Dialogue and Decision Games for Information Exchanging Agents

SIKS Dissertation Series No. 2006-13

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Printed by Ponsen & Looijen B.V., Uitgeverij, Wageningen
Cover illustration: Mammatus clouds Tulsa 1973
ISBN-10: 90-6464-020-3
ISBN-13: 978-90-6464-020-9
NUR: 984 Kunstmatige intelligentie

Dialogue and Decision Games for Information Exchanging Agents

Dialoog- en Beslisspellen voor Informatie Uitwisselende Agenten

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. W.H. Gispen,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op 18 september 2006 des middags te 12.45 uur

door

Henk-Jan Lebbink

geboren op 6 maart 1976, te Hoogeveen

Promotoren: Prof.dr. J.-J.Ch. Meyer
Prof.dr. C.L.M. Witteman

Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van Emotional Brain B.V.

To my parents

Contents

1 Introduction **1**

2 Epistemological Foundations **9**
- 2.1 Propositions and Beliefs . . . 9
- 2.2 Theories of Truth . . . 10
  - 2.2.1 Correspondence theory of truth . . . 11
  - 2.2.2 Coherence theory of truth . . . 12
  - 2.2.3 Consensus theory of truth . . . 14
  - 2.2.4 Wittgenstein's account of truth . . . 14
- 2.3 Agents and a Use Account of Truth . . . 16
  - 2.3.1 Agents and a use account of belief . . . 16
  - 2.3.2 Agents and a use account of knowledge . . . 18
- 2.4 Agents and Meaning . . . 20
  - 2.4.1 Agents understanding meaning . . . 20
  - 2.4.2 Changing the meaning . . . 21
  - 2.4.3 Dispute the meaning . . . 22
- 2.5 Concluding Remarks . . . 23

3 Paraconsistent Logics **25**
- 3.1 An Unavoidable Inconsistent World Picture . . . 25
- 3.2 Sources of Inconsistencies . . . 27
- 3.3 A Use Interpretation of Believing Contradictions . . . 28
- 3.4 Paraconsistent Logics . . . 31
- 3.5 Concluding Remarks . . . 32

4 Multi-valued Logics **35**
- 4.1 Bilattice Structure . . . 35
  - 4.1.1 Rejecting the principle of the excluded middle . . . 36
  - 4.1.2 Lattices . . . 36

4.1.3 Bilattices 36
4.1.4 Truth values 37
4.1.5 Truth values orderings, and maximal and minimal elements . . 38
4.1.6 Truth value successor 39
4.2 A Language of Multi-valued Logic 40
4.3 Theories of Multi-valued Logic 40
4.3.1 Normal and complete theories 41
4.3.2 Inverse theories 42
4.4 Multi-valued Logic Theory Descriptions 44
4.5 Changing Theories of Multi-valued Logic 45
4.5.1 Adding propositions to theories 45
4.5.2 Retracting propositions from theories 47
4.6 Concluding Remarks 50

5 Cognitive Agents **51**
5.1 The Agent's Cognitive State 52
5.1.1 The mental state structure 52
5.1.2 The language of mental state 54
5.1.3 Descriptions of the mental states 56
5.1.4 The agent's knowledge base 59
5.1.5 The agent's ontology and epistemology 60
5.1.6 Definition of the cognitive state 64
5.2 Coherence of an Agent's Mental State Structure 64
5.2.1 Coherence of beliefs about beliefs 64
5.2.2 Coherence of beliefs about desires 68
5.2.3 Coherence of desires 69
5.3 Agent System Dynamics 72
5.3.1 The basic actions of change 72
5.3.2 The complex actions of change 73
5.3.3 Updating the agent's mental state structure 76
5.3.4 The agent's deliberation cycle 76
5.4 Concluding Remarks 79

6 Decision Games **81**
6.1 Making Decisions 82
6.1.1 Symbolic AI 82
6.1.2 Meaning and use of decisions 83
6.1.3 Decision games 84
6.1.4 The activity of decision-making 85
6.1.5 Cognitive entrenchment 88
6.2 Generic Criteria to Change Beliefs 88
6.2.1 Generic criteria for decisions to believe 88
6.2.2 Generic criteria for decisions to be ignorant 89
6.3 Inference Rules and Meaning to Change Beliefs 90

6.3.1 Deciding to add beliefs using inference rules 90
6.3.2 Deciding to retract beliefs using inference rules 91
6.3.3 Deciding to add and retract beliefs using meaning 91
6.3.4 Deciding that other agents add and retract beliefs 92
6.4 Conformism to Change Beliefs 94
6.4.1 Conformism to believe 94
6.4.2 Conformism to be ignorant 95
6.5 Desiring to Change Beliefs 96
6.5.1 The motivations to make decisions 96
6.5.2 Cognitive preconditions 97
6.5.3 Coherent desires 99
6.5.4 Indirect coherent desires 101
6.5.5 Desiring to believe 104
6.5.6 Desiring to be ignorant 105
6.6 Concluding Remarks 105

7 Dialogue Games 107
7.1 A Background to Dialogue Games 108
7.1.1 Communications in speech act theory 108
7.1.2 Agent communication 110
7.1.3 Dialogue games 112
7.1.4 The activity of communicating 114
7.2 Questions for a Belief Addition 115
7.2.1 Intuitive readings 115
7.2.2 An example dialogue 117
7.2.3 Posing questions for belief addition 118
7.2.4 Interpreting communicative acts 120
7.2.5 Positive answers to questions for belief addition 121
7.2.6 Negative answers to questions for belief addition 122
7.2.7 Distributed chains of justifications 123
7.3 Questions for a Belief Retraction 126
7.3.1 Intuitive readings 126
7.3.2 Reasons for asking approval 127
7.3.3 An example dialogue 127
7.3.4 Posing questions for belief retraction 129
7.3.5 Positive answers to questions for belief retraction 130
7.3.6 Negative answers to questions for belief retraction 131
7.3.7 Distributed belief revision 131
7.4 Requesting a Belief Addition 133
7.4.1 Intuitive readings 133
7.4.2 An example dialogue 134
7.4.3 Requesting a belief addition 135
7.4.4 Granting a request for belief addition 136
7.4.5 Declining a request for belief addition 137

7.5 Requesting a Belief Retraction . . . 137
7.5.1 Intuitive readings . . . 137
7.5.2 An example dialogue . . . 139
7.5.3 Requesting a belief retraction . . . 140
7.5.4 Granting a request for belief retraction . . . 141
7.5.5 Declining a request for belief retraction . . . 141
7.6 Inform Statement about Changed Belief and Desire . . . 142
7.6.1 Intuitive reading . . . 142
7.6.2 An example dialogue . . . 143
7.6.3 Inform statement . . . 144
7.6.4 Correct representations . . . 144
7.7 Run Out of Communicative Acts . . . 146
7.8 Concluding Remarks . . . 147

8 Agreeing to Disagree 149
8.1 Disagreements . . . 150
8.1.1 Contradictory propositions . . . 150
8.1.2 Disagreeing and agreeing agents . . . 150
8.1.3 Resolving disagreements . . . 151
8.1.4 Disagreement-awareness . . . 154
8.1.5 Resolving disagreement-awareness . . . 157
8.2 Acting to resolve Disagreements . . . 162
8.2.1 Questions to resolve disagreements . . . 162
8.2.2 Requests to resolve disagreements . . . 163
8.3 To Agree to Disagree . . . 164
8.3.1 Irresolvable disagreements . . . 164
8.3.2 Believing irresolvable disagreements . . . 165
8.3.3 Communicating irresolvable disagreements . . . 166
8.3.4 Irresolvable disagreement-awareness . . . 166
8.3.5 Agreeing or disagreeing to disagree . . . 167
8.4 Interpretation of an Agreement to Disagree . . . 170
8.4.1 The 'no disagreement' theorem . . . 170
8.4.2 Alleged common knowledge . . . 171
8.5 Concluding Remarks . . . 173

9 Conclusions and Future Research 175
9.1 Conclusions . . . 176
9.1.1 Our design of a multiagent system . . . 176
9.1.2 Justified beliefs only . . . 176
9.1.3 Software providing decision support . . . 178
9.2 Dialogues about Grounds of Justification . . . 179
9.3 Dialogues about Meaning . . . 181

Bibliography 183

A Proofs **195**
A.1 Useful Propositions . . . 195
A.2 Proofs from Chapter 4 . . . 195
A.3 Proofs from Chapter 5 . . . 197
A.4 Proofs from Chapter 7 . . . 199
A.5 Proofs from Chapter 8 . . . 200

B Samenvatting **205**

C Dankwoord **211**

D SIKS Dissertation Series **215**

Chapter 1

Introduction

The company Emotional Brain is developing a software program that supports physicians to diagnose patients with psycho-pathological syndromes. Additionally, the software can predict how medicinal interventions will change the patient's physiological state. This diagnostic and predictive instrument may support physicians to assess medical needs of patients with vague and indistinct symptoms.

The domain knowledge of the software program comes from experts in psycho-neurological endocrinology and specifically in clinical psychological endocrinology, psycho-pharmacological endocrinology, and bio-physiological endocrinology. These experts are affiliated with the research group psychopharmacology of the faculty of Pharmaceutical Sciences at Utrecht University in the Netherlands. Central to the software program is the knowledge about the neurotransmitters dopamine, serotonin, noradrenalin and acetylcholine that comprise neurological-psychological personality profiles. These neurotransmitter systems describe which people are susceptible to depression, and which medication is likely to give desirable results in treating depression.

An important feature of this knowledge is that certain pieces are shared by experts, while other pieces are not. Another important feature is that this knowledge is scientific. Scientific knowledge is often subject to discussion and change, and such discussion and change becomes manifest as disagreements among experts over essential or less essential aspects of the medical diagnosis and treatment. These two properties may make the knowledge contradicting, because—knowingly or not—experts can have different opinions about aspects of the knowledge.

In this thesis, we report about software agents that are designed to represent knowledge of human professionals. Equipped with knowledge from a medical expert, these agents can assist physicians with medical assessments. Separate software agents, each with knowledge from a different medical expert, can potentially assist physicians even better. Without human mediations, agents can act autonomously and consult each other to provide joint assessments. Emotional Brain has decided to design the software program as a multiagent system in which several agents each represent one medical expert or one domain of relevant knowledge. These software agents may consult each other to reach well-founded and justified assessments, and may request others to adopt and retract beliefs to reach joint assessments.

Our agents can be programmed to report their lack of information to conclude on an assessment as well as their joint assessments to the physicians.[1] Agents can also be programmed to report their disagreements with other agents to their responsible human experts. If agents were to report every encountered disagreement to their experts, this would necessitate profuse human consulting. In general, humans should only be consulted if agents cannot conclude or cannot resolve disagreements themselves. Our agents will have operational definitions for when their disagreements are irresolvable. That is, agents will be provided with explicit criteria when attempts to resolve a disagreement have failed. If two agents have an irresolvable disagreement, they may propose to agree to disagree. Because expert time is expensive, agents should only consult them as a last resort, that is: if agents agree to disagree.

We will use an interpretation by Ellenbogen [Ell03] of a use account of truth that has been described by the later Wittgenstein[2]. This account of truth describes when our agents are justified to have beliefs, beliefs about other agent's beliefs, or when they are justified not to have beliefs. We will not contribute to a theory of truth that entitles humans to call their beliefs justified and true, but we will describe whether our agents are, according to the responsible human experts, justified to have true beliefs. Our agents will be justified to call their beliefs true, if, in the domain of knowledge that the agents are said to represent, they are entitled to call their beliefs true. Explicit criteria will provide agents with operational definitions about when other agents (and they themselves) are justified to have beliefs. We will develop communication strategies that entitle agents to have beliefs that are based on distributed chains of justifications. Likewise, we will develop decision strategies that implement multiagent belief revision. These decision strategies will prescribe that our agents may retract their beliefs only after they have conferred with others. Retracting beliefs will occur in a fashion in which beliefs remain justified in a distributed fashion. We will argue that if our agents have initial beliefs that are justified in the domain of knowledge that the agents represent, then their conclusions will also be justified in this domain.

> Given that we design a multiagent system in which agents each can represent a medical expert or a domain of relevant knowledge, our research aim is to model the agents' decision-making and communication, such that the justifications of the agent's beliefs are preserved.

[1] The multiagent system we will design in this thesis is implemented by the company Emotional Brain in the software language Java; however, we will not discuss the software implementation. Nor will we discuss the medical domain that our agent's may represent.

[2] There are two commonly recognized stages of Wittgenstein's thought—the early and the later. The early Wittgenstein provided new insights into the relations between world, thought and language. The later Wittgenstein took the step in critiquing all of traditional philosophy including its climax in his own early work.

Cognitive Agents and Multiagent Systems

We may dispute whether everyday systems, ranging from thermostats to human beings, are cognitive systems with beliefs and desires. We adopt the intentional stance[3] and agree with the view that human beings are taken as classic examples of cognitive systems. However, a sceptical view might endorse that humans are just very successful in convincing others to believe they are cognitive, while, in fact, they are only pretending. A more plausible view is that if systems can convince us that they hold beliefs, we take it that they are worthy of being classified and treated as having cognitions.[4] An operational definition whether software agents have beliefs should specify a process of inquiry in which the presence or absence of beliefs are measured. This inquiry will lead up to a 'verdict' whether or not the agent allegedly holds a belief. Agents communicate to convince other agents that they have certain cognitive attitudes, such as beliefs. Therefore, an agent's judgement whether another agent believes a proposition is based on certain criteria. Interpreted with a Wittgensteinian use-semantics (from Ellenbogen [Ell03]), according to the agent's community, such criteria are sufficient to correctly predicate that other agents believe propositions.

If convincing arguments are sufficient criteria to believe correctly that another agent believes a proposition, what would then be the sufficient criteria for an agent to believe correctly a proposition. Whichever philosophical stance we adopt for humans (see Latour [Lat99]), for agents we take it that it is their deliberate decision-making that explicitly defines the criteria that allow them to be explicitly justified to believe and know propositions. Thus, agents have to be equipped with rules that explicitly state the criteria when they are allowed to believe propositions. Additionally, they have to be equipped with rules that explicitly state the criteria when they are correct to believe that other agents have beliefs.

Agents capitalise on their ability to interact with other agents to co-operate, co-ordinate and negotiate to pursue their collective goals, and they act in competitive ways with the intent to achieve their private goals (see Wooldridge and Jennings [WJ95]; Russell and Norvig [RN03]; Weiss [Wei99]; Jennings et al. [JSW98]). Agents also have the ability to make autonomous actions and the ability to make autonomous decisions. These decisions are about how to interact with other agents and about which actions to take. To effectively establish their objectives, agents have to de-

[3] Dennett, in *True Believers: The Intentional Strategy and Why It Works* [Den81], asserts that if an agent adopts the intentional stance, she follows a predictive strategy of interpretation that presupposes the rationality of the entities the agent hopes to understand and predict. This strategy is very useful and powerful because it addresses understanding and prediction at a level of abstraction useful for agent communication.

[4] For if convincing is not enough, what would be? Turing has put forward a compelling analogous argument in his seminal article *Computing Machinery and Intelligence* [Tur50] in which he held that computers would in time be programmed to acquire abilities rivalling human intelligence. He put forward the idea of an 'imitation game' in which a computer and a human being would be interrogated and if the interrogator could not distinguish them by questioning them, it would be unreasonable not to call the computer intelligent.

ploy the appropriate processes, such as learning, problem solving and planning (see Wooldridge and Jennings [WJ97]; d'Inverno et al. [dLW97]; Grosz and Kraus [GK96]).

Multiagent systems are systems composed of multiple interacting agents (Jennings [Jen01]). These agents are considered autonomous and computational entities that perceive their environment, model their perceptions and act upon their models in a fashion that is intended to achieve their prime (design) objectives (see Lesser [Les95]; Sycara [Syc98]; Huhns and Singh [HS97a]; Moulin and Chaib-Draa [MCD96]).

Wooldridge [Woo02], among others, distinguishes two perspectives on multiagent systems. The micro perspective centres on an agent's individual abilities, such as autonomy over her actions and decisions without the intervention of humans and other agents. The principal engineering objective from this perspective is to design agent systems that can reason and draw decisions autonomously. The macro perspective centres on issues related to groups of agents for who communication is vital to co-ordinate their actions and decisions to pursue their collective goals. The ultimate objective from the macro perspective is to design agents that effectively communicate and collaborate to pursue otherwise unattainable goals. We adopt the latter perspective to model the communication between software agents that have the goal to reach collaborative assessments.

To model such agent communication, the agent system literature distinguishes three different approaches. In the *mental model* approach, the agent's cognitive state is central to the agent's justification and generation of communicative utterances. These utterances have a direct effect on the receiver's cognitive state (cf. Cohen and Perrault [CP79]). Another approach is based on the *social structure* that is presumed to exist before the agents engage in conversation. The validity of communicative utterances depends on the speaker's perceived social structure in which she is allowed to make them (cf. Colombetti [Col00] and Singh [Sin00]). The third approach is based on *argumentation theory* in which communicative utterances are allowed as in argumentation (cf. Amgoud et al. [APM00, AMP00]).

