $r \in \mathcal{R}_i$ telle que r satisfait le contenu observé o_i. On étudie successivement ces deux cas. Pour la clarté de l'exposition, on les aborde ici dans l'ordre inverse de celui adopté dans la section 6.2.1.

1. *Apprentissage induit par la nouveauté*

Si pour toute $r \in \mathcal{R}_i$, $\mathcal{F}_{o_i} \not\subseteq r$, c'est à dire s'il n'y a aucune représentation mentale dans l'esprit de $\mathcal{A}$ qui satisfait o_i, alors le contenu o_i est nouveau pour lui. Ceci est représenté dans le modèle $\mathcal{M}_i$ par le fait qu'aucun monde partiel de $\mathcal{U}_i$ ne satisfait o_i. Du côté syntaxique, on a $o_i \parallel\!\!\!\sim_{\overline{\mathcal{U}_i, \mathbf{L}}} \bot$, ce qui par *Supra-$\mathcal{U}$-Consequence* implique $o_i \parallel\!\!\!\sim_{\mathcal{M}_i} \bot$. Ceci est à interpréter comme le fait que $\mathcal{A}$ est étonné, puisque cela signifie que le contenu de son observation actuelle contredit son savoir sur les choses, l'amenant à conclure que 'quelque chose ne va pas'. Cette inadéquation entre l'observation actuelle de $\mathcal{A}$ et sa vision du monde déclenche un processus d'apprentissage.

Comme on l'a vu dans la section 6.2.1, celui-ci consiste essentiellement en l'ajout d'une nouvelle représentation $r' = \mathcal{F}_{o_i}$ à $\mathcal{R}_i$, afin de former l'ensemble $\mathcal{R}_{i+1}$ des représentations mentales les plus précises de $\mathcal{A}$ à l'instant t_{i+1}. Mais ici il nous faut à nouveau distinguer deux cas :

- Si pour toute r de $\mathcal{R}_i$, $r \not\subseteq r'$, c'est à dire, si la nouvelle représentation à ajouter ne complémente aucune des représentations mentales pré-existantes de $\mathcal{A}$, alors nous sommes dans le premier cas d'apprentissage induit par la nouveauté décrit ci-dessus (p. 194), et r' est simplement ajoutée à $\mathcal{R}_i$. Autrement dit, $\mathcal{R}_{i+1} = \mathcal{R}_i \cup \{r'\}$.
- Si au contraire il existe r_1, ... , $r_k \in \mathcal{R}_i$ tels que pour tout

j ($1 \leq_{\mathbb{N}} j \leq_{\mathbb{N}} k$), $r_j \subseteq r'$, c'est à dire si la nouvelle représentation r' à ajouter complémente une ou plusieurs représentations mentales pré-existantes de $\mathcal{A}$, alors nous sommes dans le cas d'apprentissage induit par la nouveauté dit de complémentation (voir p. 195). Dans ce cas, la nouvelle représentation r' remplace dans $\mathcal{R}_{i+1}$ toutes les représentations mentales qu'elle complémente, c'est à dire toutes les représentations $r \in \mathcal{R}_i$ telles que $r \subseteq r'$.

En rassemblant ces deux cas, on obtient que $\mathcal{R}_{i+1} = (\mathcal{R}_i - \{r \in \mathcal{R}_i \ / \ r \subseteq r'\}) \cup \{r'\}$.

Passons maintenant à la question de la prégnance des éléments de $\mathcal{R}_{i+1}$ dans l'esprit de $\mathcal{A}$ à l'instant t_{i+1}. Selon l'analyse proposée dans la section 6.2.1, la prégnance de la nouvelle représentation r' à l'instant t_{i+1} devrait dépendre de $\mathcal{I}(o_i)$, et aussi, s'il en existe, de la prégnance des représentations de $\mathcal{R}_i$ que r' complémente. Soit $\{r_1, \ldots, r_k\}$ l'ensemble (possiblement vide) de ces représentations, c'est à dire que $\{r_1, \ldots, r_k\} = \{r \in \mathcal{R}_i \ / \ r \subseteq r'\}$. De plus, soit $x = max(\{0, v_i(r_1), \ldots, v_i(r_k)\})$, et soit $\epsilon \in]0, 1[$ le taux d'apprentissage de l'agent [11]. On va supposer que $v_{i+1}(r') = max(\{\mathcal{I}(o_i), x + \epsilon.\mathcal{I}(o_i)(1 - x)\})$. Ce choix peut s'expliquer comme suit :

- Si $\{r_1, \ldots, r_k\} = \varnothing$, alors $x = 0$ et $v_{i+1}(r') = max(\{\mathcal{I}(o_i), \epsilon.\mathcal{I}(o_i)\})$, donc puisque $\epsilon <_{\mathbb{R}} 1$, $v_{i+1}(r') = \mathcal{I}(o_i)$. Autrement dit, lorsque la nouvelle représentation r' ne complémente aucune des représentations mentales pré-existantes de $\mathcal{A}$ (ce qui correspond au premier

11 La notion de taux d'apprentissage est issue de la modélisation neuronale et des réseaux de neurones artificiels, où un taux d'apprentissage est un très petit nombre réel qui contrôle la fonction d'actualisation par laquelle les poids des connexions entre les neurones d'un réseau sont révisés au cours du temps. Dans ce contexte, un taux d'apprentissage est conçu comme rendant compte du taux selon lequel les connexions inter-neuronales d'un agent se renforcent sous l'effet d'un mécanisme de plasticité synaptique, comme par exemple la potentialisation à long terme évoquée plus haut. Plus généralement, le taux d'apprentissage d'un agent peut être vu comme une mesure de la labilité de son savoir : plus ce taux est élevé, et plus l'agent est sensible à l'information entrante. C'est dans ce sens plus général qu'on utilise ici ce terme.

cas d'apprentissage induit par la nouveauté examiné dans la section 6.2.1), on suppose que la prégnance de r' dépend seulement de l'intérêt que $\mathcal{A}$ porte à o_i.