Our approach to model agent communication cannot be uniquely classified as one of these three approaches. We take it that an agent is entitled to use the predicate 'to believe' a proposition, and to utter communicative utterances based on her beliefs, desires, and her beliefs and desires about other agents' beliefs and desires. In this respect, we may view our agents from the mental model approach. However, our agents will be entitled to use the predicate 'to believe' and to use utterances according to criteria that have been agreed upon by their community. In this respect, the community provides a social structure that agents use to be justified to communicate and to have beliefs. In a sense, after an agent has uttered a communicative act that allows the intended receiver to believe that the speaker believes a proposition, the speaker makes a commitment to the receiver to believe the proposition. In anticipation, the receiver may use the manifested belief of the speaker in collaborative plans, knowing that the speaker has committed and thus will consult her if she wishes to change her beliefs. Agents will use privately held beliefs and desires, and manifested beliefs and desires of other agents, as arguments to be correct, according to her community, to make decisions and to utter communicative acts.

Aim and Scope

Our aim is to model agents' decision-making and communication, such that justifications of the agent's beliefs are preserved. We discuss the following subjects.

- *Use-semantics.* We provide a Wittgensteinian use-semantics for epistemic propositions such that we will be able to express what it means for an agent to believe propositions. This use-semantics will allow our agents to believe propositions conform to the conventions of the human experts that our agents will represent. With his language games, the later Wittgenstein described that the meaning of a statement is determined by a community's agreement on the criterion for its correct usage. We will use games that are similar to Wittgenstein's language games to provide semantics for decision-making and communication. Our agents' use of language will be described by criterion rules. These rules describe the meaning of the words in the language, including when to call propositions a truth. Our agents will be justified to regard propositions truths if the criteria to do so have been met. The criteria for decision-making and communication will have to adhere to the conventions of the human experts that our agents represent, making their application correct in the community of the human expert to make decisions and communicate information.

  In Chapter 2, we briefly review the prevailing truth-theories such as Correspondence theory of truth, Coherence theory of truth, and a theory of truth based on Wittgensteinian use-semantics. We provide the ontological commitment of the logic we use in subsequent chapters to model an agent's cognitive state. The ontological commitment of a language is what it assumes about the nature of reality (cf. Russell and Norvig [RN03, p. 242]). Unlike propositional logic that assumes that there are facts that either hold or do not hold in the world and that facts thus are either true or false, we will assume that propositions have an agreed-upon use that determines their truth value. It is such use that we take to be the truth-conditions for agents to have correct beliefs and knowledge.

- *Decision games.* The situations in which a decision is correctly drawn will be specified as the decision's principal determinant. We will use such determinants as criteria in criterion rules that will define the agent's decision-making, resulting in criterion-governed decision behaviour, that, according to the agent, is conform the shared use of the decision in her community. Thus, our agent's ability to believe correctly justified propositions is provided by criteria that entitle her to make decisions to come to believe such propositions. The shared use of the epistemic predicates *to believe* and *to be ignorant* is provided with decision games.

  In Chapter 6, we provide the general structure of criterion-governed decision-making. Two decision games will be presented that enable our agents to decide to add and retract beliefs. Both decision games will consist of three semantically different decisions, providing the meaning of believing a proposition. In one decision, the agent is justified to decide to add a belief based on her private knowledge; in the second decision, the agent is justified to decide to add a belief based on the meaning of believing a particular proposition; the third decision allows an agent to conform to other agents' beliefs. The consequences of making the semantically distinct decisions are the same: the agent is justified to predicate *to believe* a particular proposition. Similar semantics will be provided for the decision to be ignorant about a proposition. The consequence of making this decision is that the agent is justified to predicate *to be ignorant* about a particular proposition.

- *Dialogue games.* To communicate beliefs and desires, we will define speech acts that allow our agents to exchange information about their mental states. To use speech acts correctly, communication will also be governed by criterion rules: the semantics of speech acts will be provided by dialogues games. Our agents can communicate information with speech acts because they regard the use of speech acts shared in their community. The exchange of information between our agents will be the result of moves in a dialogue game that our agents make and interpret as utterances of speech acts.

  In Chapter 7, we present five dialogue games that provide agents with arguments to make justified decisions. Two games will define the use of speech acts that agents may utter with the aim of providing themselves with information. With this information, agents can become justified to decide to add beliefs or to retract beliefs. Two other dialogue games will define the use of speech acts that agents may utter with the aim of providing other agents with information. With this information, other agents may become justified to decide to add or retract beliefs. An extra dialogue game will inform agents about changed cognitive states.

- *Agreeing to disagree.* If an agent has a disagreement with another agent, and the five dialogue games could not make either agent justified to decide to change her beliefs such that the disagreement would be resolved, then, as seen from the agent's perspective, the disagreement is irresolvable. Because communication is governed by dialogue games, and decision-making to change beliefs is governed by decision games, the agents have an operational definition when they run out of options to resolve their disagreements.

  In Chapter 8, we use the decision and dialogue games to implement agent behaviour that may resolve the disagreement. However, if the disagreement turns out to be irresolvable, the agents settle their dispute by agreeing to disagree.

- *Paraconsistency.* We will argue that an agreement to disagree is tantamount to believing an inconsistent proposition, and that an agreement to disagree between two agents can be a third agent's justification to believe an inconsistent proposition. Because the meaning of epistemic propositions, such as beliefs, is provided by a use-semantics, our agents can give a sensible meaning to inconsistent propositions, just as to any other consistent proposition.

In Chapter 3, we will describe inconsistent and paraconsistent propositions. In short, a paraconsistent proposition is a consistent representation of a proposition with inconsistent information. We will describe several sources of inconsistencies, and that our agents may sometimes not be able to avoid believing paraconsistent propositions. But foremost, we will describe the meaning of paraconsistent propositions and describe how our agents will use these propositions to guide their behaviours.

- *Multi-valued logic.* If we are to enable our agents to believe inconsistent propositions, we need a logic that allows them to have such propositions consistently. To cater for additional epistemic modalities that our agents may need, such as an unknown belief state, a multi-valued logic is defined in which a proposition's truth value is not restricted to denote truth and falsity. We define a formal method that allows propositions with epistemic modality such as ignorance, inconsistency and bias. Because agents are entities with epistemic attitudes, logics with the latter epistemic modality are particularly suitable to describe the cognitive state of agent systems.

  In Chapter 4, we define multi-valued logic to represent inconsistencies as well as other epistemic attitudes agents may have. Examples of such epistemic attitudes are inconsistent and unknown belief states, as well as epistemic attitudes biased towards true, false or inconsistent belief states. We provide the epistemological commitment of the logic we use to model an agent's cognitive state. A logic's epistemological commitment provides the possible states of knowledge that the logic allows for each state of affairs (cf. Russell and Norvig [RN03, p. 244]). For our logic, the epistemological commitment provides the possible epistemic states an agent may associate with each state of affairs, such as believing a proposition to be true, to be false, to be inconsistent, or having no opinion. The multi-valued logics are used to define the agent's cognitive state on which the preconditions i.e. the truth-conditions for correct use of decision and speech acts are defined. The rules that define correct decision-making and correct communication are designed to handle inconsistent information properly.

Outline of this Thesis

After a brief general discussion of truth-theories, in Chapter 2, we present and contrast the Wittgensteinian use-account of truth with Correspondence theory and Coherence theory of truth. We then in Chapter 3 describe paraconsistent logics, and describe the meaning of paraconsistent propositions with the Wittgensteinian use-account of truth. The next chapter deals with the definition of multi-valued logics that will allow our agents to believe paraconsistent propositions. Then in Chapter 5 we define our cognitive agents: they will consist of a mental state structure that represents the agent's beliefs and desires, a set of inference rules that constitute the agent's knowledge base, and a set of inferences that constitute the agent's view on the meaning of epistemic propositions. Chapter 6 is devoted to the definition of the two decision games our agents will have. The following chapter describes the five dialogue games. Chapter 8 will deal with agreeing to disagree, and with the concept of common knowledge mentioned in Chapter 2. We provide conclusions and future research in Chapter 9.

Chapter 2

Epistemological Foundations

Meaning is use.
LUDWIG WITTGENSTEIN [Wit01]

Philosophers have given a great deal of thought to the subject of beliefs, justifications for beliefs, and to the question of whether people are allowed to think that their beliefs are knowledge. This chapter lays the epistemological foundations for our agents, who will be justified to have beliefs, in accordance with a use-based theory of truth. We will present a brief overview of three theories of truth, including a specific interpretation by Ellenbogen [Ell03] of Wittgenstein's account of truth and meaning. To allow our agents to have beliefs and knowledge, we will use a Wittgensteinian interpretation of epistemic statements that denote an agent's beliefs and knowledge.

In Section 2.1, we describe the notions of propositions, beliefs, and belief justifications. We then in Section 2.2 briefly review three well-known theories of truth: Correspondence, Coherence, and Consensus theory of truth. In Section 2.3, we provide the epistemological foundations of belief, knowledge and common knowledge that we will use throughout the remainder of this thesis. In the last section (Section 2.4), we will provided agents with a foundation for an explicit account of meaning.

2.1 Propositions and Beliefs

We are primarily concerned with propositions that can be the object of an agent's epistemic modality, such as her beliefs and desires. Whereas sentences are linguistic entities in some language, propositions are often taken to be non-linguistic, abstract and timeless entities. Because we will use propositions to represent beliefs and desires, we take it that propositions are the bearers of truth values. Additionally, we take it that propositions can be expressed by statements, and that statements can be expressed by sentences.

For ease of reference, we will often use personal names to denote arbitrary agents. Agent Sarah will often be a speaker in a dialogue, and she will be speaking to a listening agent John. Sometimes we need a third agent Fred to make the conversations more interesting. We will abbreviate Sarah with an *'s'*, John with a *'j'* and Fred with an *'f'*.

If reasons exist to believe a proposition, then Sarah is said to be *justified* to believe the proposition, and if Sarah is aware of these reasons, then she is said to be *explicitly* justified to believe the proposition. The reasons for believing a proposition are called *belief justifications*, and Sarah's beliefs that are in fact justified are called *justified beliefs* (see Audi [Aud98, p. 2]). The following case illustrates that Sarah may fail to believe propositions while she is in principle justified to believe them. Fermat's last theorem expresses an a priori proposition which proves to be true after painstaking analysis (see Aczel [Acz96]). The proposition is that equation x^n plus y^n equals z^n and has no non-zero integer solutions for x, y and z when $n > 2$. Sarah may believe this proposition to be true based on, for example, a sequence of other true propositions. This sequence provides her with a reason to believe the proposition; making her belief in the proposition explicitly justified. Even though the truth of the proposition was long disputed and not known, Sarah had always been justified in principle.

The opposite of failing to believe justified beliefs can also occur: Sarah can have unjustified beliefs. Several reasons can be given why Sarah can have unjustified beliefs. She can be explicitly justified to believe a proposition ψ; however, if she retracts the belief justifications, ψ becomes unjustified. These unjustified beliefs can persist as long as their unjustified nature is not revealed. To complicate matters, Sarah can be explicitly justified to believe a proposition ψ, while in fact, her believe in ψ is *not* justified. For example, she can be explicitly justified to believe a false belief in ψ if she justifies her belief in ψ on the grounds that a proposition ϕ is believed by John. If John retracts his belief in ϕ, then ψ becomes unjustified and possibly false, and if he does not inform Sarah about the retraction, then Sarah remains explicitly justified to believe ψ.

2.2 Theories of Truth

A theory of truth seeks to describe what it means for a proposition to be true. Philosophers and logicians distinguish different categories of truth theories, each defining the concept of truth differently. Early approaches to the semantics of propositions in the philosophy of language centre on the view that semantics is truth-conditional. Truth-conditions are thought of as theoretical entities that make propositions true or false, or, stated differently, truth-conditions are the entities that determine a statement's truth value. A theory of truth defines the general properties of propositions and their truth-conditions.

The oldest known theory of truth, Correspondence theory, dates back to Antiquity. As we will briefly describe in Section 2.2.1, this theory stems from the idea that truth corresponds to reality, and thus that truth-conditions are facts in reality. Because agents will not have direct access to reality, it is doubtful whether Correspondence theory can be useful for agents, who cannot have access to truth-conditions. Two competing theories in which the truth is in principle accessible to agents are more adequate to model an agent's conception of the truth. These two theories describe truth-conditions that are of an epistemic nature: Coherence and Consensus theories. The truth-conditions in Coherence theory (Section 2.2.2) state that the truth of a proposition is its coherence with other true propositions. The truth-conditions in Consensus theory (Section 2.2.3) state that the truth of a proposition is based on conventions of a group of agents.

Two other frequently used theories that we will only mention are the Redundancy theory and the Semantic theory of truth (see Horwich [Hor98]). These theories provide an account of truth that is non-epistemic, that is, the truth can be described without the need of an agent to be aware of it. Most classical logics adopt the Semantic theory, which defines truth as a function in a meta-language. The familiar example by Tarski [Tar56], " 'Snow is white' is true if and only if snow is white" expresses the idea that a proposition is true only if its translation in a meta-language is true.

2.2.1 Correspondence theory of truth

Dating back to Aristotle, many philosophers adhere to the idea that truth consists in a relation to reality. Correspondence theory states that truth-conditions are facts in reality and hence the theory defines truth as correspondence to a segment of reality. In this theory, a proposition is said to be true if it corresponds to a fact in reality, and is false if there are no such corresponding fact.

In epistemology, the doctrine of realism states that reality exists independently of its perception. Because the proponents of Correspondence theory presume that reality provides the truth-conditions for propositions, propositions can be true without ever being observed. Today, many philosophers believe in mathematical realism or mathematical Platonism, which holds that mathematical relationships exist independently of human thought. Realism is easy to adopt in relation to chairs and tables. The question of whether it is a truth that an object is a chair would be answered affirmatively if the object were a concrete example of the abstract chair or 'chairness'. The concrete object is called an individual; the abstract object is called a universal (see Klima [Kli04]). A realist is a philosopher who believes that universals are real and exist independently of anyone thinking of them. A realist adheres to the view that individual are instances of their universals, and that this is independent of whether agents know about the existence of these universals. An instance of a truth is taken to be a proposition that instantiates the universal 'Truth' in an objective manner. An agent could classify an instance into a wrong universal, or sometimes may not even know an instance's universal.

Problems with Correspondence theory concern the ancient metaphysical problem of universals, described by Klima [Kli04], which addresses the question what universals are supposed to be, or whether they exist. Because according to Correspondence theory, truth-conditions are facts in reality, and that reality exists independently of its perception, the truth may transcend an agent's capacity for knowledge: the truth need not be known by agents. An agent is justified to believe that a proposition is a truth if the agent knows that the proposition corresponds to a fact in reality. Demanding a strict correspondence relation between believing a truth and knowing a fact makes it rather difficult—if not impossible—for agents to have justified beliefs. See Latour [Lat99] for engaging reading about his question *"do you believe in reality?"* As defended by classical sceptics like Pyrrho[1], agents may never know reality, and, consequently, may never know the truth (see Derose [DW98]).

A practical problem with Correspondence theory is that some propositions have an inherent uncertainty. Most propositions referring to the past and future are uncertain because what they are said to refer to does not exist anymore or does not exist yet. In principle, the proposition 'John is a brave man' can refer to John's behaviour, which Sarah can observe and classify as truly brave. However, the statement that Fred, who lived in Antiquity and who is dead by now, was brave, can pose a challenge for agents, because there may be nothing left in reality to which the statement can be said to refer. Another statement that Sarah, who is still alive, is brave, can also pose a problem. To what does the statement that 'Sarah is brave' refer if Sarah had not yet encountered situations in which she can display brave behaviour? As long as Sarah does not need to be brave, she cannot be said to be brave; while if she has to be brave, she can be said to be brave. Inconsistent propositions, as will be discussed in Chapter 3, also fail to have a physical counterpart.

In *Truth and other enigmas* [Dum78], Dummett says that an agent is a realist if she thinks that propositions can be true or false even though she may have no way of ever discovering their truth. For a realist, the truth is considered objective and can transcend the agent's capacity for knowledge. The main problem with Correspondence theory is that some propositions do not have truth-conditions, which means that, in general, agents will be denied access to know the truth of propositions. This problem led philosophers to consider alternative ways to answer the question of what constitutes the truth. An alternative conception of truth is addressed in the next section on Coherence theory of truth.

2.2.2 Coherence theory of truth

Coherence theory defines truth as coherence with a specified set of propositions. That is, the truth-condition of a proposition is its coherence with a specified set. Coherence theory has several versions, differing with regard to two major issues: what the contents of the specified set should be, and who is supposed to know the

[1] Pyrrho, c. 360–c. 270 BC was a Greek philosopher from Elis, founder of the Greek school of scepticism.

specified set. An agent is justified to believe that a proposition constitutes a truth if what the agent believes coheres with the specified set (see Young [You02]).

According to some early versions of Coherence theory, the coherence relation is consistency (Davidson [Dav86]; Rescher [Res73]). To say that a proposition coheres with a certain specified set of propositions, in this view, is to say that the proposition is consistent with the specified set. This account of coherence is unsatisfactory for the following reason. If a proposition is consistent with the specified set, and the negation is consistent with the same specified set, then even if the proposition and its negation are inconsistent with each other, they are both truths. If coherence were consistency, the coherence theorist would have to claim that both propositions are true. In general, more sophisticated coherence relations are considered, such as strict logical entailment. According to these more sophisticated relations, a proposition coheres with a set of propositions if and only if it is entailed by elements of the set.

In contrast with Correspondence theory's emphasis on an independent reality, the coherence view supposes that true beliefs constitute an inter-related system in which each belief adds justification to all other beliefs, independent of a reality. The truth is in a sense arbitrary and not objective. This does not seem to do justice to the intuition that truth and reality are at least related.