- Si $\{r_1, \ldots, r_k\} \neq \varnothing$, alors $v_{i+1}(r') = \mathcal{I}(o_i)$ si $\mathcal{I}(o_i) \geq_{\mathbb{R}} x + \epsilon.\mathcal{I}(o_i)(1-x)$, et $v_{i+1}(r') = x + \epsilon.\mathcal{I}(o_i)(1-x)\}$ sinon. C'est à dire que dans le cas dit de complémentation, on suppose que si l'interêt de $\mathcal{A}$ pour o_i est 'suffisamment fort' par rapport à la prégnance de la plus prégnante des représentations mentales à complémenter, alors la prégnance de r' dépendra totalement de la force de cet intérêt, mais que dans le cas contraire elle dépendra principalement de la prégnance de la plus prégnante des représentations mentales à complémenter. Plus précisément, on suppose que dans ce dernier cas la prégnance de r' sera celle de la plus prégnante des représentations mentales à complémenter, augmentée du gain de prégnance $\epsilon.\mathcal{I}(o_i)(1-x)$ qui résulte de la réactivation de cette dernière. On suppose de plus que ce gain de prégnance dépend d'une part du taux d'apprentissage ϵ de l'agent, et d'autre part de son intérêt $\mathcal{I}(o_i)$ pour le contenu appris. Le facteur $(1-x)$ quant à lui rend compte du fait que la prégnance d'une représentation ne peut pas croître indéfiniment, en faisant progressivement décroître $\epsilon.\mathcal{I}(o_i)(1-x)$ à mesure que x grandit. Comme $\epsilon.\mathcal{I}(o_i) <_{\mathbb{R}} 1$, on aura toujours $\epsilon.\mathcal{I}(o_i)(1-x) <_{\mathbb{R}} (1-x)$. Il s'ensuit que $v_{i+1}(r')$ tend vers 1 lorsque x grandit, et donc $v_{i+1}(r')$ appartiendra toujours à l'intervalle $]\,0,\,1\,[$.

Pour ce qui est des autres représentations mentales qui font partie de $\mathcal{R}_{i+1}$, on peut supposer qu'elles n'ont pas été réactivées par l'observation $\mathcal{F}_{o_i}$, puisqu'elles ne satisfont pas o_i. Soit $\epsilon' \in]\,0,\,1\,[$ le 'taux d'oubli' de l'agent, c'est à dire une mesure de la tendance générale de ses représentations mentales à perdre de leur prégnance lorsqu'elles ne sont pas réactivées. Suivant l'analyse proposée dans la section 6.2.1, on va sup-

poser que pour toute $r \in \mathcal{R}_{i+1} - \{r'\}$, $v_{i+1}(r) = v_i(r) - \epsilon'.v_i(r)(1-\mathcal{I}(o_i))$. Le facteur $(1-\mathcal{I}(o_i))$ fait décroître $\epsilon'.v_i(r)(1-\mathcal{I}(o_i))$ à mesure que $\mathcal{I}(o_i)$ grandit, rendant compte de la plus grande résistance à l'oubli des représentations mentales dont le contenu intéresse davantage l'agent. Le facteur $v_i(r)$ fait que $\epsilon'.v_i(r)(1-\mathcal{I}(o_i))$ décroît en même temps que $v_i(r)$, et que donc $v_{i+1}(r)$ tend vers 0 lorsque $v_i(r)$ décroît, ce qui garantit que $v_{i+1}(r) \in]\,0,\,1\,[$.

2. *Apprentissage induit par la répétition*

Si il existe $r \in \mathcal{R}_i$ telle que $\mathcal{F}_{o_i} \subseteq r$, alors l'agent connaît au moins un objet/situation qui, selon lui, satisfait o_i. Dans le modèle $\mathcal{M}_i$, ceci est rendu par le fait qu'il y a au moins un monde partiel w $\in \mathcal{U}_i$ tel que w $\Vdash o_i$. On a donc $o_i \;\|\!\not\sim_{\mathcal{U}_i, \boldsymbol{L}^{\Vdash}} \bot$, et par $(\boldsymbol{L}^{\Vdash})$ $\mathcal{U}$-*Consistence*, $o_i \;\|\!\not\sim_{\mathcal{M}_i} \bot$, ce qui correspond au fait que o_i ne provoque pas d'étonnement chez $\mathcal{A}$. De ce fait, le processus inférentiel peut aller à son terme, et l'agent opère les inférences automatiques auxquelles il est disposé.

De plus, puisqu'il y a au moins une représentation r dans $\mathcal{R}_i$ qui satisfait o_i, l'ensemble des objets et situations que $\mathcal{A}$ connaît n'est pas modifié par l'observation $\mathcal{F}_{o_i}$, c'est à dire que $\mathcal{R}_{i+1} = \mathcal{R}_i$.

Selon l'analyse proposée dans la section 1.1.2, le processus inférentiel déclenché par $\mathcal{F}_{o_i}$ se ramène au fait que parmi les représentations mentales de $\mathcal{A}$ appartenant à $\mathcal{R}_i$, seules les plus prégnantes de celles qui satisfont o_i sont réactivées par $\mathcal{F}_{o_i}$. Cependant il nous faut considérer ici les valeurs de prégnance arrondies plutôt que les valeurs exactes puisque, comme on l'a déjà dit, les petites différences de prégnance sont masquées par le bruit neuronal. Soit $y = max(\{rd(v_i(r))/r \in \mathcal{R}_i \text{ et } \mathcal{F}_{o_i} \subseteq r\})$, et $\mathcal{Y} = \{r \in \mathcal{R}_i \;/\; \mathcal{F}_{o_i} \subseteq r \text{ et } rd(v_i(r)) = y\}$. $\mathcal{Y}$ est l'ensemble des représentations mentales les plus précises de $\mathcal{A}$ qui, d'après notre analyse, sont réactivées à l'instant t_i par l'observation $\mathcal{F}_{o_i}$. Reprenant les estimations précédentes concernant le gain ou la perte de prégnance des représenta-

tions suite à leur réactivation ou leur non-réactivation, on va supposer que :

- Pour toute $r \in \mathcal{Y}$, $v_{i+1}(r) = v_i(r) + \epsilon.\mathcal{I}(o_i)(1 - v_i(r))$, et
- Pour toute $r \in \mathcal{R}_{i+1} - \mathcal{Y}$, $v_{i+1}(r) = v_i(r) - \epsilon'.v_i(r)(1 - \mathcal{I}(o_i))$.

À ce stade nous connaissons $\mathcal{R}_{i+1}$, aussi bien dans le cas où il existe une représentation $r \in \mathcal{R}_i$ telle que $\mathcal{F}_{o_i} \subseteq r$, que dans celui où il n'en existe pas, et dans chacun de ces cas nous avons une estimation de v_{i+1}. Nous sommes donc en capacité de définir ρ'_{i+1}, $\mathcal{U}_{i+1}$ et $<_{i+1}$, comme indiqué page 200. $\mathcal{M}_{i+1} = (\mathcal{U}_{i+1}, <_{i+1})$ est un modèle à mondes partiels fini, rangé et sans-précisifications, et le modèle de la vision du monde de $\mathcal{A}$ à l'instant t_{i+1}. Puisque suivant la procédure ci-dessus, un modèle à mondes partiels fini rangé et sans-précisifications $\mathcal{M}_i$ est toujours révisé en un autre modèle à mondes partiels fini, rangé et sans-précisifications $\mathcal{M}_{i+1}$ défini de manière unique, le processus de révision ainsi conçu est indéfiniment itérable.

Cependant, une fois que l'interprétation attendue du dispositif logique est claire, on peut souhaiter se débarrasser autant que possible des éléments non-logiques de la modélisation ci-dessus. Ceci peut se faire de la manière suivante. Supposons que pour quelque $i \in \{1, ..., n\}$ on connaisse $\mathcal{R}_i$ et v_i, et qu'on ait construit $\mathcal{M}_i = (\mathcal{U}_i, <_i)$ comme suggéré page 200. On définit une fonction $v*_i : \mathcal{U}_i \longrightarrow]0, 1[$ par : pour tout $\mathrm{w} \in \mathcal{U}_i$, $v*_i(\mathrm{w}) = v_i({\rho'}_i^{-1}(\mathrm{w}))$. Soit maintenant o_i le contenu de l'observation totale de $\mathcal{A}$ à l'instant t_i. Les deux cas d'apprentissage examinés plus haut se ramènent à :

1. Si $o_i \Vdash_{\overline{\mathcal{U}_i, \boldsymbol{L}}} \bot$:

Soit $\mathrm{w'} \in \mathcal{W}^{\mathrm{p}}_{\boldsymbol{L}}$ le monde partiel tel que $o_i = \delta(\mathrm{w'})$, et $x^* = max(\{0\} \cup \{v*_i(\mathrm{w}) / o_i \vdash \delta(\mathrm{w})\})$. On définit :

- $\mathcal{U}_{i+1} = (\mathcal{U}_i - \{\mathrm{w} \in \mathcal{U}_i / o_i \vdash \delta(\mathrm{w})\}) \cup \{\mathrm{w'}\}$,
- $v*_{i+1}(\mathrm{w'}) = max(\{\mathcal{I}(o_i), x^* + \epsilon.\mathcal{I}(o_i)(1 - x^*)\})$,
- Pour tout $\mathrm{w} \in (\mathcal{U}_{i+1} - \{\mathrm{w'}\})$, $v*_{i+1}(\mathrm{w}) = v*_i(\mathrm{w}) - \epsilon'.v*_i(\mathrm{w})(1 - \mathcal{I}(o_i))$.

2. Si $o_i \not\Vdash_{\overline{\mathcal{U}_i, \boldsymbol{L}}} \bot$, alors $\{\mathrm{w} \in \mathcal{U}_i \;/\; \mathrm{w} \Vdash\!\!= o_i\} \neq \varnothing$.

Soit $y^* = max(\{rd(v*_i(\mathrm{w})) / \mathrm{w} \in \mathcal{U}_i \text{ et } \mathrm{w} \Vdash\!\!= o_i\})$, et

$\mathcal{Y}^* = \{\text{w} \in \mathcal{U}_i \ / \ \text{w} \Vdash o_i \text{ et } rd(v*_i(\text{w})) = y^*\}$. On définit :

. $\mathcal{U}_{i+1} = \mathcal{U}_i$,

. Pour tout w $\in \mathcal{Y}^*$, $v*_{i+1}(\text{w}) = v*_i(\text{w}) + \epsilon.\mathcal{I}(o_i)(1 - v*_i(\text{w}))$,

. Pour tout w $\in (\mathcal{U}_{i+1} - \mathcal{Y}^*)$, $v*_{i+1}(\text{w}) = v*_i(\text{w}) - \epsilon'.v*_i(\text{w})(1 - \mathcal{I}(o_i))$.

Dans les deux cas, on définit $<_{i+1}$ par : pour tous w et w' $\in \mathcal{U}_{i+1}$, w $<_{i+1}$ w' ssi $rd(v*_{i+1}(\text{w})) >_{\mathbb{R}} rd(v*_{i+1}(\text{w}'))$. $\mathcal{M}_{i+1} = (\mathcal{U}_{i+1}, <_{i+1})$ est un modèle à mondes partiels fini, rangé et sans-précisifications, et le modèle de la vision du monde de $\mathcal{A}$ à l'instant t_{i+1}.

Conclusion et perspectives

Dans ce travail, on s'est inspiré des sciences cognitives pour aborder la question de la modélisation logique du raisonnement et de l'apprentissage, et on a développé un modèle bio-inspiré d'un type d'inférences très simples qu'on a appelé les *inférences automatiques*. Ce modèle repose sur un certain nombre d'hypothèses quant à la manière dont les cerveaux fonctionnent, et pour cette raison il ne peut prétendre être vrai, mais seulement être suffisamment plausible. Une meilleure connaissance de l'organisation et des processus cérébraux rendront sans doute nécessaire dans l'avenir un certain nombre de modifications du modèle proposé. Cependant les hypothèses sur lesquelles il repose semblent pour la plupart suffisamment bien établies pour laisser penser que ces ajustements ne porteront que sur des points de détail.

Par exemple, une compréhension plus fine des faits neuronaux qui sous-tendent le phénomène de prégnance des représentations mentales dans l'esprit d'un agent pourraient nous amener à revoir les formules qui définissent la fonction d'actualisation de la mesure de la prégnance dans la modélisation de l'apprentissage proposée section 6.2.2. Ou encore, une meilleure compréhension des mécanismes neuronaux qui assurent la liaison des caractéristiques entre elles et du rôle de ces mécanismes dans la formation des représentations mentales pourraient nous pousser à abandonner l'hypothèse simplificatrice que nous avons faite selon laquelle toutes les caractéristiques qui participent à une représentation donnée sont également importantes, ce qui fait que les représentations mentales peuvent être vues comme de simples conjonctions de caractéristiques. Si tel était le cas, nous ne pourrions plus figurer les représentations mentales par des mondes partiels, et il nous faudrait alors remplacer ceux-ci dans le

modèle par des structures sémantiques plus complexes. Cela impliquerait bien sûr un renouvellement complet de l'appareillage logique, mais pour autant cela ne changerait rien à l'idée générale à l'origine du modèle, ni à la dynamique globale de celui-ci.