Putnam, in *Reason, Truth and History* [Put81], defends coherence theories against accusations of arbitrariness by attempting to reconcile them with pragmatic theories. To briefly comment on Pragmatism, Peirce, who is considered the founder of Pragmatism, and later James (Menand [Men97]), formulated the pragmatic theory in which truth is defined as the success and utility of the practical consequences of an idea (James [Jam95, Jam97]). The idea that a belief is a truth is that acting upon the belief yields desirable and satisfactory results. Under this account of truth, an agent will be justified to call something true if inquiry has ceased to provide new insights into reality. Agents are said to know the truth if the agent's methods of enquiry concerning reality have been perfected and thus allow agents to know reality objectively. Putnam states that the specified set consists of those propositions that will be believed when agents with finite cognitive capacities have reached some limit of inquiry. Thus, the propositions that an agent may be justified to believe in the future, is the specified set, which provides the conditions for truths. This may make the conception of truth robust against accusations of arbitrariness; however, on this account of truth, just as with pragmatics, agents can never be sure whether propositions can be treated as knowledge or as possibly false beliefs that need further inquiry. In the next section, we address consensus theories of truth in which, in principle, agents can distinguish truths from false beliefs.

In a broad sense, truth maintenance systems (Doyle [Doy79]) implement the idea that the truth of propositions is the propositions' coherence with other propositions. Informally, a proposition in a truth maintenance system is true when sufficient other true propositions exist that add to the justification of the proposition. If the truth-status of a particular proposition is changed, the system will recursively update the truth-status of other propositions that depend on that particular proposition, thus maintaining the justification of true propositions. The truth of propositions is enforced by testing whether propositions that have lost parts of their justification still have sufficient justification such that they may be called true. Just as in Coherence theory, the truth of a particular proposition in a truth-maintenance system depends on whether a relation exists with other propositions that add justification to that particular proposition. Our agents could be closely modelled upon a truth-maintenance system in which an agent's belief would constitute a truth if it were based on other beliefs that are held true by other agents or the agent herself. Such a model would have profound implications for our agents because it would assume that agents have direct access to other agents' beliefs. This flies in the face of a fundamental assumption of agent systems, that a cognitive state is inaccessible to other agents. Moreover, such an approach does not explain how a community of agents can agree on the use and truth of propositions; a truth-maintenance system would only provide a method to calculate that propositions have sufficient justification to be regarded as truths. In the following section, we will explain how a community defines the use and truths of propositions through Consensus theory.

2.2.3 Consensus theory of truth

Consensus theory of truth defines the truth as something that has been agreed upon by a group. On this account, the truth-condition of a proposition is an agreement of a group of agents. This theory provides a subjective account of truth comparable to the subjective nature of Coherence theory: for both theories, the truth is independent of what may be said to occur in reality. Coherence theory allows the truth for one agent to be different from that of another agent. However, because the whole group agrees on what the truth is taken to be, Consensus theory does not permit agent specific accounts of truth. Yet it is still possible for a group to agree on something, while another group agrees differently. On a consensus account of truth, what an agent may call the truth may not be fully subjective, but may still be regarded arbitrary and irrespective of what can be said to constitute reality.

Wittgenstein's account of truth can be said to be a special type of Consensus theory. This theory is based on the dictum 'meaning is use', which states that agents learn what it means to call something 'the truth' by agreeing on the linguistic practices of their community. Ellenbogen [Ell03] gives a cunning twist to the agents' agreements on when they are justified to call a proposition a truth, while avoiding an arbitrary account of truth. Wittgenstein's account of truth will be discussed in the following section.

2.2.4 Wittgenstein's account of truth

Wittgenstein rejected the realist conception that the truth-conditions of propositions are facts in reality, and that the meaning of propositions is a reference to some segment of reality. He adopted an account of meaning which is based on use. This change of meaning led to his rejection of the realist conception of truth in which agents come to know their environments. Wittgenstein reasoned that if there is nothing more to meaning than use, then the concept of truth is nothing more than what agents can grasp through their use of those statements, which they treat as true. Ellenbogen, in *Wittgenstein's Account of Truth*, argues that

> the dictum "meaning is use" that what makes it correct to call statements "true" is our agreement on the criteria whereby we call them "true". [Ell03, p. xii]

This interpretation of meaning centres on the view that language users establish the meaning of statements with the criteria to use these statements. The dictum 'meaning is use' states that someone comes to understand what it is for a state of affairs to obtain by a theoretical description of the features involved. That is to say that a theoretical description provides a principal determinant, i.e. a criterion, to use a statement correctly. People use these criteria to understand which properties need to obtain for the correct application of statements.

According to this view, truth is acquired by learning how to *use* the predicate 'is true' for statements. It only makes sense to think of 'the truth' in terms of what humans can know, because it is only given such knowledge that they can learn what counts as establishing their truths. People learn when propositions establish a truth by participating in the linguistic practices of their community. That is to say, they become aware of the criteria that their community uses to establish truths. Within linguistic communities, people agree on correct uses of statements; it is this practice of agreeing upon criteria which statements they may treat as being true or false. The meaning of statements is not that some people use them in a particular way, but rather that their use is shared by a community. By learning a language, members of a community adopt the criteria to use the statements of the language by tacitly agreeing on the conventions of the community. Thus, the meaning of a statement is a community's tacit agreement on a criterion for its correct use.

Someone may adopt a rule that states in which situations she can use a proposition. Because this rule expresses *a* criterion to use a proposition, we will call this rule a *criterion rule*. We will say that someone has represented the meaning of a proposition if she has a criterion rule with a criterion that specifies the situation in which the proposition may be used, and this use has been agreed upon by her community. Because this rule uses *the* criterion for the correct use of a proposition, and this use is conventional, we will call this rule a *conventional rule*. People learn that criterion rules are conventional rules when they engage in the conversational activities of their community. Thus, a conventional rule states that the criterion associated with it has been agreed upon by the community to establish the truth values of the proposition.

For each type of statement, a community needs to have methods to determine whether they call statements true or false. A community agrees about what counts as an adequate test of any given type of statement. Different types of inquiry reflect different agreements on what is taken to count as adequate tests; these different agreements are referred to as language games. The members of a community are said to agree on the criteria to determine when it is correct to predicate '*is true*' of statements within a language game. Therefore, an agent who has mastered a language has learned when the sentence of the language can be used to denote statements, and because she is conversant with the rules of the language game, she has learned when a statement is called a truth. By learning the conventional rules for predicating *is true* of the sentences in a language that her community treats as being true or false, the agent acquires a conception of truth based on use. That is to say, as agents participate in the rule-following practices of a community, they play the language games of the community.

A use-based approach of meaning allows agents to construct explanations dependent on the truth or falsity of putatively undecidable propositions. An agent may have an understanding of an undecidable proposition because she knows when, according to her community, it is said to be true. An agent could even tell why attempts at verification are blocked. The statement 'Sarah is brave' and 'Sarah is not brave' can both be undecidable for John because Sarah may not yet have encountered situations in which she can display her braveness (cf. Section 2.2.1). John however may understand that both statements cannot be true together. John can be said to have a grasp of what would make each proposition true. Therefore, John's grasp of the truth-conditions of undecidable propositions is not exclusively his ability to tell whether they are true. The fact that John has a grasp of how to investigate the truth value suffices to attribute him an understanding of its truth-conditions.

Different language games provide different uses of statements. The language game of *is true* describes the situations in which agents are, according to their community, correct to use the predicate *is true*. Other games describe when agents are entitled to correctly predicate *is real* and *is fair*. Note that the language games of *is real* and *is true* need not coincide. That is, the theory neither denies nor presumes reality. Other language games may describe when agents are entitled to predicate epistemic propositions correctly, such as *to know* and *to believe* propositions. These will be presented in the following section.

2.3 Agents and a Use Account of Truth

Wittgenstein, in *Philosophical Investigations* [Wit01], describes the idea that meaning is use. According to Ellenbogen [Ell03], this should be interpreted to import that a criterion to correctly use a statement is the criterion to determine the truth of the statement. We will extend this idea and describe what it means for agents to predicate that they believe and know propositions. In the remainder of this thesis, we will use this use account of truth to provide our agents with operational definitions for justified beliefs.

2.3.1 Agents and a use account of belief

We adopt the view that the meaning of statements equals the conventional rules that a community has tacitly agreed upon to establish their truth (see Section 2.2.4). Similar to the language game to correctly predicate that statements are true, in the following paragraphs, we describe when agents are explicitly justified to predicate epistemic statements, such as *to believe*, correctly according to their community's agreed upon use. Next, we provide the mechanism that describes the meaning of beliefs.

An agent will be explicitly justified to believe a statement if she regards herself entitled, according to her community, to predicate *to believe* the statement. Just like any other statement, an agent is correct, according to her community, to use a statement if the criterion of its application has been met. We take it that the criterion to predicate *to believe* a statement is, among others, the criterion to establish the truth of the statement.[2] However, being justified in principle to believe a statement, as illustrated in Section 2.1 with Fermat's last theorem, is not a sufficient condition for believing the statement. Whether an agent actually has a belief when she is justified to have it, depends on whether she applies the processes of her cognitive state that can change her mental state to reflect that she believes the statement. If, in addition to being explicitly justified to believe a statement, an agent has the opportunity to come to believe it, she will be said to be in the state of believing the statement.

An agent need not retract her beliefs if, according to the agent, the criteria to establish the beliefs cease to hold. If a criterion ceases to hold, the agent will be considered to have an unjustified belief. We consider it good practice to provide different language games to predicate *to believe* a statement, and *to be ignorant* about a statement. An agent is explicitly justified to be ignorant about a statement if the agent regards herself entitled, according to her community, to predicate *to be ignorant* about the statement. If an agent is explicitly justified to be ignorant about a statement, and the agent is given the opportunity to change her cognitive state accordingly, she is said to have forgotten the belief. We assume that the criteria that entitle agents to predicate *to believe* and *to be ignorant* about a statement cannot be met simultaneously. That is to say, agents cannot be said to believe a statement and to be ignorant about that statement at the same time.

Under Ellenbogen's interpretation of Wittgenstein's account of meaning, and taking an agent's mental processes into account, Sarah may regard herself explicitly justified to believe a statement if:

1. She regards herself entitled, according to her community, to predicate *to believe* the statement, and she has had the opportunity to change her mental state accordingly.

2. She did not have the opportunity to change her mental state, even though she regards herself entitled, according to her community, to predicate *to be ignorant* about the statement.

If an agent were asked what it means to believe a certain statement, then she can reply with the criteria that entitled her to predicate *to believe* and *to be ignorant* about the statement. However, if an agent were asked *why* she believes a certain statement, then she can reply with the circumstances that made her explicitly justified to adopt

[2] In Chapter 6, we will broaden the criterion to believe statements to include that agents may conform to beliefs of other agents, and that agents may infer beliefs from existing beliefs.

the belief. More specifically, she could answer that at a certain moment in the past, the criterion to believe the statement had been met, and that ever since, she has not been justified (or additionally not given the opportunity) to become ignorant about the statement.

We could provide Sarah with an operational definition when she is explicitly justified to predicate that John is said to have beliefs he is justified to have, and beliefs he is not justified to have. If Sarah has an operational definition, then she can be said to have an understanding in which situations John has justified and unjustified beliefs. If Sarah is justified to believe that, first, the criterion to retract a belief has been met for John; second, John had the opportunity to change his mental state accordingly, and third, John did not change his cognitive state accordingly, then arguably, Sarah is explicitly justified to believe that John is not justified to have the belief. Even though John can be said to have this unjustified belief, he can be explicitly justified to have it, because it could be that he cannot know better. The same holds for Sarah.

Similar to the language game to predicate correctly to believe a proposition, a community can agree upon what it means for Sarah to believe that John has a belief. Sarah can be said to be explicitly justified to have a belief regarding a belief of John, when John has attested to have the belief. Similar to Sarah's own beliefs, Sarah predicates correctly that John has a belief if the criterion to determine that John has such a belief has been met. Sarah explicitly justifies her belief that John has a belief with John's behaviour.[3] The behaviour of a system may convince Sarah to adopt a belief if the agent's community has agreed that this behaviour satisfies the criterion to predicate that the system 'believes'. Analogous to Turing's idea of an 'imitation game' that can be said to provide a criterion to call a system intelligent (cf. footnote 4 on page 3), in a similar fashion, a community could agree on a criterion when a system is said to believe a proposition. If a human who according to the criterion of an interrogator's community believes a statement cannot be distinguish (by the interrogator) from a machine that attests to believe a statement, then the interrogator could be said to be justified to believe that the machine believes the statement. Note however that because the truth is not grounded in reality, the criterion to predicate *is true* and to predicate *is real* need not be equal. Consequently, if Sarah is justified to believe that John believes a statement, she can still meaningfully question whether John really believes the statement. We will further address the decoupling of the grammars of *is true* and *is real* in Section 3.3 on a use interpretation of believing inconsistencies.

2.3.2 Agents and a use account of knowledge

Analogous to what it means for an agent to have a belief, an agent is justified to know a statement if, according to the agent, the criterion to predicate *to know* the statement has been met. If the agent has had the opportunity to change her cognitive

[3] It could even be argued that if Sarah cannot remember that she believes a statement, but she observes she behaves as if she believes it, she may be justified to predicate *to believe* the statement.

state accordingly, she can be said to know the proposition. The only difference between an agent knowing a statement and her believing it, is that, in the former case, the criterion allows her, in conformity with her community, to predicate *to know* a statement, while in the latter case, the criterion allows her to predicate *to believe* a statement. In popular speech, beliefs are often said to be refutable while knowledge is said not to be; nevertheless, the criteria for believing or knowing define when agents are entitled to believe or know something.

An agent may regard her knowledge that another agent knows or believes a statement as correct if the agent has acted (to change her mental state to reflect that she knows) in accordance with the use of the predicate *to know*. If a community agrees that it is impossible for agents to know other agents' mental states, then agents are not allowed to say that they know such things, and if they do, they are not seriously saying they *know* what other agents' mental states are, they are merely joking. Other communities may agree differently and provide a use in which agents are allowed to predicate *to know* that other agents have beliefs, for example by adopting the intentional stance (Dennett [Den81]). The intentional stance states that agents may ascribe mental predicates to entities with the aim of understanding and predicting these entities (cf. footnote 3 on page 3). What matters for the agent designers is that with these epistemological foundations, the criteria for predicating to know that another agent believes or knows a statement is constructive. For example, Sarah could be correct to know that John did not commit a certain crime, based on his testimony that he did not do it. Sarah's community could augment the truth-conditions—and thus alter the meaning of knowing the truth in this particular case—with a positive result from a polygraph that John is not lying. A community could also agree that a polygraph is not such a useful and predictive instrument after all, and, become to use a brain-scanner to define what John 'really' believes.

Whether the epistemic modalities, such as believing and knowing, are formalised with a certain formal system is not important for the question when agents are justified to believe and know propositions. We could use systems of knowledge satisfying the axiomatic system S4 that corresponds to modal logic (Kripke [Kri63]) to formalise an agent's knowledge, or we could use axiomatic system KD45 often referred to as doxastic logic (see Meyer and van der Hoek [MvdH95]) to represent the agent's belief state. What does matter is that a community agrees which criterion defines the correct use of the predicate *to believe* and *to know*. In Chapter 4 we will describe our formal system that can handle epistemic modalities such as beliefs and ignorance.

Next, we briefly describe knowledge held by groups of agents. Informally, a statement is *mutually known* among a set of agents if each agent knows the proposition. Common knowledge is a little more complicated. Lewis, in *Convention* [Lew02], gives an intuitive and explicit analysis of common knowledge as a hierarchy of statements of the form 'agent *i* knows that agent *j* knows that ... knows a proposition.' For our current analysis it would not matter whether agents use: either Schiffer's [Sch72] approach in which epistemic logic (Hintikka [Hin62]) is used to define common knowledge as a hierarchy of statements; or Barwise's [Bar88] formal approach in which common knowledge is analysed as fixed points; or Aumann's [Aum76] approach with algorithms for determining what information is commonly known by Bayesian agents. Just as for the modalities of believing and knowing propositions, for the remainder of this chapter, it does not matter how the concept of common knowledge is formalised. What does matter is that a community of agents has agreed upon the criterion to call a piece of knowledge common knowledge.

2.4 Agents and Meaning

2.4.1 Agents understanding meaning

According to Wittgenstein, an agent is said to grasp the meaning of a statement if she uses the statement according to the criterion that has been agreed upon by her community. An agent need not be aware that she uses a statement in accordance with a use that her community has agreed upon. That is to say, she has a criterion rule that allows her to use the statement conform to the conventional use. An agent becomes aware of the meaning of a statement if she may predicate, according to her community, that her use of the statement is conventional.

Although agents may be justified to know the meaning of statements, other agents in their community can still use these statements differently. Knowing the meaning, i.e. the agreement on the use of a statement shared by a community, still allows the members of a community to have different conceptions of what this shared use is supposed to be. This is a direct consequence of the dictum 'meaning is use' which allows agents to know a statement if the criterion to do so has been met. If an agent regards herself entitled to know that a statement is used in a certain situation that is described by a certain criterion that everybody tacitly agrees upon, then the agent can be said to know the meaning of the statement. The agent will then be said to know that a criterion is part of a conventional rule that governs the correct use of the statement. If criteria are agreed upon tacitly, agents can be said to know that other agents use the same conventional rules although they never explicitly agreed upon them.