D'autre part, il ne faut pas perdre de vue que la modélisation des inférences automatiques et des processus d'apprentissage associés qui est proposée ici n'est que le premier élément d'un projet plus général de modélisation du raisonnement et de l'apprentissage. Comme on l'a souligné au début de ce mémoire, chez les agents naturels le raisonnement est composé d'un grand nombre de facultés cognitives distinctes, parmi lesquelles la capacité à opérer des inférences automatiques est sans doute l'une des plus élémentaires. Un prolongement naturel du présent travail consisterait donc à articuler une modélisation de certaines de ces autres facultés à celle des inférences automatiques. Parmi celles qui pourraient être étudiées avec profit, la première qui vient à l'esprit est l'utilisation des concepts généraux dans le raisonnement. Par exemple, un agent qui observe un cygne noir pour la première fois mais qui connait déjà d'autres espèces d'oiseaux peut s'attendre à ce que celui-ci soit capable de voler, bien qu'il n'ait aucune expérience préalable des cygnes noirs (ceci ne peut bien sûr se faire qu'après que la surprise de l'agent soit retombée, c'est à dire, après qu'il ait admis que les cygnes noirs existent). Pour autant qu'on puisse en juger à ce stade de notre réflexion, il semble que pour ce faire l'agent recherche d'abord ses concepts généraux les plus spécifiques parmi ceux qui sont compatibles avec l'information entrante, et qu'il en extrait le contenu informationnel commun pour s'en servir comme prémisse d'inférences automatiques. La capacité à remonter ainsi à des concepts plus généraux pour raisonner est variable selon les espèces, et dépend probablement de la capacité de ces dernières à disposer d'un vaste ensemble de concepts généraux.

Un autre cas intéressant à étudier serait celui des processus de raisonnement qui mettent en jeu des séquences ordonnées de représentations mentales plutôt que des représentations uniques. Si on accepte l'hypothèse d'Eichenbaum

selon laquelle les souvenirs épisodiques sont encodés par l'hippocampe sous la forme de séquences ordonnées de représentations mentales discrètes [1], on peut imaginer un processus d'abstraction par lequel les caractéristiques communes de séquences suffisamment semblables serait abstraites d'une manière assez similaire à celle par laquelle le contenu de représentations mentales uniques suffisamment semblables est abstrait pour former des concepts plus généraux. Or, mathématiquement parlant, un ensemble de séquences ordonnées n'est rien d'autre que le graphe d'une relation, et donc une telle capacité à représenter les caractéristiques communes des éléments d'un ensemble de séquences pourrait bien se ramener à la capacité à représenter des relations. Ceci n'est bien sûr qu'une hypothèse qui demanderait à être vérifié plus avant, cependant le fait que la capacité à représenter les relations ne semble exister que chez les espèces ayant des capacités cognitives assez développées, ce qui est aussi le cas de la capacité à posséder des souvenirs épisodiques riches et détaillés, en renforce la plausibilité.

Pour finir, il pourrait être intéressant de se pencher sur les conséquences de la conscience de soi chez les agents naturels. Une idée assez largement répandue est que la conscience de soi consiste en la capacité pour un agent cognitif de collecter de l'information sur ses propres opérations neuronales, et de traiter cette information exactement comme l'information sensorielle en provenance du monde extérieur. Si cette idée est juste, alors un agent conscient de lui-même aurait la possibilité de former des représentations mentales de (certaines de) ses propres opérations neuronales, et d'en abstraire ensuite les caractéristiques communes pour former des représentations de certaines classes d'opérations. Parmi les classes d'opérations qui pourraient être ainsi abstraites et représentées, se trouvent en premier lieu la conjonction, la négation et la $\|$-disjonction de contenus d'information. Ce type d'opérations est en effet constamment réa-

1 [Eichenbaum, 2004]. Voir aussi p. 136.

lisé par les agents cognitifs[2], et on peut donc imaginer que les classes d'opérations correspondantes devraient être les premières à être abstraites. Remarquons que puisque d'un point de vue mathématique les connecteurs logiques sont des fonctions et que les fonctions sont des relations, la capacité à représenter des relations ferait ici figure de pré-requis. D'autre part, une telle représentation explicite des connecteurs logiques — qui est à opposer à la réalisation implicite des connecteurs qui prévaut dans les autres cas[3] — pourrait à son tour permettre l'émergence de la représentation verbale syntaxiquement structurée, laquelle caractérise les capacités supérieures de langage et de raisonnement.

Ainsi, partant de la faculté élémentaire de former des représentations mentales et de les utiliser comme support d'inférences et de prise de décisions, on pourrait par l'adjonction progressive de facultés additionnelles reconstruire de larges pans du raisonnement humain. Dans un tel modèle, raisonnement et apprentissage seraient naturellement articulés, et l'itération de l'apprentissage ne poserait pas de problème.

2 Les opérations neuronales supportant vraisemblablement les conjonctions on été brièvement décrites p. 131 ci-dessus. Pour ce faire une idée de ce à quoi pourraient ressembler les opérations neuronales qui correspondent à la négation et à la ‖-disjonction, voir p. 140 (négation) et 166 (‖-disjonction).

3 Voir p. 146 ci-dessus.

Annexes

Annexe A

Pour tout w $\in \mathcal{U}$ et toute $\boldsymbol{L}\Vdash$-formule α, δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ α ssi w $\Vdash\!\models \alpha$.

Démonstration :

$\Longrightarrow$) Soit w $\in \mathcal{U}$ et soit α une $\boldsymbol{L}\Vdash$-formule telle que δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ α.
w $\Vdash\!\models \delta$(w), donc par définition de $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$, w $\Vdash\!\models \alpha$.

$\Longleftarrow$) Soit w $\in \mathcal{U}$ et soit α une $\boldsymbol{L}\Vdash$-formule telle que w $\Vdash\!\models \alpha$. On procède par induction sur la construction des $\boldsymbol{L}\Vdash$-formules :

i) (Cas de base) Si α est une $\boldsymbol{L}$-formule, alors δ(w)$\vdash \alpha$ (voir section 2.1). Donc par supra-classicalité de $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$, δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ α.

ii) (Pas de récurrence) Si α n'est pas une $\boldsymbol{L}$-formule, alors, ou bien $\alpha = \beta \parallel \gamma$ ou bien $\alpha = \beta \curlywedge \gamma$. Supposons qu'on a déjà prouvé que pour tout w'$\in \mathcal{U}$, w' $\Vdash\!\models \beta$ implique δ(w') $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ β et w' $\Vdash\!\models \gamma$ implique δ(w') $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ γ (hypothèse de récurrence).

1. Si $\alpha = \beta \parallel \gamma$, alors par hypothèse
w $\Vdash\!\models \beta \parallel \gamma$, ssi (par déf. de $\parallel$)
w $\Vdash\!\models \beta$ ou w $\Vdash\!\models \gamma$, donc par hypothèse de récurrence
δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ β or δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ γ.

- Si δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ β, alors par la déf. de $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$,
  δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ $\beta \parallel \gamma$, *i.e.* δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ α.
- Sinon, δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ γ, donc par la déf. de $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$,
  δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ $\beta \parallel \gamma$, *i.e.* δ(w) $\|\!\sim_{\mathcal{U},\boldsymbol{L}\Vdash}$ α.

2. Si $\alpha = \beta \curlywedge \gamma$, alors par hypothèse
w $\Vdash\!\models \beta \curlywedge \gamma$, ssi (par déf. de $\curlywedge$)

w $\Vdash \beta$ et w $\Vdash \gamma$, donc par hypothèse de récurrence
δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\shortparallel}}} \beta$ et δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\shortparallel}}} \gamma$, ssi
δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\shortparallel}}} \beta \curlywedge \gamma$, *i.e.* δ(w) $\Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\shortparallel}}} \alpha$.

□

Annexe B

***Injectivity* :**

Pour toute $\boldsymbol{L}^{\shortparallel}$-formule $\boldsymbol{\alpha}$ maximale-consistante dans $\boldsymbol{\mathcal{U}}$,
si $\boldsymbol{\alpha} \parallel \boldsymbol{\beta} \parallel \boldsymbol{\gamma} \Vdash_{\bowtie} \boldsymbol{\beta} \parallel \boldsymbol{\gamma}$ et $\boldsymbol{\alpha} \parallel \boldsymbol{\beta} \not\Vdash_{\bowtie} \boldsymbol{\beta}$, alors $\boldsymbol{\alpha} \parallel \boldsymbol{\gamma} \Vdash_{\bowtie} \boldsymbol{\gamma}$.

Démonstration :

Soit α, β et γ des $\boldsymbol{L}^{\shortparallel}$-formules telles que α est maximale-consistante dans $\mathcal{U}$, $\alpha \parallel \beta \parallel \gamma \Vdash_{\bowtie} \beta \parallel \gamma$ et $\alpha \parallel \beta \not\Vdash_{\bowtie} \beta$, et soit w $\in \mathcal{U}$ $<$-minimal pour $\alpha \parallel \gamma$ (s'il n'existe pas de tel w, alors, trivialement, $\alpha \parallel \gamma \Vdash_{\bowtie} \gamma$). Supposons que w $\not\Vdash \gamma$:

i) Alors w $\Vdash \alpha$.

ii) Par hypothèse α est maximale-consistante, donc α est une $\boldsymbol{L}$-formule. Puisque w $\Vdash \alpha$, δ(w) $\vdash \alpha$.

iii) w $\Vdash \alpha \wedge \delta$(w), donc $\alpha \wedge \delta$(w) $\not\Vdash_{\overline{\mathcal{U},\boldsymbol{L}^{\shortparallel}}} \bot$. Par définition des formules maximales-consistantes, $\alpha \vdash \delta$(w). Donc par ii), $\alpha \equiv \delta$(w).

iv) Par hypothèse, $\alpha \parallel \beta \not\Vdash_{\bowtie} \beta$, ssi
$\exists$ w' $\in \mathcal{U}$ t.q. w' est $<$-minimal pour $\alpha \parallel \beta$ et w' $\not\Vdash \beta$. Donc w' $\Vdash \alpha$, ssi δ(w') $\vdash \alpha$. Puisque α est maximale-consistante, $\alpha \vdash \delta$(w'), donc $\alpha \equiv \delta$(w').

v) Par iii) et iv), w = w', donc par iv), w est $<$-minimal pour $\alpha \parallel \beta$ et w $\not\Vdash \beta$.

vi) Puisque w $\Vdash \alpha$, w $\Vdash \alpha \parallel \beta \parallel \gamma$. On montre que w est $<$-minimal pour $\alpha \parallel \beta \parallel \gamma$:
Supposons le contraire, alors il existe w" $<$ w tel que w" $\Vdash \alpha \parallel \beta \parallel \gamma$. Par $\boldsymbol{L}^{\shortparallel}$-smoothness, w $\not<$ w, donc w" $\neq$ w, donc puisque α est maximale-consistante, w" $\not\Vdash \alpha$. Donc w" $\Vdash \beta \parallel \gamma$. Si w" $\Vdash \beta$, alors w" $\Vdash \alpha \parallel \beta$, donc w n'est pas $<$-minimal pour $\alpha \parallel \beta$, ce qui contredit v). Si w" $\Vdash \gamma$,

alors w" $\Vdash \alpha \parallel \gamma$, donc w n'est pas $<$-minimal pour $\alpha \parallel \gamma$, ce qui contredit l'hypothèse. Donc w est $<$-minimal pour $\alpha \parallel \beta \parallel \gamma$.

vii) Par hypothèse, $\alpha \parallel \beta \parallel \gamma \mid\!\sim_{\mathcal{M}} \beta \parallel \gamma$, donc par vi), w $\Vdash \beta \parallel \gamma$.
Par v) w $\nVdash \beta$, donc w $\Vdash \gamma$, ce qui contredit l'hypothèse.

□

Annexe C

$\parallel$-*Disjunct Equivalence* :

Si $\alpha_1 \mid\!\sim_{\mathcal{M}} \alpha_2$, $\alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_1$ et $\alpha_1 \parallel \beta \mid\!\sim_{\mathcal{M}} \gamma$, alors $\alpha_2 \parallel \beta \mid\!\sim_{\mathcal{M}} \gamma$.