Our agents need not doubt their knowledge about the meaning of statements as long as they do not observe difference in the way they use the propositions. This is an important feature, because it could allow meaning to change gradually, and to be learned by new members of the community. Agents who are learning to use statements, e.g. to predicate that they know that other agents believe something, may make mistakes as seen from another agent's perspective. If Sarah and John observe that they have a different criterion for the use of a statement, the agent with the most authority may correctly predicate that the other has made a 'mistake'. Assume Sarah teaches John, and Sarah is a leading authority on some subject. Sarah may encourage John to change his rules such that he will correctly predicate, e.g. *to believe* certain statements in more situations. That is to say, John will know that a different criterion is conventional, and this criterion will make him more often correctly use a certain predicate.

In which situation can an agent know that what she knows is the meaning of a statement is in fact just a belief about the meaning of the statement? Stated differently, when can agents conclude that what they assumed to be tacitly agreed upon by their community, is in fact not a convention, and is not agreed upon explicitly by everyone?

2.4.2 Changing the meaning

An agent may revise her judgement that a certain criterion rule is conventional and thus change the meaning of a statement associated with that rule by judging that another criterion rule is conventional. An agent's judgement that a criterion rule is shared and conventional can be made only after the agent has come (as we have defined) to know that the rule's criterion can be used to establish the associated proposition. An agent cannot come to know that a criterion is used in her community to establish a statement if she does not know that the criterion can be used to determine the proposition in the first place. That is, an agent must first come to learn a relation between a criterion and a statement, i.e. a criterion rule, and later come to learn that this relation is conventional and in Wittgenstein's terminology a 'form of life'.

> "Agreement in form of life is logically prior to agreement in opinions. For agreement in form of life is our agreement on a shared world picture. And this world picture forms the inherited background against which we distinguish true and false (O.C. #94)"[4] Ellenbogen [Ell03, p. 4]

Agents can change conventional rules. Consider the following example.

> "From the beginning of the seventeenth to the late nineteenth century, the defining criterion of gold was solubility in *aqua regia*. But in the nineteenth century it was discovered that gold had the atomic number 79, a feature not exclusively correlated with the feature of solubility in *aqua regia*; that is, the former criterion was discovered to have room for other non-"noble" metals whereas the latter criterion defined 'gold'." [Ell03, pp. 9–10]

The first person to discover the symptom that gold has the atomic number 79 did not know that this symptom would become the criterion to tell gold from other metals. This person had to convince herself that this atomic number is a symptom more indicative than the criterion of solubility in *aqua regia*. Only then could this person try to convince other people who were still using the criterion of solubility in *aqua regia* that this criterion was less useful than the atomic number 79. The precise number of renowned chemists that had to be convinced before the old criterion was degraded to a symptom is hard to tell; nevertheless, the precise criterion to adopt new use may not be of particular importance. Wittgenstein's observation is that

[4] O.C. #94, Wittgenstein in *On Certainty* [Wit72].

these changes of language games give a predictive and understandable account of how humans change the meaning of sentences and statements. For example, students in chemistry may not be able to distinguish fool's gold (iron sulphide) from what their lecturer calls 'gold' because it has the lustre of gold ore. If the students accept the lecturer to be an authority, then they accept that the atom number equal to 79 is the criterion for what their community calls 'gold'.

2.4.3 Dispute the meaning

Next we describe that even though agents are entitled to know the meaning of a statement, they may still dispute whether their criteria to use a statement is *the* criterion to correctly use the statement. Agents may be entitled to know how their community uses a statement; nevertheless, they may encounter other agents who use the statement differently.

If Sarah believes something that contradicts something that John believes, and both are aware they disagree, then both may desire to resolve their disagreement. In such a situation, Sarah has an indication that either of them holds a belief that they are not justified to have, or either of them may adopt a belief that they are justified to have. The same holds for John. Sarah may try to convince John to retract or adopt certain beliefs with the intent to resolve the disagreement. Sarah could also test her own beliefs, which may result in her being justified to retract or adopt beliefs that would resolve the disagreement. If all four possibilities have been pursued, yet have not resolved the disagreement, then, from Sarah's perspective, the disagreement can be said to be irresolvable. These contradictions, disagreements and irresolvable disagreements will be further addressed in Chapter 8.

The language game of predicating *to believe* a statement consists of conventional rules, and thus, as seen from an agent's perspective, and all agents use the same criterion rules, all agents should be entitled to believe the same statements. If, according to the rules of this game, Sarah is justified to believe that a statement is true, and, as far as she can tell, John is justified to believe that the statement is false, and thus they disagree over this statement, then the following two situations can be the case. Either one of the agents has used knowledge to infer beliefs that resulted in their disagreement. Or, the agents have a different account of believing, that is, they have different criteria to come to believe statements. That is, as seen from Sarah's perspective, the assumed conventional status of the criterion rules that both John and Sarah use to predicate 'to believe' are not conventional.

If agents are taken to represent a domain of relevant knowledge, then, on behalf of their domains, they may provide epistemic judgements about what is considered true and false. However, these agents may not change the knowledge of their domains. If agents encounter irresolvable disagreements, then, from their perspective, this results from an incorrect assumption (made by domain experts) that a conventional rule is shared by all the agents. In such a situation, the responsible experts must be consulted because the knowledge is their responsibility.

2.5 Concluding Remarks

In this chapter we presented the philosophical foundations that we use to understand what it means for our agents to believe a statement, to have a piece of knowledge, and to know the meaning of statements, and as we will later describe, the meaning of decisions and speech acts.

We described three theories that define what it means for a proposition to be true. Correspondence theory of truth states that a proposition's truth-conditions are facts, and hence defines truth as correspondence to a segment of reality. In this theory, the truth-conditions are not part of an agent's cognitive state, and consequently, the agent does not have direct access to the truth-conditions to verify the truth. The two other theories are epistemic theories of truth that define that truth-conditions are part of the agent's cognitive state. Coherence theory defines the truth-condition of a proposition as the coherence with a specified set of propositions held by an agent. The truth is arbitrary and not objective in the sense that different agents can have different specified sets and thus have different accounts of what it is that constitutes the truth. Our agents are taken to adhere to Wittgenstein's use account of truth. Under this account, the truth-condition of a statement is the criterion that an agent uses to predicate that the statement is true. According to the agent, this criterion has been agreed upon by her community as to specify the situations in which she is correct to predicate that the statement is true. The statement is true for the agent if the statement's criterion has been met in her cognitive state. Such truths could be considered arbitrary, independent of what could be called reality, but unlike Coherence theory, in which agents can have different conceptions of what the truth is supposed to be, under a Wittgensteinian use account of truth, for agents the criterion when to predicate *is true* is shared.

We extended Ellenbogen's interpretation of the Wittgensteinian use account of truth to allow agents to predicate *to believe* and *to be ignorant* about statements. Additionally, we described what it means for our agents to have knowledge and what it means to know the meaning of statements.

Chapter 3

Paraconsistent Logics

> *Doublethink means the power of holding two contradictory beliefs in one's mind simultaneously, and accepting both of them.*
>
> George Orwell, *"1984"* [Orw48]

More often than they may wish for, agents and humans are confronted with contradictory information. Such pieces of information may guide agents and humans to believe things that they assume cannot be real. For example, a doctor may find no overt signs that a patient suffers from a certain disease, yet lab results indicate that the patient does suffer from that disease. We may sometimes jump to conclusion, in this situation that either the doctor's judgement or the lab result is wrong. We may sometimes also deliberately hold contradictory beliefs and doublethink if this serves a purpose. In this chapter, we describe what it means for our agents to have contradictory beliefs and what we are to do if our agents become explicitly justified to have such contradictory beliefs.

In Section 3.1, we describe several circumstances in which our agents become explicitly justified to decide to believe inconsistent statements, and come to have an inconsistent world picture. In Section 3.2, we provide a rationale to justify the existence and application of inconsistencies. We then turn in Section 3.3 to a use interpretation of contradictory beliefs, after which we in Section 3.4 describe paraconsistent propositions and two fundamental assumptions of logic that need to be challenged to consistently represent inconsistent propositions in logics.

3.1 An Unavoidable Inconsistent World Picture

If Sarah is explicitly justified to believe a statement, that is, she regards herself entitled, according to her community, to predicate *to believe* a statement, then she may, as described in Section 2.3.1, change her cognitive state such that she can be said to believe the statement. Because a community agrees on the criterion for the correct use of statements, the community can agree that Sarah correctly uses a statement based on her beliefs about John's beliefs. Thus, agents may use their beliefs about other agents' beliefs as grounds to justify their own beliefs. The question we address is how Sarah should cope with situations in which different sources supply justifications that would lead her to believe contradictory statements. If a source supplies Sarah with a justification for a statement, and another source supplies a justification for the statement's negation, then, Sarah may be justified to believe both statements, and thus be explicitly justified to have an inconsistent belief.

A possible argument for Sarah not to come to believe inconsistent statements is that she could postpone believing one of the statements, and thus postpone believing an inconsistency. If a source that supplied Sarah with a justification cannot justify its own information, Sarah's postponement may give the source time to revise its unjustified information, which would obviate Sarah's need to believe the inconsistent statement. The validity of this argument is predicated under the assumption that at least some of the source's information is unjustified, and this unjustified nature will be revealed, and resolved, eventually. Sarah, who is justified to believe inconsistent statements, may inform the sources that they provide her with contradictory information. While the sources are engaged in resolving their disagreement, Sarah is arguably justified to postpone her decision until the sources have agreed on their opinions. The net effect would be that Sarah does not have to entertain inconsistent statements, although she is justified to have them, when the inconsistency is going to be refuted in the future. If the source's definitive answer would take a long time, contradictory propositions can be considered to have an undetermined epistemic status, associated with an unknown truth value.[1]

How should Sarah act if she believes the disagreeing information is *not* temporary but permanent? It would not be satisfactory if Sarah postponed her decision to believe both propositions because this would be equal to renouncing to adopt a belief indefinitely, while she is explicitly justified to believe both contradictory statements. The question is rephrased as to whether situations can occur in which sources disagree and their disagreement is final.

Assume that the two sources that provided Sarah with evidence to believe inconsistent statements are John and Fred. If Fred and John have pursued every method at their disposal to resolve their differing belief, but failed to do so, then they may settle their differing opinion with an agreement. In this agreement Fred and John acknowledge that they have differing beliefs, that is, they agree to disagree. Informally, if two agents have a disagreement and their opinions will not change when they are confronted with new information, then it can be considered rational for them to agree to disagree. That is to say, the two agents agree that the information

[1] Or instead of the epistemic status equal to 'unknown', such statements can be considered to posses truth values associated with truth value 'true by default'. Such default truth values are considered to be either refuted and recalled, because they only have a tentative epistemic truth status, or they are confirmed in which case their tentative status is eventually overwritten by a permanent one.

that could have resolve the disagreement, as seen from their perspectives, has not helped to convince themselves, and the other, to adopt or retract beliefs such that the disagreement is resolved. In Chapter 8, we will discuss the precise criterion to be justified to agree to disagree. If John and Fred have agreed to disagree about a proposition, then their difference of opinion is permanent. If Sarah is justified to believe inconsistent statements based on John and Fred's belief in a proposition, and Sarah believes that John and Fred have agreed to disagree about the statement, then she is aware their differing opinion is final, and thus that she is justified to believe an inconsistency.

3.2 Sources of Inconsistencies

Several philosophical considerations have been put forward to comprehend why true contradictions or *dialetheias* exist. In the fourth century B.C., the Greek philosopher Eubulides of Miletus described a paradigmatic example of a true contradiction in the Liar Paradox. The paradox results from evaluating the truth value of the following self-referential sentence *'I am lying now'*. This sentence and equivalent sentences, such as in (3.1), are called Liar Sentences.

"this statement is not true" (3.1)

Logicians, among others, have often treated inconsistency as an undesirable property of logics. Other people with practical concerns often treat inconsistency as a state of a system that must be contained and handled with care. Software system designers for example accept that large amounts of data contain inconsistencies. These inconsistencies result from, for example, human errors when entering data such as patient information. Inconsistencies may also originate from sensor malfunctions. If software is to handle data that result from interaction with an environment such as vision or audio, inconsistencies are bound to emerge eventually when the data are processed. These inconsistencies do not reflect an inconsistent environment, but technical or human error that can, in principle, be prevented and resolved. While such errors are being investigated and resolved by either agents or their responsible human experts, the software, i.e. agents, should handle inconsistencies in a satisfactory fashion. Bertossi et al. [BHS04] for example, design databases that are tolerant to inconsistencies. These databases should manage information with inconsistencies without generating spurious answers to queries. Paraconsistent logics should behave similarly; they should allow inconsistent propositions without generating spurious proofs.

The domain knowledge that our agents represent originates from experts in psycho neurological endocrinology. In this domain, contradictory beliefs can emerge when explaining human depression. The following example may exaggerate and oversimplify the scientific knowledge from these domains; however, it captures the essence that scientific terms may be used in similar situations while having different meanings. Domain experts working in the area of clinical psychology may describe depression in terms of the person's psyche in relation to her environment and especially to other humans. For this expert, a depression is the result of chronic stress. Another expert from the area of bio-psychological endocrinology would not use psyche-related terms to understand and explain depressions. Such an expert understands depression as a function of neurotransmitters produced by the brain and the effects on a person's behaviour when these neurotransmitters bind to receptors. For this expert, a depression is the result of excessive secretion of neurotransmitter cortisol in the brain. If both experts were to consult each other, which knowledge would be most valid? They may come to understand that cortisol and stress are somehow related[2], but they could just as well consider that the rule of the other is incompatible with theirs. In particular, for the company Emotional Brain, such heterogeneity of domain experts' opinions necessitates a software system that handles possible inconsistencies.

3.3 A Use Interpretation of Believing Contradictions

In this section, we provide an interpretation of believing contradictory statements based on use. We then describe when beliefs in contradictory statements can be used to steer an agent's behaviour.

Correspondence theory of truth, as described in Section 2.2.1, defines truth as correspondence with a segment of reality. Under this account, an agent is justified to call a statement a truth if what the statement corresponds to is real. That is, if an agent can call a state of affairs real, then she may call its corresponding statement a truth. Conversely, if an agent may call a statement a truth, then she may call its corresponding state of affairs real. Analogously, if an agent can call a state of affairs not real, then she may call its corresponding statement false, and if she calls a statement false, then she may call its corresponding state of affairs not real. Let us call the rules that agents use for the application of these predicates their grammars.

Assume that Sarah adheres to the view that reality is unique, that is, she may not call a state of affairs real and not real simultaneously. In other words, according to Sarah, reality exists only in one particular way. Additionally, assume that Sarah adheres to an account of truth as described by Correspondence theory. Under these assumptions, if Sarah is to call a statement inconsistent, then she should call a state of affairs both real and not real. According to Correspondence theory, the coupling of the grammars of *is real* and *is true*, and the grammars of *is not-real* and *is false* prohibit statements to be both true and false, because this would fly in the face of the assumption of a unique reality. Thus, under the assumption of a unique reality, Correspondence theory cannot meaningfully allow our agents to have inconsistent beliefs.

We therefore adopt a use interpretation of meaning, as described in Section 2.2.4. This allows a community of agents to agree on the criteria to use the predicates *is*

[2] Contemporary psychology, e.g. Kolb and Wishaw [KW03], relates cortisol and stress; cortisol can thus be called a stress hormone.

true, *is false*, *is real* and *is not-real*. If agents agree to use the predicates *is true* and *is real* differently from usage imposed by Correspondence theory, then agents could meaningfully call a statement true and false without predicating to call a state of affairs real and not real simultaneously. We assume that our community of agents agreed that the grammar of *is true* is only *intended* to coincide with the grammar of *is real*. Thus, the truth-condition for a proposition is not that it corresponds to something real, but that the criterion to call a statement a truth is intended to be equal to the criterion to call a state of affairs real. If the intended equality of both criteria fails, then agents can meaningfully predicate that a statement is inconsistent without necessarily predicating that that reality is not unique. Thus, under a use account of truth, a community can agree on the criterion to call something a truth that allows agents to meaningfully believe inconsistent statements, and at the same time meaningfully believe that their world is unique.

Given that with a use account of truth, agents can meaningfully believe inconsistent statements, agents can use these inconsistencies to avert them. An agent's belief in inconsistent statements does not declare her world inconsistent, but merely refutes the assumption that the rules to talk about the truth never lead to contradictions. We take it that our agents assume a unique reality. If, as described in Section 3.1, Sarah is justified to believe that a statement is both true and false, then, under the assumption of a unique reality, she may conclude that either the sources have provided her with unjustified information, or that the criteria that entitle her to believe the inconsistency cannot be used to talk about reality. Because we assumed that our community of agents has agreed that the grammar of *is true* and *is real* is intended coincide, that is, agents should be justified to predicate *is true* and *is real* of a statement in the same circumstances, then an agent's belief in inconsistent statements is a motive to enquire into the sources of her beliefs with the aim to refute the inconsistency.

The theory of belief revision provides a systematic analysis of *preventing* databases or sets of statements from becoming inconsistent. Two theories prevail on how to modify descriptions of a world in the light of new information: the theory of revision by Gärdenfors [Gär88], and the theory of update by Katsuno and Mendelzon [KM91]. Both theories formalise how an agent has to change a theory or her belief state when new information has to be incorporated. However, a fundamental distinction exists between the two types of modifications. The first type, *revision*, is used when agents obtain new information about a static environment. The second type of modification, *update*, consists of bringing a belief state up-to-date when an environment that it intends to describe, changes. As observed by Katsuno and Mendelzon [KM91], the difference is temporal and focal: revision describes the change of the agent's description of her environment which, presumably, *has not* changed, while update describes the change of the agent's description of her environment which, presumably, *has* changed.