Démonstration :

Soit α_1, α_2, β et γ des $\boldsymbol{L}^{\parallel}$-formules telles que $\alpha_1 \mid\!\sim_{\mathcal{M}} \alpha_2$, $\alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_1$ et $\alpha_1 \parallel \beta \mid\!\sim_{\mathcal{M}} \gamma$.

i) Par hypothèse, $\alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_1$ donc par $\mathcal{U}$-*Right Weakening*, $\alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.

ii) Par *Reflexivity*, $\alpha_1 \mid\!\sim_{\mathcal{M}} \alpha_1$, donc par $\mathcal{U}$-*Right Weakening*, $\alpha_1 \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.
Par $\parallel$-*Or* sur i), $\alpha_1 \parallel \alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.

iii) Par *Reflexivity*, $\alpha_1 \parallel \beta \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$, donc par $\parallel$-*Or* sur ii),
$(\alpha_1 \parallel \alpha_2)\|(\alpha_1 \parallel \beta) \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.
Puisque $(\alpha_1 \parallel \alpha_2)\|(\alpha_1 \parallel \beta) \cong_{\mathcal{U},\boldsymbol{L}^{\parallel}} \alpha_1 \parallel \alpha_2 \parallel \beta$, par $\mathcal{U}$-*Left Equivalence*
$\alpha_1 \parallel \alpha_2 \parallel \beta \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$.

iv) Par *Reflexivity*, $\alpha_1 \parallel \beta \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \beta$, donc par $\mathcal{U}$-*Right Weakening*,
$\alpha_1 \parallel \beta \mid\!\sim_{\mathcal{M}} \alpha_1 \parallel \alpha_2 \parallel \beta$.

v) Par hypothèse $\alpha_1 \parallel \beta \mid\!\sim_{\mathcal{M}} \gamma$, donc par la règle dérivée *Equivalence* sur iii) et iv), $\alpha_1 \parallel \alpha_2 \parallel \beta \mid\!\sim_{\mathcal{M}} \gamma$.

vi) Par hypothèse, $\alpha_1 \mid\!\sim_{\mathcal{M}} \alpha_2$ et par *Reflexivity* $\alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_2$, donc par $\parallel$-*Or* $\alpha_1 \parallel \alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_2$. Donc par $\mathcal{U}$-*Right Weakening*, $\alpha_1 \parallel \alpha_2 \mid\!\sim_{\mathcal{M}} \alpha_2 \parallel \beta$.

vii) Par *Reflexivity* $\beta \mid\!\sim_{\mathcal{M}} \beta$, donc par $\mathcal{U}$-*Right Weakening*, $\beta \mid\!\sim_{\mathcal{M}} \alpha_2 \parallel \beta$.
Donc par $\parallel$-*Or* sur vi), $\alpha_1 \parallel \alpha_2 \parallel \beta \mid\!\sim_{\mathcal{M}} \alpha_2 \parallel \beta$.

viii) Par *Cautious Monotony* sur v) and vii), $(\alpha_1 \parallel \alpha_2 \parallel \beta) \wedge (\alpha_2 \parallel \beta) \mid\!\sim_{\mathcal{M}} \gamma$.
Puisque $(\alpha_1 \parallel \alpha_2 \parallel \beta) \wedge (\alpha_2 \parallel \beta) \cong_{\mathcal{U},\boldsymbol{L}^{\parallel}} \alpha_2 \parallel \beta$, par $\mathcal{U}$-*Left Equivalence*,
$\alpha_2 \parallel \beta \mid\!\sim_{\mathcal{M}} \gamma$.

□

Annexe D

$\|$-*Transitivity* :

Si $\boldsymbol{\alpha_1} \mid\!\sim_{\mathcal{M}} \boldsymbol{\alpha_2}$, ... , $\boldsymbol{\alpha_{n-1}} \mid\!\sim_{\mathcal{M}} \boldsymbol{\alpha_n}$, alors $\boldsymbol{\alpha_1} \| \boldsymbol{\alpha_n} \mid\!\sim_{\mathcal{M}} \boldsymbol{\alpha_n}$.

Démonstration :

Soient α_1, α_2, ... , α_n des $\boldsymbol{L}^{\|}$-formules telles que $\alpha_1 \mid\!\sim_{\mathcal{M}} \alpha_2$, ... , $\alpha_{n-1} \mid\!\sim_{\mathcal{M}} \alpha_n$.

i) Par hypothèse, pour tout $i(1 \leq_{\mathbb{N}} i <_{\mathbb{N}} n)$, $\alpha_i \mid\!\sim_{\mathcal{M}} \alpha_{i+1}$, et, par *Reflexivity* $\alpha_{i+1} \mid\!\sim_{\mathcal{M}} \alpha_{i+1}$.
Donc par $\|$-*Or*, pour tout $i(1 \leq_{\mathbb{N}} i <_{\mathbb{N}} n)$, $\alpha_i \| \alpha_{i+1} \mid\!\sim_{\mathcal{M}} \alpha_{i+1}$.

ii) Par *Reflexivity*, pour tout $i(1 \leq_{\mathbb{N}} i <_{\mathbb{N}} n)$, $\alpha_{i+1} \mid\!\sim_{\mathcal{M}} \alpha_{i+1}$, donc par $\mathcal{U}$-*Right Weakening*, pour tout $i(1 \leq_{\mathbb{N}} i <_{\mathbb{N}} n)$, $\alpha_{i+1} \mid\!\sim_{\mathcal{M}} \alpha_i \| \alpha_{i+1}$.

iii) Par i), $\alpha_{n-1} \| \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_n$, donc par des applications successives de la règle dérivée $\|$-*Disjunct Equivalence* sur i) et ii), on obtient
$\alpha_1 \| \alpha_2 \| \ldots \| \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_n$.

iv) Par $\mathcal{U}$-*Right Weakening* sur iii), $\alpha_1 \| \alpha_2 \| \ldots \| \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_1 \| \alpha_n$.

v) Par *Cautious Monotony* sur iii) et iv),
$(\alpha_1 \| \alpha_2 \| \ldots \| \alpha_n) \wedge (\alpha_1 \| \alpha_n) \mid\!\sim_{\mathcal{M}} \alpha_n$.
$(\alpha_1 \| \alpha_2 \| \ldots \| \alpha_n) \wedge (\alpha_1 \| \alpha_n) \cong_{\mathcal{U}, \boldsymbol{L}^{\|}} \alpha_1 \| \alpha_n$, donc par $\mathcal{U}$-*Left Equivalence*, $\alpha_1 \| \alpha_n \mid\!\sim_{\mathcal{M}} \alpha_n$.