An example of revision is when Sherlock Holmes learns that the murder weapon was used left-handedly, and he thus revises the set of possible suspects to those who are left-handed, leaving, for example, the butler as the last suspect. The world in which the murder was committed did not change, only Sherlock's beliefs regarding the murderer have changed. A theory is updated when data from a sensor overwrite the sensor's previous data: a sensor is said to update its data. Other examples of this variety are database updates: if Sarah enters "transfer 200K euros to John's account", then she updates John's account information.

In contrast, instead of preventing these undesired inconsistencies, agents may need to accept them and deal with them later. Our agents will be equipped with a use account of truth, which will enable them to be explicitly justified to believe statements in accordance with their community, that is, conform to the tacit agreement of the experts they are representing. From the assumption that the grammar of *is true* intends to coincide with the grammar of *is real*, agents can meaningfully believe that a statement is inconsistent while also meaningfully assuming that their environment cannot be inconsistent. When Sarah revises her cognitive state to believe inconsistent statements, she can be taken to admit, knowingly, to have a belief about her environment that does not tally with her assumption of a unique reality. Her inconsistent belief may motivate her to enquire into the sources of her belief with the aim to refute her inconsistent belief. When new information about her environment becomes available, Sarah may revise her beliefs and possibly retract the inconsistent belief, and replace it with a consistent one. Sarah was never justified to believe that her environment is inconsistent.

Assume that Sherlock Holmes does some belief revision. Mr. Holmes starts out with two suspects for his murder case: the butler and the gardener. Because Holmes does not believe that the butler has done it, nor does he believe that the gardener has committed the crime, the situation is characterised by an uninformed Holmes. We could say that Holmes is entitled to predicate *to be ignorant* that the gardener committed the crime, and the same for the butler. Because Holmes knows that somebody must have committed the crime, and because his suspects are the only two left who could have done it, he is motivated to enquire with the aim to learn new information such that he is justified to adopt the belief that either the butler or the gardener has done it. However, the situation is characterised as over-informed if Holmes has information that justifies him to believe that both the butler and the gardener have done it. For example, Holmes may believe that the murder weapon is used left-handedly and the butler is the only left-handed suspect, thus Holmes is explicitly justified to believe that the butler has done it. Additionally, the gardener has confessed that he has committed the crime, Holmes is thus explicitly justified to believe that the gardener has done it. Holmes knows that not both could have committed the crime, thus he is motivated to enquire with the aim to learn new information such that he becomes explicitly justified to retract either the belief that the butler has done it, or the belief that the gardener has.

As said, an agent's belief in inconsistent statements is a motive to enquire into the sources of her belief with the aim to refute the inconsistent belief. If this inquiry does not resolve the situation, as described in Section 3.1 on an unavoidable inconsistent world picture, then the inconsistent statement is a motive to revise the grammar of *is true*. That is to say, belief in inconsistent statements has a second use: such beliefs motivate communities to revise their use of the predicate *is true* and *is false* such that they cannot be predicated simultaneously.

3.4 Paraconsistent Logics

The previous section described that agents may meaningfully believe inconsistent statements, and that such beliefs may serve the purpose of guiding the agents' activities to resolve such beliefs. In this section, we describe how to formalise inconsistent statements. When dealing with logics, we will use the term 'proposition' interchangeably with the term 'statement'.

Greek philosophers found contradictions of great interest; however, a formal analysis only dates back to Vasiliev's imaginary logic (Bazhanov [Baz98]) at the start of the twentieth century. Vasiliev, among others, challenged the logical principle that anything can be proven from absurdity, *ex contradictione quodlibet* (ECQ). In standard logics, such as classical, intuitionist and modal logic, if inconsistent propositions are asserted, then with ECQ any proposition can be proven. A paraconsistent logic is taken to be a logic that allows inconsistent propositions to be asserted without becoming trivial. That is to say, a paraconsistent logic does not allow that anything can be proven from an absurdity. In standard logics, contradictory assertions have no model, viz. no representation of contradictory information. To allow logical theories with inconsistent propositions while preserving a model, inconsistencies have to be represented consistently. A paraconsistent proposition refers to inconsistent propositions that are part of a consistent theory.

Suber [Sub97] and Andrews [And96] discuss two fundamental assumptions underlying the very fabric of logic, which are revealed by investigating the negation operator: the principle of non-contradiction and the principle of excluded middle. Given a proposition p and its classical negation $\neg p$, the principle of non-contradiction asserts that *at most* one statement is true and that both can be false. Thus, the principle says that if a proposition p is true, then $\neg p$ cannot also be true. This flies in the face of any attempt to model a paraconsistent logic that allows both p and $\neg p$ to be true, and can represent inconsistent propositions without absurdity following. For our logic to cope with inconsistent propositions, we will introduce a special truth value that will represent a proposition's inconsistent truth-nature consistently. To allow such an extra truth value, another fundamental assumption of classical logic, the principle of excluded middle, has to be challenged.

Inverse to the principle of non-contradiction is the principle of excluded middle, which asserts that *at least* one statement p or $\neg p$ is true and that *not* both can be false. The principle says that if p is not true, then $\neg p$ must be true, and if p is false, then it cannot be that $\neg p$ is also false. Combining the two principles leads to the classical two-valued logics in which at most and at least one proposition p or $\neg p$ is true. Thus either p is true and $\neg p$ is false, or p is false and $\neg p$ is true. Rejecting the principle of excluded middle will allow logics with additional non-classical truth values. The truth values to represent a lack of information that is associated with an information state 'undefined' or 'unknown', and truth values that represent an inconsistent information state can then be added. Kleene [Kle50] provided a three-valued logical system in which a new truth value to represent a lack of information is introduced. A truth value can be added, in a similar fashion, to a logical system to represent inconsistent information. Belnap, in *A useful four-valued logic* [Bel77], does precisely this: he adds two truth values to represent unknown and inconsistent information states.

In Chapter 4, we will reject the principle of non-contradiction and the principle of excluded middle to construct logics that allow paraconsistent propositions. The theories of these logics will have the following properties. If a theory rejects the principle of non-contradiction, then a proposition p, and, at the same time, the negation $\neg p$ may follow from the theory. In addition, if the theory rejects the principle of excluded middle, then if a proposition p does not follow from the theory, then its negation, $\neg p$, does not need to follow from the theory at the same time. That is to say that the theory allows a third truth value associated with a proposition other than true and false. Such a third, non-classical truth value can be used, for example, to represent the contradictory information state of inconsistent propositions. If a proposition p follows from a theory, and, at the same time, the negation $\neg p$ follows from the same theory, then the corresponding paraconsistent proposition is also said to follow from such a theory. Equally, if a paraconsistent proposition follows from a theory, then both the corresponding p and $\neg p$ follow from that theory. A theory that consistently represents inconsistent propositions need not be inconsistent itself. The theories that we will define in the following chapter will not allow situations in which p follows from a theory, and, at the same time, p does not follow from the theory. Analogously, it is not allowed that $\neg p$ follows from a theory, and, at the same time, $\neg p$ does not follow from the theory. The formal semantics of propositions, including paraconsistent propositions, will be presented in Chapter 4 on multi-valued logics.

3.5 Concluding Remarks

We described when our agents become explicitly justified to decide to believe inconsistent statements. Sarah is explicitly justified to decide to believe an inconsistent statement if two sources provide her with information that, conform to Sarah's community, entitles one source to predicate that the proposition *is true*, and simultaneously entitles the other to predicate that the statement *is false*. If, in addition to Sarah's justification to believe that the statement is both true and false, the sources have agreed to disagree about information, that allows Sarah to believe the inconsistent statement, Sarah has no choice but to believe the inconsistency.

Just as any consistent proposition, the 'meaning is use' interpretation provides paraconsistent propositions with a meaning. The meaning of a statement is the criterion for its correct use that a community has agreed upon. If a community agrees that the use of believing inconsistent statements is that it can be established when a certain criterion has been established, then an agent can use this belief, for example, as a motive to revise beliefs that resulted in believing the inconsistency.

However before the agent can revise her belief in the inconsistent belief, she may need to represent this belief consistently. Two fundamental properties of logic, viz., the principle of non-contradiction and the principle of excluded middle, have to be challenged to allow a logic to represent an agent's inconsistent belief consistently.

Chapter 4

Multi-valued Logics

A multi-valued logic is a logical system in which the truth value of a proposition is not restricted to the classic truth values 'true' and 'false'. Other values can be introduced to represent alethic modalities of necessarily and contingently true and false, or probabilistic modalities of certainly and possibly true and false. A different type of epistemic modality is ignorance, inconsistency and bias. Because agents are entities with epistemic attitudes, logics with the latter epistemic modality are particularly suitable to describe (multi) agent systems. In this chapter, we define a multi-valued logic that allows our agents to have consistently, among other attitudes: beliefs, ignorance and desires about inconsistent information.

In Section 4.1, we will define bilattice structures, which provide formal relations between the classical and novel truth values. These truth values will be used in Section 4.2 to construct a logical language of multi-valued logics. We then turn in Section 4.3 to subsets of such languages which will be denoted as theories of multi-valued logic. With these theories, we can represent an agent's epistemic attitudes, such as her lack of beliefs and her inconsistent beliefs. To allow efficient representation of these theories, in Section 4.4 we provide theory descriptions that allow straightforward implementation. To allow theories to represent changing environments, the activity of changing theories has to be dealt with, this is done in Section 4.5. This chapter is partly based on published work by Lebbink et al. [LWM03b, LWM03c].

4.1 Bilattice Structure

A bilattice is an algebraic structure that formalises a space of generalised truth values. This structure provides the formal treatment of the relations between truth values that will be needed to construct multi-valued logics.

4.1.1 Rejecting the principle of the excluded middle

To allow logics to represent inconsistent propositions in a non-trivial manner, as described in Section 3.4, the assumptions of non-contradiction and the excluded middle have to be challenged. In principle, we only have to reject the assumption of non-contradiction if we are to allow a logic to represent paraconsistent propositions. The additional rejection of the assumption of the excluded middle is justified by the conceptual argument that agents, in order to be ignorant, should be allowed to fail to assign propositions a classical truth value. An addition to the conceptual argument is that such rejection will help us construct logics that can consistently represent inconsistent propositions. By waiving the assumption of the excluded middle, non-classic truth values can be defined and added to a logic (see Rescher [Res69]); such truth values can represent an agent's ignorance or consistently represent inconsistent propositions. Logics can be kept from becoming inconsistent if the inconsistent propositions that are to follow from the logic are represented as consistent formulae with a new truth value that represents the proposition's inconsistent truth status.

We will denote the classical truth value 'true' and 'false' by t and f respectively. To represent the inconsistent modality of a proposition, we use i; informally, i represents the information of both t and f. To represent the information associated with an undetermined or unknown information state of neither t nor f, we use u. Other truth values that represent biased information states will be identified in Section 4.1.4.

4.1.2 Lattices

Before exploring the bilattice structure, it is worth reviewing some general notions of ordered sets first. Given a set P, a subset $S \subseteq P$, and a partial-order[1] $\leq$ on P, $a \in P$ is an *upper bound* of S if $s \leq a$ for all $s \in S$. The inverse to the upper bound is the lower bound, i.e. $a \in P$ is a *lower bound* of S if $s \geq a$ for all $s \in S$. The set of all upper bounds of S is denoted $up(S)$, and the set of all lower bounds is denoted $lo(S)$. $a \in P$ is a *least upper bound* (or join or supremum) of S, denoted $\sqcup S = a$, if $a \in up(S)$ and $a \leq s$ for all $s \in up(S)$. Analogously, $a \in P$ is a *greatest lower bound* (or meet or infimum) of S, denoted $\sqcap S = a$, if $a \in lo(S)$ and $s \leq a$ for all $s \in lo(S)$ (cf. Lipschutz [Lip64]).

Definition 4.1 (lattice). *A* lattice *is a structure $\langle B, \leq \rangle$ such that B is a non-empty set of elements, relation $\leq$ is a partial order over B, and for all finite $S \subseteq B$ exists a greatest lower bound $\sqcap B$ and a least upper bound $\sqcup B$. (cf. Davey [DP02]).*

A lattice is called *complete* if, and only if, for all subsets exists a least upper bound and a greatest lower bound. A lattice is *finite* if B is finite.

4.1.3 Bilattices

The notion of a bilattice was first introduced by Ginsberg [Gin88] as a general framework for many AI applications, such as truth maintenance systems (Doyle [Doy79])

[1] A partial-order is a reflexive, transitive and antisymmetric relation.

and default inferences (Reiter [Rei80]). A bilattice is an algebraic structure that formalises a space of generalised truth values with two lattice orderings (Ginsberg [Gin88]; Fitting [Fit90]). Bilattices can be used for reasoning about the semantics of logic programs (Fitting [Fit91]), for pragmatics in linguistics (Schöter [Sch96]), or for semantics of logical systems (Arieli and Avron [AA98]).

Definition 4.2 (bilattice)**.** *Given two lattices $\langle B, \leq_b \rangle$ and $\langle D, \leq_d \rangle$, the structure $\mathcal{B} = \langle B \times D, \leq_k, \leq_t \rangle$ is a* bilattice *if the partial orders are defined as follows (with $b_1, b_2 \in B$ and $d_1, d_2 \in D$): $\langle b_1, d_1 \rangle \leq_k \langle b_2, d_2 \rangle$ if $b_1 \leq_b b_2$ and $d_1 \leq_d d_2$, and $\langle b_1, d_1 \rangle \leq_t \langle b_2, d_2 \rangle$ if $b_1 \leq_b b_2$ and $d_2 \leq_d d_1$ (cf. Fitting [Fit91]).*

Notation 4.1. *Instead of truth value $\langle b, d \rangle$, we write $b \times d$. Additionally, we abuse notation and write that θ is a truth value from bilattice $\mathcal{B}$ by writing $\theta \in \mathcal{B}$.*

4.1.4 Truth values

The bilattice $\mathcal{B}$ that we will use throughout this thesis will provide use with truth values that are taken from the Cartesian product of the values of the lattices $\langle B, \leq_b \rangle$ and $\langle D, \leq_d \rangle$. Let B and D be finite sets of rational numbers from $\mathbb{Q}$ between and including 0 and 1, that is, $\{0, 1\} \subseteq B \subseteq \mathbb{Q}$ and $D = B$. Truth value $b \times d \in B \times D$ has the following interpretation: The numbers from B quantify the presence or lack of *positive* support for some expression; in contrast to the numbers from D that quantify the presence or lack of *negative* support for some expression. 0 stands for lack of support and 1 stands for full presence of support.

The four-valued logic that has been proposed by Belnap [Bel77] corresponds with the bilattice in figure 4.1. This bilattice is the smallest complete bilattice and has four truth values. The truth value 1×0 stands for the classical truth value 'true' and is denoted t. In accord with our intuition, only positive support and no negative support exists for t. The truth value 0×1 stands for the classical truth value 'false' and is denoted f. For truth value f no positive support and full negative support exists. The non-orthodox truth value 0×0 represents a complete lack of information; it is denoted u and in accord with intuition holds that neither positive support nor negative support exist. The truth value 1×1 represents the inconsistency of the information; it is denoted i and it expresses that both full positive and full negative support exist.

All complete bilattices have by definition at least the four truth values present in the smallest complete bilattice. Other bilattices with more truth values are constructed by taking the sets of elements B and D with more than two elements. With $B = D = \{0, {}^1\!/_2, 1\}$ the bilattice from figure 4.2 with nine truth values is constructed. Truth values from bilattices with more than four truth values can be used to represent biased information or probabilities (see Ginsberg [Gin88]). The truth value ${}^1\!/_2 \times 0$ denotes that some positive support and no negative support exists; in a figurative sense, ${}^1\!/_2 \times 0$ is more true than false and thus *biased* to true. A partial inconsistent truth value ${}^1\!/_2 \times {}^1\!/_2$ denotes an unbiased information state in which some positive and some

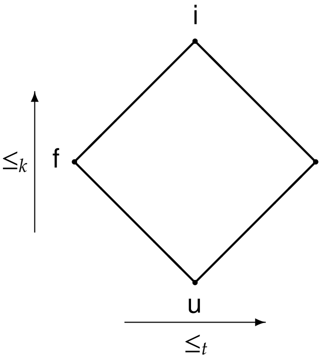

Figure 4.1: Smallest complete bilattice for a four-valued logic.

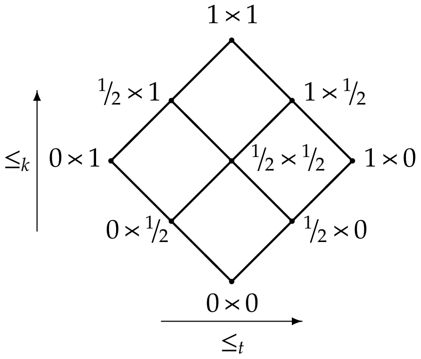

Figure 4.2: Second smallest complete bilattice having nine truth values.

negative support for some expression exists. Another partial truth value would be $1 \times 1/2$ which is biased towards true and $1/2 \times 1$ which is biased towards false.

4.1.5 Truth values orderings, and maximal and minimal elements

Truth values are ordered by the amount of truth and by the amount of information. A truth value θ_1 represents less information than truth value θ_2, denoted $\theta_1 \leq_k \theta_2$, as expressed in equation (4.1). A truth value θ_1 represents less truth than truth value θ_2, denoted $\theta_1 \leq_t \theta_2$, as expressed in equation (4.2).

$$b_1 \times d_1 \leq_k b_2 \times d_2 \quad = \quad (b_1 \leq_b b_2) \wedge (d_1 \leq_d d_2) \tag{4.1}$$
$$b_1 \times d_1 \leq_t b_2 \times d_2 \quad = \quad (b_1 \leq_b b_2) \wedge (d_2 \leq_d d_1) \tag{4.2}$$

The truth value $\mathsf{u} = 0 \times 0$ has less positive support and equal negative support than $\mathsf{t} = 1 \times 0$, and t has equal positive and less negative support than $\mathsf{i} = 1 \times 1$, that is, u has less information than t, and t has less information than i, i.e. $\mathsf{u} \leq_k \mathsf{t} \leq_k \mathsf{i}$. The truth value $\mathsf{f} = 0 \times 1$ has equal positive support and more negative support than $\mathsf{u} = 0 \times 0$, and u has less positive support and equal negative support than $\mathsf{t} = 1 \times 0$, that is, f has a lower amount of truth than u, and u has a less amount of truth than t, i.e. $\mathsf{f} \leq_t \mathsf{u} \leq_t \mathsf{t}$.

Two truth values θ_1 and θ_2 are comparable in the information order $\leq_k$, denoted $\theta_1 \gtrless_k \theta_2$, if $\theta_1 \leq_k \theta_2$ or $\theta_2 \leq_k \theta_1$. Moreover, two truth values θ_1 and θ_2 are not comparable in order $\leq_k$, denoted $\theta_1 \ngtrless_k \theta_2$, if $\theta_1 \nleq_k \theta_2$ and $\theta_2 \nleq_k \theta_1$. Comparability in order $\leq_t$ is defined analogously. For instance, the truth values t and f are not comparable in the k-order, that is, $\mathsf{t} \nleq_k \mathsf{f}$ and $\mathsf{f} \nleq_k \mathsf{t}$, i.e. $\mathsf{t} \ngtrless_k \mathsf{f}$.[2]

Because we will only be dealing with finite bilattices, we will use the maximal and minimal elements of sets of truth values instead of upper and lower bounds.

[2] In classical logics, truth value 'false' is often defined as the negation of truth value 'true', i.e. $\neg \mathsf{t} = \mathsf{f}$. For the truth values t and f from a bilattices structure the negation $\neg$ is not defined.

The maximal elements of a set of truth values $\Theta \subseteq \mathcal{B}$ in order $\leq_k$ is denoted $max_k(\Theta)$, and the set of minimal elements of Θ is denoted $min_k(\Theta)$. The maximal and minimal elements in order $\leq_t$ can be defined analogously; however, they will not be used.

$$\begin{aligned} max_k(\Theta) &= \{\theta \in \Theta \mid \forall \theta' \in \Theta \setminus \theta\ (\theta \nleq_k \theta')\} && (4.3) \\ min_k(\Theta) &= \{\theta \in \Theta \mid \forall \theta' \in \Theta \setminus \theta\ (\theta' \nleq_k \theta)\} && (4.4) \end{aligned}$$

If $max_k(\Theta)$ and $min_k(\Theta)$ are singleton sets, then we will speak of *the* k-maximum and k-minimum respectively.

To each order $\leq_k$ and $\leq_t$ join $\oplus$ and meet $\otimes$ operations are associated according to the following equations:

$$\begin{aligned} b_1 \times d_1 \otimes_k b_2 \times d_2 &= (b_1 \sqcap_b b_2) \times (d_1 \sqcap_d d_2) && (4.5) \\ b_1 \times d_1 \oplus_k b_2 \times d_2 &= (b_1 \sqcup_b b_2) \times (d_1 \sqcup_d d_2) && (4.6) \\ b_1 \times d_1 \otimes_t b_2 \times d_2 &= (b_1 \sqcap_b b_2) \times (d_1 \sqcup_d d_2) && (4.7) \\ b_1 \times d_1 \oplus_t b_2 \times d_2 &= (b_1 \sqcup_b b_2) \times (d_1 \sqcap_d d_2) && (4.8) \end{aligned}$$

The k-meet $\otimes_k$ can be thought of as the truth value representing the information that is shared by the two truth values, i.e. the mutual information of the two truth values, e.g. $\mathsf{f} \otimes_k \mathsf{t} = \mathsf{u}$, or $0 \times 1 \otimes_k 1 \times 0 = 0 \times 0$. Likewise, the k-join $\oplus_k$ is thought of as the information that results after combining two truth values, e.g. $\mathsf{f} \oplus_k \mathsf{t} = \mathsf{i}$, or $0 \times 1 \oplus_k 1 \times 0 = 1 \times 1$. The t-meet $\otimes_t$ can be thought of as the minimal positive support and the maximum negative support both truth values have, e.g. $\mathsf{u} \otimes_t \mathsf{i} = \mathsf{f}$, or $0 \times 0 \otimes_t 1 \times 1 = 0 \times 1$. The t-join $\oplus_t$ is thought of as the maximum positive support and the minimum negative that both truth values have, e.g. $\mathsf{u} \otimes_t \mathsf{i} = \mathsf{t}$, or $0 \times 0 \otimes_t 1 \times 1 = 1 \times 0$. The k-join and k-meet for sets of truth values $\Theta \subseteq \mathcal{B}$ are denoted $\bigoplus_k(\Theta)$ and $\bigotimes_k(\Theta)$ respectively. The operators for the order $\leq_t$ are denoted analogously. See Ginsberg [Gin88] and Fitting [Fit90] for formal treatments of bilattices and their operators.

Remark 4.1. *Because our bilattices are products of finite linear orders with themselves, our bilattices are distributive. A complete bilattice is (infinitely) distributive if* $\theta_1 \otimes \bigoplus(\Theta) = \bigoplus(\{\theta_1 \otimes \theta' \mid \theta' \in \Theta\})$. *In the finite case, this reduces to finite distributivity:* $\theta_1 \oplus (\theta_2 \otimes \theta_3) = (\theta_1 \oplus \theta_2) \otimes (\theta_1 \oplus \theta_3)$.

4.1.6 Truth value successor

A successor relation on truth values of a bilattice identifies pairs of adjacent truth values in either the knowledge or truth ordering. These relations are useful when procedures have to be developed to execute functions on truth values. Especially retraction, as will be defined in Section 4.5.2, will depend on the following successor relation.

The successor relation is defined with respect to the bilattice orderings $\leq_k$ and $\leq_t$; only the k-successor relation will be used and discussed. Informally, θ_1 is a k-successor of θ_2, denoted $\theta_2 \olessthan_k \theta_1$, if θ_2 is adjacent to θ_1, $\theta_2 \leq_k \theta_1$, and no other truth values are in between.

Definition 4.3 (k-successor)**.** *Given a complete bilattice $\mathcal{B}$, relation $\lessdot_k \subseteq \mathcal{B} \times \mathcal{B}$ is a* k-successor *if for $\theta_1 \lessdot_k \theta_2$ holds:*

- $\theta_1 \leq_k \theta_2$;
- $\forall \theta' \in \mathcal{B} \Big((\theta_1 \leq_k \theta') \wedge (\theta' \leq_k \theta_2) \Big) \Rightarrow \Big((\theta' = \theta_1) \vee (\theta' = \theta_2) \Big)$.

Both truth values t and f from the smallest complete bilattice (see figure 4.1) are adjacent to u, i.e. $\mathsf{u} \lessdot_k \mathsf{t}$ and $\mathsf{u} \lessdot_k \mathsf{f}$. Thus, relation $\lessdot_k$ is not a function.

4.2 A Language of Multi-valued Logic

The bilattices from the previous section will be used to construct logics in which propositions can be assigned 'classical' and non-classical truth values. This section provides the language for these logics, which are multi-valued because, as we will define in the next section, multiple truth values can be assigned to propositions simultaneously. The propositions from this language of multi-valued logic are constructed in a fashion that is considered truth value bearing, capable of being the object of an agent's belief, ignorance or desire.

Throughout this thesis, we assume that $\mathcal{F}$ is a finite set of propositional letters or formulae. A proposition of multi-valued logic is a pair $p : \theta$ with formula p from a set of formulae $\mathcal{F}$ and a truth value θ from a (finite) bilattice $\mathcal{B}$. A language of multi-valued logic consists of the finite number of propositions of multi-valued logic that can be construed in such a manner. If ambiguity is unlikely to occur, we will speak of propositions instead of propositions of multi-valued logic.

Definition 4.4 (a language of multi-valued logic $\mathcal{L}_{\mathcal{B},\mathcal{F}}$)**.** *Given a complete bilattice $\mathcal{B}$ and a set of formulae $\mathcal{F}$, a* language of multi-valued logic *$\mathcal{L}_{\mathcal{B},\mathcal{F}}$ is the smallest set that satisfies: if $p \in \mathcal{F}$ and $\theta \in \mathcal{B}$ then $p : \theta \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$.*

The reading of proposition $p : \theta$ is that *'formula p has at least truth value θ'*. The adverb 'at least' refers to the information order $\leq_k$ and states that p has been assigned possibly more information than represented by θ, but certainly *at least* the information of θ. A proposition $p : \theta_1$ is said to represent more or equal information than $p : \theta_2$ if θ_1 represents more information than θ_2, i.e. $\theta_2 \leq_k \theta_1$. Two propositions $p : \theta_1$ and $p : \theta_2$ are said to be comparable in order $\leq_k$ if θ_1 and θ_2 are comparable in order $\leq_k$; moreover, the two propositions are not comparable in order $\leq_k$ if their truth values are not comparable in order $\leq_k$.

4.3 Theories of Multi-valued Logic

With the language from the previous section, sets of propositions are constructed in which multiple truth values can be assigned to propositions simultaneously. This section provides rules to construct four different types of these sets, which we will call theories of multi-valued logic. These theories have properties that resemble an agent's belief, ignorance or desire, and are used in Chapter 5 to describe parts of the agents' cognitive state.

A theory of multi-valued logic is a set of multi-valued logic propositions $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$. The properties of theories are expressed as rules under which the sets are closed. In the remainder, and only if confusion is unlikely to occur, we speak of theories instead of theories of multi-valued logic. In Section 4.3.1 we will define normal and complete theories, in Section 4.3.2 we will define inverse and complete inverse theories.

4.3.1 Normal and complete theories

The complete lack of information associated with truth value u always applies for all propositions present in a theory. The set of formulae $\mathcal{F}$ defines the syntactic structures for which the theory can represent information. For all formulae $p \in \mathcal{F}$ a theory assigns *at least* the unique information state u.

$$p \in \mathcal{F} \quad \Longrightarrow \quad p : \mathsf{u} \in \mathcal{T} \tag{R1}$$

The reading of proposition $p : \theta$ that formula p has at least truth value θ enforces that if a proposition is part of a theory, then the propositions with less information are also part of the theory. If formula p has at least truth value θ_1 in theory $\mathcal{T}$, i.e. $p : \theta_1 \in \mathcal{T}$, and θ_2 represents less or an equal amount of information than θ_1, i.e. $\theta_2 \leq_k \theta_1$, then formula p also has at least truth value θ_2 in theory $\mathcal{T}$. The information in $p : \theta_2$ is said to be *subsumed* under $p : \theta_1$.

$$p : \theta_1 \in \mathcal{T} \quad \& \quad \theta_2 \leq_k \theta_1 \quad \Longrightarrow \quad p : \theta_2 \in \mathcal{T} \tag{R2}$$

Information is closed in a theory if the k-join of truth values of the same formula present in the theory is also present. Recall that the k-join is thought of as the information that results from combining two truth values.[3] Compare figure 4.3 in which a formula has (at least) two truth values θ_1 and θ_2, and figure 4.4 in which truth value θ_3 is present which is the k-join of θ_1 and θ_2.

$$p : \theta_1 \in \mathcal{T} \quad \& \quad p : \theta_2 \in \mathcal{T} \quad \Longrightarrow \quad p : (\theta_1 \oplus_k \theta_2) \in \mathcal{T} \tag{R3}$$

Definition 4.5 (closure of sets of propositions Cn). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, the closure of $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ under a set of rules $\mathcal{R}$, denoted $Cn(\Psi, \mathcal{R}) \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, has the following properties:*

1. *idempotent: $Cn(Cn(\Psi, \mathcal{R})) = Cn(\Psi, \mathcal{R})$;*
2. *increasing: $\Psi \subseteq Cn(\Psi, \mathcal{R})$;*
3. *monotone: $\Psi_1 \subseteq \Psi_2 \Rightarrow Cn(\Psi_1, \mathcal{R}) \subseteq Cn(\Psi_2, \mathcal{R})$.*

[3] For example, as we will discuss in Chapter 5, interpret the theory as an agent's belief state. If the agent believes that formula p is t, and, at the same time, that p is f, then the agent also believes that p is i.

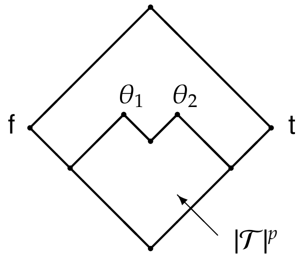

Figure 4.3: Normal theory $\mathcal{T}$ restricted to formula p.

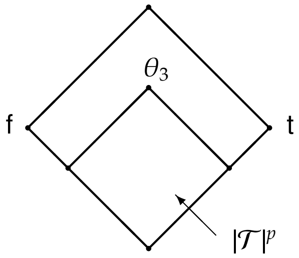

Figure 4.4: Complete theory $\mathcal{T}$ restricted to formula p.

We define theories of multi-valued logic as sets of propositions $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ that are closed under sets of rules.

Definition 4.6 (multi-valued logic theory). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, a set $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ is called a:*

- *normal theory if $\mathcal{T} = Cn(\mathcal{T}, \{R1, R2\})$;*
- *complete theory if $\mathcal{T} = Cn(\mathcal{T}, \{R1, R2, R3\})$.*

Notation 4.2. *By $|\mathcal{T}|^F \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ we denote a (unique) theory that is restricted to the formulae from $F \subseteq \mathcal{F}$, i.e. $|\mathcal{T}|^F = \{p : \theta \in \mathcal{T} \mid p \in F\}$. If a set of formulae F equals a singleton set, we substitute F with its element, e.g. instead of $|\mathcal{T}|^{\{p\}}$ we write $|\mathcal{T}|^p$.*

Figure 4.1 and 4.2 depict bilattices with four and nine truth values respectively; however, in general, complete bilattices can have any (finite) number of truth values equal to the cube of a natural number starting from two, i.e. 2^2, 3^2, 4^2, etc. Because we need not assume a specific number of truth values, we will use pictures of generic bilattices with an unspecified number of truth values. The truth values i and u have been omitted for clarity. Figure 4.3 depicts a bilattice with truth values from a theory that has been restricted to one formula. The surface between and including θ_1 and θ_2 and u is the set of truth values that hold for an example formula, say p. In a figurative sense, the surface is said to equal $|\mathcal{T}|^p$. Truth values θ_1 and θ_2 are the maxima for p. Figure 4.4 depicts a bilattice in which the surface between θ_3 and u equals a complete theory that is restricted to p. The difference with the normal theory from figure 4.3 is that the maxima of a normal theory, in this case θ_1 and θ_2, need not be a singleton set, whereas the maxima of a complete theory, due to closure under rule R3, is a singleton set.

4.3.2 Inverse theories

Inverse theories are used to represent information about propositions that are not present in theories. The opposite to the presence of propositions is the absence of propositions. To represent propositions that are absent in a theory, one either

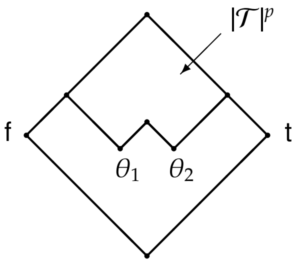

Figure 4.5: Inverse theory $\mathcal{T}$ restricted to formula p.

explicitly states which propositions are present in the theory, and leaves implicit the propositions that are absent, or one explicitly states the propositions that are absent, by making them present in an inverse theory. The latter approach is adopted in the definition of inverse theories.

The propositions that, due to closure under rule R1, are always present in normal and complete theories cannot be present in inverse theories. The least amount of information, which is represented in proposition $p : \mathsf{u}$, is not present in an inverse theory.

$$p \in \mathcal{F} \quad \Longrightarrow \quad p : \mathsf{u} \notin \mathcal{T} \tag{R1d}$$

Under the assumption that inverse theories are closed under rule R1d, a significant asymmetry surfaces with non-inverse theories: Non-inverse theories cannot be empty while inverse theories can. In a non-inverse theory, all $p \in \mathcal{F}$ have at least truth value u, which is the reason $p : \mathsf{u}$ cannot be part of an inverse theory. The asymmetry reveals a hitherto implicit assumption that because theories are constructed to represent the truth modality of formulae, if a (normal and complete) theory represents a formula, then the theory has to store the formula with a non-informative truth value.

The opposite of subsumed information from R2 states that if a proposition $p : \theta_1$ is not part of a theory, i.e. $p : \theta_1 \notin \mathcal{T}$, then also the propositions with more information, $p : \theta_2$ with $\theta_1 \leq_k \theta_2$, are not part of that theory, i.e. $p : \theta_2 \notin \mathcal{T}$. Stated differently, if a proposition is part of an inverse theory, i.e. $p : \theta_1 \in \mathcal{T}$, then all propositions with more or equal information, i.e. $p : \theta_2$ and $\theta_1 \leq_k \theta_2$, are also present in an inverse theory $p : \theta_2 \in \mathcal{T}$.

$$p : \theta_1 \in \mathcal{T} \quad \& \quad \theta_1 \leq_k \theta_2 \quad \Longrightarrow \quad p : \theta_2 \in \mathcal{T} \tag{R2d}$$

Definition 4.7 (multi-valued logic inverse theory). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, a set $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ is called an inverse theory if $\mathcal{T} = Cn(\mathcal{T}, \{R1d, R2d\})$.*

Figure 4.5 depicts the bilattice that results from restricting an example inverse theory to formula p. Intuitively, the surface between and including θ_1 and θ_2 and i is the restricted (inverse) theory.