□

Annexe E

***Rankedness* :**

Pour toutes $\boldsymbol{L}^{\shortparallel}$-formules $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$ and $\boldsymbol{\alpha_3}$ maximales-consistantes dans $\boldsymbol{\mathcal{U}}$ (toutes $\not\equiv$), si $\boldsymbol{\alpha_1 \parallel \alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_2}$ et $\boldsymbol{\alpha_1 \parallel \alpha_3 \mid\!\!\not\sim_{\mathcal{M}} \alpha_3}$, alors $\boldsymbol{\alpha_2 \parallel \alpha_3 \mid\!\!\sim_{\mathcal{M}} \alpha_2}$.

Démonstration :

Soit $\mathcal{M} = (\mathcal{U}, <)$ un modèle à mondes partiels fini rangé, et soit α_1, α_2 et α_3 des $\boldsymbol{L}^{\shortparallel}$-formules maximales-consistantes dans $\mathcal{U}$ (toutes $\not\equiv$) et telles que $\alpha_1 \parallel \alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_2$ et $\alpha_1 \parallel \alpha_3 \mid\!\!\not\sim_{\mathcal{M}} \alpha_3$. De plus, soit w $\in \mathcal{U}$ t.q. w est $<$-minimal pour $\alpha_2 \parallel \alpha_3$ (s'il n'existe pas de tel w dans $\mathcal{U}$, alors, trivialement, $\alpha_2 \parallel \alpha_3 \mid\!\!\sim_{\mathcal{M}} \alpha_2$). Supposons que w $\shortparallel\not\models \alpha_2$:

i) Alors w $\shortparallel\models \alpha_3$.
Donc w $\shortparallel\models \alpha_1 \parallel \alpha_3$.

ii) Par hypothèse, $\alpha_1 \parallel \alpha_3 \mid\!\!\not\sim_{\mathcal{M}} \alpha_3$, ssi
il existe w' appartenant à $\mathcal{U}$ tel que w' est $<$-minimal pour $\alpha_1 \parallel \alpha_3$ et w' $\shortparallel\not\models \alpha_3$.
Puisque $<$ est modulaire, by i), w' $\leq$ w.

iii) Puisque w' $\shortparallel\models \alpha_1 \parallel \alpha_3$ et w' $\shortparallel\not\models \alpha_3$, w' $\shortparallel\models \alpha_1$. Donc w' $\shortparallel\models \alpha_1 \parallel \alpha_2$.

iv) Puisque w' $\shortparallel\models \alpha_1$ et α_1 est maximale-consistante, $\alpha_1 \equiv \delta$(w').

v) On montre que w' $\shortparallel\not\models \alpha_2$: supposons le contraire, alors puisque α_2 est maximale-consistante, on a $\alpha_2 \equiv \delta$(w'), donc par iv) $\alpha_1 \equiv \alpha_2$, ce qui contredit l'hypothèse.

vi) Par hypothèse $\alpha_1 \parallel \alpha_2 \mid\!\!\sim_{\mathcal{M}} \alpha_2$, donc par iii) et v), w' n'est pas $<$-minimal pour $\alpha_1 \parallel \alpha_2$. Puisque $\mathcal{M}$ is fini, $<$ est $\boldsymbol{L}^{\shortparallel}$-smooth, donc il existe w" $<$ w' tel que w" est $<$-minimal pour $\alpha_1 \parallel \alpha_2$. w" $\shortparallel\models \alpha_2$, donc w" $\shortparallel\models \alpha_2 \parallel \alpha_3$.

vii) Par vi) w" $<$ w', et par ii), w' $\leq$ w, donc par modularité de $<$, w" $<$ w.

viii) Par vi), w" $\shortparallel\models \alpha_2 \parallel \alpha_3$, donc par vii) w n'est pas $<$-minimal for $\alpha_2 \parallel \alpha_3$, ce qui contredit l'hypothèse.

□

Index des symboles

Les principaux symboles utilisés dans ce travail sont listés ci-dessous. Les références sont celles des pages où un symbole est introduit ou ré-introduit, ou encore où la notion correspondante est expliquée.

Bibliographie

Alchourrón, C., Gardenfors, P. et Makinson, D. (1985). On the logic of theory change : partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530.

Aminoff, E. M., Kveraga, K. et Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, 17(8):379–390.

Barnett, P. D., Nordström, K. et O'Carroll, D. C. (2007). Retinotopic organization of small-field-target-detecting neurons in the insect visual system. *Current Biology*, 17(7):569–578.

Bartos, M. (2008). Hunting prey with different escape potentials — alternative predatory tactics in a dune dwelling salticid. *The Journal of Arachnology*, 35:499–508.

Bayley, P. J., Hopkins, R. O. et Squire, L. R. (2006). The fate of old memories after medial temporal lobe damage. *The Journal of Neuroscience*, 26(51):13311–13317.

Bayley, P. J. et Squire, L. R. (2005). Failure to acquire new semantic knowledge in patients with large medial temporal lobe lesions. *Hippocampus*, 15(2):273–280.

Bear, M. F. et Abraham, W. C. (1996). Long-term depression in hippocampus. *Annual Review of Neuroscience*, 19:437–462.

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology*, 61:27–48.

Cahill, L. et McGaugh, J. L. (1995). A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and Cognition*, 4(4):410–421.

Chittka, L. et Niven, J. (2009). Are bigger brains better ? *Current Biology*, 19:995–1008.

Dubois, D. (2008). On ignorance and contradiction considered as truth-values. *Logic Journal of the IGPL*, 16(2).

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience*, 1(1):41–50.

Eichenbaum, H. (2004). Hippocampus : Cognitive processes and neural representations that underlie declarative memory. *Neuron*, 44(1):109–120.

Freund, M. (2004). On the revision of preferences and rational inference processes. *Artificial Intelligence*, 152(1):105–137.

Gabbay, D. M. et Schlechta, K. (2008). Cumulativity without closure of the domain under finite unions. *The Review of Symbolic Logic*, 1(3):372–392.

Gabbay, D. M. et Schlechta, K. (2009). Roadmap for preferential logics. *Journal of Applied Non-classical Logics*, 19(1):43–95.

Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R. et Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, 322(5898):96–101.

Giurfa, M. (2007). Behavioral and neural analysis of associative learning in the honeybee : a taste from the magic well. *Journal of Comparative Physiology A*, 193(8):801–824.

Holdstock, J. S., Hocking, J., Notley, P., Devlin, J. T. et Price, C. J. (2009). Integrating visual and tactile information in the perirhinal cortex. *Cerebral Cortex*, 19(12):2993–3000.

Hollis, K. L. et Guillette, L. M. (2011). Associative learning in insects : Evolutionary models, mushroom bodies, and a neuroscientific conundrum. *Comparative Cognition & Behavior Reviews*, 6:24–45.

Hubel, D. H. et Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148(3):574–591.

Hubel, D. H. et Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154.

Hubel, D. H. et Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195(1):215–243.

Izquierdo, I. (1993). Long-term potentiation and the mechanisms of memory. *Drug Development Research*, 30(1):1–17.

Jackson, R. R., Carter, C. M. et Tarsitano, M. S. (2001). Trial-and-error solving of a confinement problem by a jumping spider, *Portia fimbriata*. *Behaviour*, 138(10):1215–1234.

Jackson, R. R. et Cross, F. R. (2011). Spider cognition. *Advances in Insect Physiology*, 41:115–174.

Konieczny, S. et Pino Perez, R. (2002). Sur la représentation des états épistémiques et la révision itérée. *In Révision des croyances*, pages 181–202. Hermes Science Publications.

Kraus, S., Lehmann, D. et Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44. (Page numbers are those of the arXiv version).

Kripke, S. (1965). Semantical analysis of intuitionistic logic I. *In* Crossley, J. et Dummett, M., éditeurs : *Formal systems and recursive functions*, pages 92–130. North-Holland Publishing Company.

Lehmann, D. et Magidor, M. (1992). What does a conditional knowledge base entail ? *Artificial Intelligence*, 55.

Leitgeb, H. (2004). *Inference on the low-level, an investigation into deduction, nonmomotonic reasoning and the philosophy of cognition*. Kluwer Academic Publishers.

Leopold, D. A. et Logothesis, N. K. (1999). Multistable phenomena : changing views in perception. *Trends in Cognitive Sciences*, 3(7):254–264.

Levy, D. A., Bayley, P. J. et Squire, L. R. (2004). The anatomy of semantic knowledge : Medial vs. lateral temporal lobe. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6710–6715.

Lin, L. *et al.* (2007). Neural encoding of the concept of nest in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14):6066–6071.

Manns, J. R., Hopkins, R. O. et Squire, L. R. (2003). Semantic memory and the human hippocampus. *Neuron*, 38(1):127–133.

Martin, A. et Chao, L. L. (2001). Semantic memory and the brain : structure and processes. *Current Opinion in Neurobiology*, 11(2):194–201.

McGaugh, J. (2000). Memory — a century of consolidation. *Science*, 287(2): 248–251.

Menzel, R. et Giurfa, M. (2001). Cognitive architecture of a mini-brain : the honeybee. *Trends in Cognitive Sciences*, 5(2):62–71.

O'Carroll, D. (1993). Feature-detection neurons in dragonflies. *Nature*, 362(6420):541–543.

Paulk, A. C., Dacks, A. M. et Gronenberg, W. (2009). Color processing in the medulla of the bumblebee (Apidae : *Bombus impatiens*). *The Journal of Comparative Neurology*, 513(5):441–456.

Paulk, A. C., Phillips-Portillo, J., Dacks, A. M., Fellous, J. M. et Gronenberg, W. (2008). The processing of color, motion, and stimulus timing are anatomically segregated in the bumblebee brain. *The Journal of Neuroscience*, 28(25):6319–6332.

Phelps, E. A. et LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing : From animal models to human behavior. *Neuron*, 48(2):175–187.

Preston, A. R. et Wagner, A. D. (2007). The medial temporal lobe and memory. *In* Kesner, R. et Martinez, J., éditeurs : *Neurobiology of Learning and Memory*, pages 305–337. Elsevier. 2nd Edition.

Quiroga, R. Q. (2012). Concept cells : the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13:587–597.

Sanes, J. R. et Zipursky, S. L. (2010). Design principles of insect and vertebrate visual systems. *Neuron*, 66(1):15–36.

Schildberger, K. (1984). Multimodal interneurons in the cricket brain : properties of identified extrinsic mushroom body cells. *Journal of Comparative Physiology A*, 154(1):71–79.

Schwartz, J., Grimault, N., Hupé, J., Moore, B. C. J. et Pressnitzer, D. (2012). Multistability in perception : binding sensory modalities, an overview. *Philosophical Transactions of the Royal Society B*, 367(1591):896–905.

Shimamura, A. P. (2002). Relational binding theory and the role of consolidation in memory retrieval. *In* Squire, L. et Schacter, D., éditeurs : *Neuropsychology of memory*, pages 61–72. The Guilford press. 3rd Edition.

Shimamura, A. P. (2010). Hierarchical relational binding in the medial temporal lobe : The strong get stronger. *Hippocampus*, 20(11):1206–1216.

Squire, L. R. et Alvarez, P. (1995). Retrograde amnesia and memory consolidation : a neurobiological perspective. *Current Opinion in Neurobiology*, 5(2):169–177.

Squire, L. R. et Bayley, P. J. (2007). The neuroscience of remote memory. *Current Opinion in Neurobiology*, 17(2):185–196.

Squire, L. R., Stark, C. E. L. et Clark, R. E. (2004). The medial temporal lobe. *Annual Review of Neurosciences*, 27:279–306.

Srinivasan, M. V. (2006). Honeybee vision : In good shape for shape recognition. *Current Biology*, 16(2):58–60.

Sterzer, P., Kleinschmidt, A. et Rees, G. (2009). The neural bases of multistable perception. *Trends in Cognitive Sciences*, 13(7):310–318.

Strausfeld, N. J., Hansen, L., Li, Y., Gomez, R. S. et Ito, K. (1998). Evolution, discovery, and interpretations of arthropod mushroom bodies. *Learning & Memory*, 5(1):11–37.

Suzuki, W. A. et Eichenbaum, H. (2000). The neurophysiology of memory. *Annals of the New York Academy of Sciences*, 911:175–191.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex : clustering of cells with similar but slightly different stimulus selectivities. *Cerebral Cortex*, 13(1):90–99.

Taylor, K. I., Moss, H. E., Stamatakis, E. A. et Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 103(21): 8239–8244.

Tsunoda, K. *et al.* (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4(8):832–838.

Wessnitzer, J. et Webb, B. (2006). Multimodal sensory integration in insects — towards insect brain control architectures. *Bioinspiration & Biomimetics*, 1(3).

Wilson, R. I. et Mainen, Z. (2006). Early events in olfactory processing. *Annual Review of Neuroscience*, 29:163–201.

Zhang, Y., Lu, H. et Bargmann, C. I. (2005). Pathogenic bacteria induce aversive olfactory learning in *Caenorhabditis elegans*. *Nature*, 438(10):179–184.