4.4 Multi-valued Logic Theory Descriptions

The previous section provided rules that define the different theories of multi-valued logic. These rules however do not lend themselves to be programmed easily. This section provides descriptions of the information state of theories that can be programmed straightforwardly.

The k-maximal elements of a set of propositions $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, $max_k(\mathcal{T}) \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ is the set of propositions with a truth value that is the k-maximal element of all propositions of $\mathcal{T}$. Analogously, the k-minimal elements of a set of propositions $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, $min_k(\mathcal{T}) \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ is the set of propositions with a truth value that is the k-minimal element of all propositions of $\mathcal{T}$.

Definition 4.8 (maximal elements of a theory). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, the set of k-*maximal elements *of $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, denoted $max_k(\mathcal{T}) \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, is defined as follows: $max_k(\mathcal{T}) = \{p : \theta \in \mathcal{T} \mid \theta \in max_k(\{\theta' \in \mathcal{B} \mid p : \theta' \in \mathcal{T}\})\}$.*

Definition 4.9 (minimal elements of a theory). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, the set of k-*minimal elements *of $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, denoted $min_k(\mathcal{T}) \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, is defined as follows: $min_k(\mathcal{T}) = \{p : \theta \in \mathcal{T} \mid \theta \in min_k(\{\theta' \in \mathcal{B} \mid p : \theta' \in \mathcal{T}\})\}$.*

Remark 4.2. *From definition 4.8 and 4.9 follows $max_k(\mathcal{T}) \subseteq \mathcal{T}$ and $min_k(\mathcal{T}) \subseteq \mathcal{T}$.*

Proposition 4.1. *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, for every set $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ we have: (proofs in Appendix A.2)*

1. $\mathcal{T} = Cn(\mathcal{T}, \{R1, R2\}) \Rightarrow \mathcal{T} = Cn(max_k(\mathcal{T}), \{R1, R2\})$ *(normal)*
2. $\mathcal{T} = Cn(\mathcal{T}, \{R1, R2, R3\}) \Rightarrow \mathcal{T} = Cn(max_k(\mathcal{T}), \{R1, R2, R3\})$ *(complete)*
3. $\mathcal{T} = Cn(\mathcal{T}, \{R1d, R2d\}) \Rightarrow \mathcal{T} = Cn(min_k(\mathcal{T}), \{R1d, R2d\})$ *(inverse)*

The information that is represented by a theory can be described by a small subset of propositions of the theory. This set will be called the theory description; it can be used to prescribe and describe theories efficiently and uniquely. A unique description of a normal and complete theory $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ is a set $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ that equals the k-maximal elements of $\mathcal{T}$, i.e. $\Psi = max_k(\mathcal{T})$. If Ψ is closed under the rules of a normal or complete theory respectively, $\mathcal{T}$ will be yielded. A unique description of an inverse theory is provided by the k-minimal elements of the propositions that are present in the inverse theory.

Definition 4.10 (theory description $\|\mathcal{T}\|$). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, the set $\|\mathcal{T}\| \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ is called a* theory description *for theory $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$:*

1. *if $\mathcal{T}$ is a normal theory then $\|\mathcal{T}\| = max_k(\mathcal{T})$;*
2. *if $\mathcal{T}$ is a complete theory then $\|\mathcal{T}\| = max_k(\mathcal{T})$;*
3. *if $\mathcal{T}$ is inverse theory then $\|\mathcal{T}\| = min_k(\mathcal{T})$.*

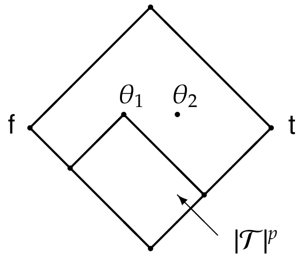

Figure 4.6: Adding $p : \theta_2$ to $\mathcal{T}$ with $||\mathcal{T}||^p = p : \theta_1$ results in figure 4.7 or 4.8.

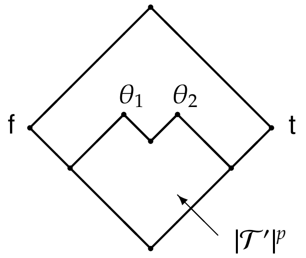

Figure 4.7: Normal theory $\mathcal{T}'$ with $||\mathcal{T}'||^p = \{p : \theta_1, p : \theta_2\}$.

Notation 4.3. *By $||\mathcal{T}||^F$ we denote the description of $\mathcal{T}$ that is restricted to the set of formulae $F \subseteq \mathcal{F}$, i.e. $||\mathcal{T}||^F = \{p : \theta \in ||\mathcal{T}|| \mid p \in F\}$. If a set of formulae F equals a singleton set, we substitute F with its element, e.g. instead of $||\mathcal{T}||^{\{p\}}$ we write $||\mathcal{T}||^p$. If a theory description $||\mathcal{T}||$ is a singleton set, we omit the set notation and state its element, e.g. instead of $||\mathcal{T}|| = \{p : \mathrm{t}\}$ we write $||\mathcal{T}|| = p : \mathrm{t}$.*

The description $||\mathcal{T}||^{\{p,q\}} = \{p : \mathrm{t}, q : \mathrm{f}\}$ states that propositions $p : \mathrm{t}$ and $q : \mathrm{f}$ describe the information state of formulae p and q in $\mathcal{T}$.

Remark 4.3. *If $\mathcal{T} = Cn(\mathcal{T}, \{R1, R2, R3\})$, then, because $\mathcal{T}$ is closed under R3 we have that $||\mathcal{T}||^p$ is a singleton set, for all $p \in \mathcal{F}$.*

4.5 Changing Theories of Multi-valued Logic

If theories and theory descriptions are to be used to represent and implement an agent's mental states, such as her beliefs, ignorance and desires, then actions to change theories have to be specified. Because agents may change their mental states, this section provides update actions of adding and retracting sets of propositions from theories. To cater for such need to change theories, procedures on theory descriptions are provided for implementing these actions efficiently.

4.5.1 Adding propositions to theories

The addition of propositions to a theory results in a possibly unchanged theory that incorporates these propositions. The action of adding $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ to $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ is to take the set-theoretical union of Ψ and $\mathcal{T}$, and to take the closure of the resulting set under the theory's rules.

Definition 4.11 (addition to theories)**.** *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, addition of $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ to theory $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, denoted add$(\Psi, \mathcal{T})$, yields theory $\mathcal{T}' \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T}' = Cn(\mathcal{T} \cup \Psi, \mathcal{R})$, with $\mathcal{R}$ the rules associated with $\mathcal{T}$.*

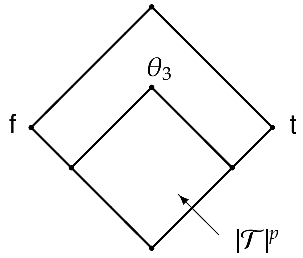

Figure 4.8: Complete theory $\mathcal{T}$ with $||\mathcal{T}||^p = p : \theta_3$.

Notation 4.4. *Instead of add($\{\psi\}, \mathcal{T}$), we write add($\psi, \mathcal{T}$) .*

Proposition 4.2. *Addition (def. 4.11) is commutative. Proof in Appendix A.2.*

Because addition is commutative, the order of adding propositions to theories does not influence the resulting theory; consequently, the addition of a set of propositions to a theory can be regarded as the addition of the set's propositions sequentially.

The addition of $p : \theta_2$ to a normal theory with theory description $||\mathcal{T}||^p = p : \theta_1$ results in one of the following three theory descriptions. See figure 4.6 and 4.7 for a depiction. If $p : \theta_2$ is subsumed under $p : \theta_1$, i.e. $\theta_2 \leq_k \theta_1$, then $p : \theta_2$ is already present in $\mathcal{T}$ and the theory description is unchanged. If $p : \theta_1$ is subsumed under $p : \theta_2$, i.e. $\theta_1 \leq_k \theta_2$, then $p : \theta_2$ is new information that is not yet present in the theory. The current information is overwritten by the new information: the new theory description is $||\mathcal{T}'||^p = p : \theta_2$. If $p : \theta_1$ and $p : \theta_2$ are not comparable in order $\leq_k$, i.e. $\theta_1 \not\lessgtr_k \theta_2$, then $||\mathcal{T}'||^p = \{p : \theta_1, p : \theta_2\}$. See figure 4.7 for a depiction, and see algorithm 4.1 for a procedure to add propositions to descriptions of normal theories.

Algorithm 4.1 (addition to a normal theory)**.** *Given $p : \theta_1 \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and (normal) theory $\mathcal{T}, \mathcal{T}' \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T}' = add(p : \theta_1, \mathcal{T})$. The procedure to compute $||\mathcal{T}'||$ from $||\mathcal{T}||$ is provided next. Take $\Psi := ||\mathcal{T}||$ and $\Phi := |\Psi|^p$. If $\exists p : \theta' \in \Phi$ $(\theta_1 \leq_k \theta')$, i.e. $p : \theta_1 \in \mathcal{T}$, then $||\mathcal{T}'|| := ||\mathcal{T}||$; else, take $\Phi_2 := \{p : \theta' \in \Phi \mid \theta' \leq_k \theta_1\}$, i.e. Φ_2 are propositions that are equally or less informed then $p : \theta_1$. Take $\Psi := \Psi \setminus \Phi_2$, and take $||\mathcal{T}'|| := \Psi \cup \{p : \theta_1\}$.*

The addition of a proposition $p : \theta_2$ to a complete theory with $||\mathcal{T}||^p = p : \theta_1$ results in one of the following three theory descriptions. See figure 4.6 and 4.8 for a depiction. If $p : \theta_2$ is subsumed under $p : \theta_1$, i.e. $\theta_2 \leq_k \theta_1$, then $p : \theta_2$ is already present in $\mathcal{T}$ and the theory description is unchanged. If $p : \theta_1$ is subsumed under $p : \theta_2$, i.e. $\theta_1 \leq_k \theta_2$, then $p : \theta_2$ is not yet present in $\mathcal{T}$, and the new theory description is $||\mathcal{T}'||^p = p : \theta_2$. If $p : \theta_1$ and $p : \theta_2$ are not comparable in order $\leq_k$, i.e. $\theta_1 \not\lessgtr_k \theta_2$, then $||\mathcal{T}'||^p = p : (\theta_1 \oplus_k \theta_2)$. See figure 4.8 for a depiction, and see algorithm 4.2 for a procedure to add propositions to theory descriptions of complete theories.

Algorithm 4.2 (addition to a complete theory)**.** *Given $p : \theta_1 \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and complete theories $\mathcal{T}, \mathcal{T}' \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T}' = add(p : \theta_1, \mathcal{T})$. The procedure to compute $||\mathcal{T}'||$ from $||\mathcal{T}||$ is provided next. Take $\Psi := ||\mathcal{T}||$ and $p : \theta_2 := |\Psi|^p$. If $\theta_1 \leq_k \theta_2$, i.e. $p : \theta_1 \in \mathcal{T}$, then $||\mathcal{T}'|| := ||\mathcal{T}||$; else, take $\Psi := \Psi \setminus \{p : \theta_2\}$ and $\theta_1 := \theta_1 \oplus_k \theta_2$. Take $||\mathcal{T}'|| := \Psi \cup \{p : \theta_1\}$.*

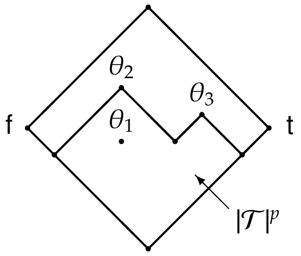

Figure 4.9: Retracting $p\!:\!\theta_1$ from normal theory $\mathcal{T}$ with $||\mathcal{T}||^p = \{p:\theta_2, p:\theta_3\}$.

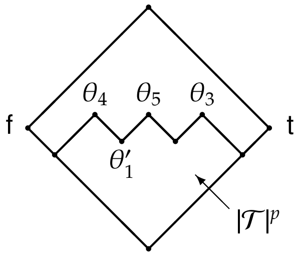

Figure 4.10: Normal theory $\mathcal{T}$ with $||\mathcal{T}||^p = \{p:\theta_4, p:\theta_5, p:\theta_2\}$.

The procedure of adding a proposition to a description of an inverse theory is similar to the procedure of adding the proposition to a description of a normal theory.

Algorithm 4.3 (addition to an inverse theory). *Given $p:\theta_1 \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and inverse theories $\mathcal{T},\mathcal{T}' \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T}' = add(p:\theta_1,\mathcal{T})$, with $\theta_1 \neq$ u. The procedure to compute $||\mathcal{T}'||$ from $||\mathcal{T}||$ is provided next. Take $\Psi := ||\mathcal{T}||$ and $\Phi := |\Psi|^p$. If $\exists p:\theta' \in \Phi$ $(\theta' \leq_k \theta_1)$, i.e. $p:\theta_1 \in \mathcal{T}$, then $||\mathcal{T}'|| := ||\mathcal{T}||$; else, take $\Phi_2 \subseteq \Phi$ with $\Phi_2 := \{p:\theta' \in \Phi \mid \theta_1 \leq_k \theta'\}$, i.e. Φ_2 are propositions that are equally or more informed then $p:\theta_1$. Take $\Psi := \Psi \setminus \Phi_2$, and take $||\mathcal{T}'|| := \Psi \cup \{p:\theta_1\}$.*

4.5.2 Retracting propositions from theories

Theories can also be changed by retracting propositions. The retraction of propositions from a normal or inverse theory possibly results in an unchanged theory that does not incorporate the propositions. However, the retraction of propositions from a complete theory need not result in a theory that does not incorporate the propositions. This ineffectiveness of retraction will be explained shortly.

Definition 4.12 (retraction from theories). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, retraction of $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ from theory $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$, denoted $retr(\Psi,\mathcal{T})$, yields theory $\mathcal{T}' \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with:*

- *normal theory: $\mathcal{T}' = \mathcal{T} \setminus \{p:\theta' \in \mathcal{L}_{\mathcal{B},\mathcal{F}} \mid (p:\theta \in \Psi) \wedge (\theta \leq_k \theta')\}$;*
- *complete theory: $\mathcal{T}' = Cn(\mathcal{T} \setminus \{p:\theta' \in \mathcal{L}_{\mathcal{B},\mathcal{F}} \mid (p:\theta \in \Psi) \wedge (\theta \leq_k \theta')\}, \{R1, R2, R3\})$;*
- *inverse theory: $\mathcal{T}' = \mathcal{T} \setminus \{p:\theta' \in \mathcal{L}_{\mathcal{B},\mathcal{F}} \mid (p:\theta \in \Psi) \wedge (\theta' \leq_k \theta)\}$.*

Notation 4.5. *Instead of $retr(\{\psi\},\mathcal{T})$, we write $retr(\psi,\mathcal{T})$.*

Remark 4.4. *Retraction of propositions from normal and inverse theories yield theories that are closed under the associated rules, i.e. $Cn(\mathcal{T},\mathcal{R}) = Cn(retr(\Psi,\mathcal{T}),\mathcal{R})$ for $\mathcal{R} = \{R1, R2\}$ or $\mathcal{R} = \{R1d, R2d\}$. Retractions of propositions from complete theories yield theories that are closed by definition.*

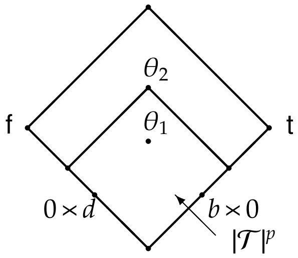

Figure 4.11: Retracting $p : \theta_1$ from complete theory $\mathcal{T}$ with $\theta_1 = 0 \times d \oplus_k b \times 0$.

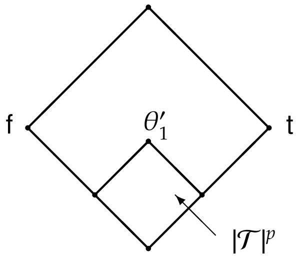

Figure 4.12: Complete theory $\mathcal{T}$ with $||\mathcal{T}||^p = p : \theta_1'$.

Taking the closure of complete theories is necessary because the resulting sets would otherwise not need to be proper theories. If a set of propositions would be the result of taking the set-theoretical difference of a theory and a set of propositions, then the resulting set need not be a proper complete theory because, if the set would be closed under the associated rules, the proposition could become an element of the set again. Stated differently, the retraction from complete theories can be ineffective if what is retracted is restored by the application of rule R3.

Proposition 4.3. *Retraction (def. 4.12) is commutative. Proof in Appendix A.2.*

Order of retracting propositions does not influence the outcome of retraction; consequently, the retraction of sets of propositions from theories can be regarded as retracting the elements of the set of propositions sequentially.

If a proposition is retracted from a normal theory, the resulting set remains closed under the rules R1 and R2. If proposition $p : \theta_1$ is retracted from a theory $\mathcal{T}$ with $||\mathcal{T}||^p = \{p : \theta_2, p : \theta_3\}$, and $p : \theta_1$ is subsumed under $p : \theta_2$, i.e. $\theta_1 \leq_k \theta_2$, and $p : \theta_1$ is not comparable to $p : \theta_3$ i.e. $\theta_1 \ngtr_k \theta_3$ (see figure 4.9), then the k-maximal element $p : \theta_2$ is replaced by two k-maximal element $p : \theta_4$ and $p : \theta_5$. Consequently, the resulting theory $\mathcal{T}'$ has description $||\mathcal{T}'||^p = \{p : \theta_4, p : \theta_5, p : \theta_3\}$ (see figure 4.10). Let θ_1' be a k-predecessor of θ_1, i.e. $\theta_1' \olessthan_k \theta_1$, then the truth value θ_4 is the t-meet of θ_2 and θ_1', i.e. $\theta_4 = \theta_2 \otimes_t \theta_1'$. truth value θ_5 is the t-join of θ_2 and θ_1', i.e. $\theta_5 = \theta_2 \oplus_k \theta_1'$. See algorithm 4.4 for a procedure to compute $||\mathcal{T}'||$ from $||\mathcal{T}||$.

Algorithm 4.4 (retraction from a normal theory)**.** *Given $p : \theta_1 \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and (normal) theories $\mathcal{T}, \mathcal{T}' \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T}' = retr(p : \theta_1, \mathcal{T})$ with $\theta_1 \neq \mathsf{u}$. The procedure to compute $||\mathcal{T}'||$ from $||\mathcal{T}||$ is provided next. Take $\Psi := ||\mathcal{T}||$ and $\Phi := |\Psi|^p$. If $\forall p : \theta' \in \Phi\ (\theta_1 \nleq_k \theta')$, i.e. $p : \theta_1 \notin \mathcal{T}$, then $||\mathcal{T}'|| := ||\mathcal{T}||$; else, take $\Phi_2 := \{p : \theta' \in \Phi \mid \theta_1 \leq_k \theta'\}$, i.e. Φ_2 are propositions that are equally or more informed than $p : \theta_1$. Take $\Psi := \Psi \setminus \Phi_2$. Take $\theta_2 := \bigoplus_k \{\theta' \in \mathcal{B} \mid p : \theta' \in \Phi_2\}$, and let θ_1' be a k-predecessor of θ_1, i.e. $\theta_1' \olessthan_k \theta_1$. Take $||\mathcal{T}'|| := \Psi \cup \{p : (\theta_2 \otimes_t \theta_1'), p : (\theta_2 \oplus_t \theta_1')\}$.*

The retraction of a set of propositions from a complete theory need not be effective. That is to say, a retracted set may remain part of the theory after its retraction, we

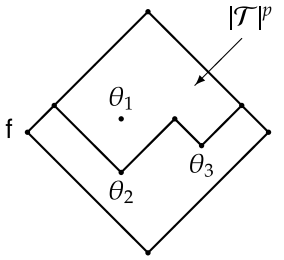

Figure 4.13: Retracting $p:\theta_1$ from inverse theory $\mathcal{T}$ with $\|\mathcal{T}\|^p = \{p{:}\theta_2, p{:}\theta_3\}$.

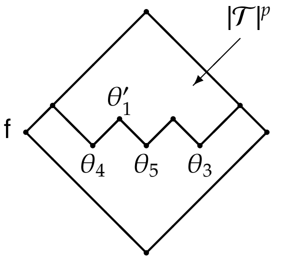

Figure 4.14: Inverse theory $\mathcal{T}$ with $\|\mathcal{T}\|^p = \{p:\theta_4, p:\theta_5, p:\theta_2\}$.

may have $\Psi \subseteq \mathit{retr}(\Psi, \mathcal{T})$. The retracted proposition may be added to the theory because a complete theory is closed under rule R3. Consider the following example.

Example 4.1. *If we have a complete theory $\mathcal{T}$ with $|\mathcal{T}|^p = p:\mathsf{i}$ and we retract $p:\mathsf{i}$ from $\mathcal{T}$, then we have $\mathcal{T}' = \mathit{retr}(p{:}\mathsf{i}, \mathcal{T})$ with $\|\mathcal{T}'\|^p = p{:}\mathsf{i}$. Because $\mathcal{T} \setminus \{p{:}\mathsf{i}\}$ yields a set of proposition Ψ with $p:\mathsf{t}, p:\mathsf{f} \in \Psi$, the closure under rule R3 will make $p:\mathsf{i}$ present in $\mathcal{T}'$ again. That is, the retraction was not effective.*

To enforce that the retraction of $p{:}\,\theta$ from a complete theory $\mathcal{T}$ is effective, instead of retracting $p:\theta$, the retraction of $p:0\times d$ or $p:b\times 0$ with $(b\times 0 \oplus_k 0\times d) = \theta_1$ will be effective. See figure 4.11 and figure 4.12. If $\mathcal{T} = Cn(\mathcal{T}, \{\text{R1}, \text{R2}, \text{R3}\})$ then for all $p:\theta \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ we have $p:\theta \notin \mathit{retr}(p:b\times 0, \mathcal{T})$ and $p:\theta \notin \mathit{retr}(p:0\times d, \mathcal{T})$. Because $b{\times}0 \leq_k \theta$ and $\forall p{:}\theta' \in \mathit{retr}(p{:}b{\times}0, \mathcal{T})$ holds $(\theta' \nleq_k \theta)$, thus $\forall p{:}\theta' \in \mathit{retr}(p{:}b{\times}0, \mathcal{T})$ holds $\neg\exists p{:}\theta'' \in \mathit{retr}(p{:}b{\times}0, \mathcal{T}) \wedge (\theta \leq_k \theta' \oplus_k \theta'')$, i.e. rule R3 cannot make $p{:}\theta \in \mathit{retr}(p{:}b{\times}0, \mathcal{T})$. A similar argument holds for $0\times d \leq_k \theta$. That is, after retracting $p:\theta$ from $\mathcal{T}$, the resulting theory $\mathcal{T}'$ will have a description $\|\mathcal{T}'\|^p = p:\theta'_1$ with θ'_1 the k-predecessor of θ_1, i.e. $\theta'_1 \olessthan_k \theta_1$. See algorithm 4.5 for a procedure to compute $\|\mathcal{T}'\|$ from $\|\mathcal{T}\|$.

Assumption 4.1. *The propositions that are to be retracted from a complete theory are chosen such that the retraction will be effective.*

Algorithm 4.5 (retraction from a complete theory)**.** *Given $p:\theta_1 \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and complete theories $\mathcal{T}, \mathcal{T}' \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T}' = \mathit{retr}(p:\theta_1, \mathcal{T})$ with $\theta_1 \neq \mathsf{u}$. The procedure to compute $\|\mathcal{T}'\|$ from $\|\mathcal{T}\|$ is provided next. Take $\Psi := \|\mathcal{T}\|$ and $p{:}\theta_2 := |\Psi|^p$. If $\theta_1 \nleq_k \theta_2$, i.e. $p{:}\theta_1 \notin \mathcal{T}$, then $\|\mathcal{T}'\| := \|\mathcal{T}\|$; else take $\Psi := \Psi \setminus \{p:\theta_2\}$ and take $\theta_1 := \theta'_1$ with $\theta'_1 \olessthan_k \theta_1$. Take $\|\mathcal{T}'\| := \Psi \cup \{p:\theta_1\}$.*

The retraction of propositions from an inverse theory is similar to retracting propositions from a normal theory. If a proposition is retracted from an inverse theory, the resulting theory remains closed under rules R1d and R2d. If proposition $p{:}\theta_1$ is retracted from a theory $\mathcal{T}$ with $\|\mathcal{T}\|^p = \{p{:}\theta_2, p{:}\theta_3\}$, and $\theta_2 \leq_k \theta_1$, and $\theta_1 \ngeq_k \theta_3$ (see figure 4.13), then the k-minima $p:\theta_2$ is replaced by two new k-minima $p:\theta_4$ and $p{:}\theta_5$. Consequently, the resulting theory $\mathcal{T}'$ has description $\|\mathcal{T}'\|^p = \{p{:}\theta_4, p{:}\theta_5, p{:}\theta_3\}$

(see figure 4.14). Let θ'_1 be a k-successor of θ_1, i.e. $\theta_1 \olessthan_k \theta'_1$, then the truth value θ_4 is the t-meet of θ_2 and θ'_1, i.e. $\theta_4 = \theta_2 \otimes_t \theta'_1$. truth value θ_5 is the t-join of θ_2 and θ'_1, i.e. $\theta_5 = \theta_2 \oplus_k \theta'_1$. See algorithm 4.6 for a procedure to compute $\|\mathcal{T}'\|$ from $\|\mathcal{T}\|$.

Algorithm 4.6 (retracting from an inverse theory)**.** *Given* $p : \theta_1 \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ *and inverse theories* $\mathcal{T}, \mathcal{T}' \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ *with* $\mathcal{T}' = retr(p : \theta_1, \mathcal{T})$*. The procedure to compute* $\|\mathcal{T}'\|$ *from* $\|\mathcal{T}\|$ *is provided next. Take* $\Psi := \|\mathcal{T}\|$ *and* $\Phi := |\Psi|^p$*. If* $\exists p : \theta' \in \Phi \ (\theta' \leq_k \theta_1)$*, i.e.* $p : \theta_1 \notin \mathcal{T}$*, then* $\|\mathcal{T}'\| := \|\mathcal{T}\|$*; else, take* $\Phi_2 := \{p : \theta' \in \Phi \mid \theta_1 \leq_k \theta'\}$*, i.e.* Φ_2 *are propositions that are equally or more informed than* $p : \theta_1$*. Take* $\Psi := \Psi \setminus \Phi_2$*. Take* $\theta_2 := \bigotimes_k \{\theta' \in \mathcal{B} \mid p : \theta' \in \Phi_2\}$*, and let* θ_1 *be a k-predecessor of* θ'_1*, i.e.* $\theta_1 \olessthan_k \theta'_1$*. Take* $\|\mathcal{T}'\| := \Psi \cup \{p : (\theta_2 \otimes_t \theta'_1), p : (\theta_2 \oplus_t \theta'_1)\}$.

4.6 Concluding Remarks

In this chapter, we defined three types of theories: normal, complete and inverse theories of multi-valued logic. Depending on the type, theories are sets of propositions that are closed under sets of rules. These theories provide a formal account of propositions that can be true, false, neither true nor false, or both true and false, as well as truth values that are biased towards either true or false. Propositions of these theories consist of a formula taken from a set of formulae $\mathcal{F}$ and a truth value taken from a bilattice $\mathcal{B}$. In this thesis, we will not commit to a specific set of formulae or to a specific bilattice. To facilitate the implementation of these theories, we defined theory descriptions that represent theories of multi-valued logics. For both theories and their theory descriptions, we defined addition and retraction operations.

Chapter 5

Cognitive Agents

The world is its own best model.
Rodney Brooks [Bro86]

We will use the term cognitive agent to denote the kind of entity for which issues of representation are central for achieving her objectives. For these agents, their representation of an environment is essential to guide their behaviours. Having a representation contrasts with subsumption architectures, which are characterised by Brooks with the slogan "the world is its own best model". Subsumption architectures work on the assumption that an explicit representation of an external reality is not needed, and that an agent's behaviour can be guided from the synergy between sensation and actuation. Whether our agents' beliefs represent: facts in reality, beliefs held by other agents, or beliefs held by human experts, does not matter for the truth of beliefs. This because our agents can become explicitly justified to believe, as described in Chapter 2, based solely on their cognitive states. These cognitive agents have no need for an external reality to justify their beliefs. If beliefs were justified by a segment of reality, then Brooks' approach would be appropriate: reality would always have more details than its representation. Because our agents justify their beliefs on their own cognitive states, arguably, agents only have interest in their own cognitive states. A cognitive state is a cognitive agent's only world.

In this chapter, we use the theories of multi-valued logic from Chapter 4 to represent cognitive states that our agents use to store their beliefs, desires and their beliefs about other agents' beliefs and desires. A cognitive state will also incorporate what an agent calls knowledge and what she calls the meaning of propositions. In Section 5.1, we will define an agent's cognitive state as a collection of mental states, a knowledge base and an epistemology. We turn in Section 5.2 to describe the relations between the different mental states. These relations describe how mental states have to be changed to retain a coherence cognitive state. In Section 5.3, we will discuss the agent's deliberation cycle that defines when changes to her mental states will be made. The deliberation cycle will provide a roadmap for the different types of rules that we will define in subsequent chapters to allow agents to make decisions and to communicate.

5.1 The Agent's Cognitive State

Our agents are said to have beliefs, desires, and beliefs about other agents' beliefs and desires. These mental states will be grouped in what we call a mental state structure, which is part of the agent's cognitive state (Section 5.1.1). To formulate conditions and properties of the agent's mental state structure, we define in Section 5.1.2 a language of mental state. We will describe in Section 5.1.3 the properties of the different mental states an agent has regarding herself and others. In addition to the mental state structure, our agents will have knowledge of some relevant domain (Section 5.1.4), and knowledge of the meaning of propositions in their community, that is, the agents are said to have an epistemology (Section 5.1.5). These three elements, viz., mental states, knowledge base and epistemology will constitute an agent's cognitive state (Section 5.1.6).

5.1.1 The mental state structure

Our agents will be said to believe propositions to be true, false, unknown, inconsistent, or some other truth value. They may also desire to believe propositions and desire to be ignorant about other propositions. In addition, agents may have beliefs regarding other agents' beliefs, beliefs regarding lack of beliefs, and beliefs about what others may or may not desire to believe. These different beliefs and desires are represented in different mental states and are stored in what will be called the agent's mental state structure. We will implement a mental state structure with several theories of multi-valued logic. As said before, we will not commit to a specific bilattice $\mathcal{B}$ nor to a specific set of formulae $\mathcal{F}$. An agent designer is free to choose whichever bilattice or set of formulae he or she wants to take. The specific properties of the agent's mental states will determine which type of theory has to be used, viz., a normal, complete or inverse theory.

Definition 5.1 (mental state structure $\mathcal{M}$). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, a* mental state structure $\mathcal{M}$ *is a tuple $\langle \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n \rangle$ with mental state $\mathcal{T}_i \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ (with $i \leq n$ and $n \leq 2$; we will shortly be dealing with $n \in \mathbb{N}$).*

Notation 5.1. *Instead of writing that $\mathcal{T}$ is an element of tuple $\mathcal{M}$, we write $\mathcal{T} \in \mathcal{M}$.*

The set $\mathcal{A}$ denotes the names of the agents that participate in the multiagent system. In general, a mental state structure $\mathcal{M}_s$ is indexed with the name of an agent $s \in \mathcal{A}$ who 'owns' the mental states. In definition 5.2, the complete set of mental state names will be defined; these mental state names correspond to the mental states that we will describe in Section 5.1.3. Later, in Section 5.2, we will provide a comprehensive treatment of the relations between the mental states.

A heuristic for the different mental state names is that Sarah has her own beliefs, denoted by B_s, her desires about her beliefs, denoted D_sB_s, and her desires about her ignorance, denoted $D_s\widetilde{B}_s$. The '$\widetilde{B}_s$' represents Sarah's ignorance, and if preceded by 'D_s', it denotes her desires about her ignorance. Sarah also has beliefs regarding other agents' beliefs. 'B_sB_j' denotes Sarah's beliefs about John's beliefs. Sarah also has beliefs about the lack of desires that other agents have, '$B_s\widetilde{D}_jB_j$ represents Sarah's beliefs about John's lack of desires to believe. The '$\widetilde{D}_j$' represents John's lack of desire. Note that the '~' is not an operator on 'B' or 'D', and that '$D_s\widetilde{B}_s$' and 'D_sB_s' are different names for a mental state structure. Next, we define the complete set of different mental state names.

Definition 5.2 (mental state names *MSN*). *Given the set of agent names $\mathcal{A}$, the finite set of* mental state names *MSN_s for agent $s \in \mathcal{A}$ with $j, a \in \mathcal{A}$, $j \neq s$, and either $a = s$ or $a = j$, has the following elements:*

- *Mental states about beliefs, beliefs about beliefs, beliefs about lack of beliefs, etc.*
  $B_s, B_sB_j, B_sB_jB_s, B_s\widetilde{B}_jB_s, B_s\widetilde{B}_j, B_sB_j\widetilde{B}_s, B_s\widetilde{B}_j\widetilde{B}_s, B_sB_jB_sB_j, B_sB_jB_s\widetilde{B}_j \in MSN_s$
- *Mental states about desires to believe, beliefs about desires to believe, beliefs about ignorance about desires to believe, beliefs about the lack of desire to believe, etc.*
  $D_sB_a, B_sD_jB_a, B_sB_jD_sB_a, B_s\widetilde{B}_jD_sB_a, B_s\widetilde{D}_jB_a, B_sB_j\widetilde{D}_sB_a, B_s\widetilde{B}_j\widetilde{D}_sB_a \in MSN_s$
- *Mental states about desires to be ignorant, beliefs about desires to be ignorant, beliefs about ignorance about desire to be ignorant, etc.*
  $D_s\widetilde{B}_a, B_sD_j\widetilde{B}_a, B_sB_jD_s\widetilde{B}_a, B_s\widetilde{B}_jD_s\widetilde{B}_a, B_s\widetilde{D}_j\widetilde{B}_a, B_sB_j\widetilde{D}_s\widetilde{B}_a, B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a \in MSN_s$

The set of mental state names of a multiagent system is denoted MSN and equals the union of all mental state names, $MSN = \bigcup_{s\in\mathcal{A}} MSN_s$.

In definition 5.1, $n \in \mathbb{N}$ is the number of different mental states that make up the agent's mental state structure. Let $d \in \mathbb{N}$ be the number of agents in the multiagent system, i.e. d is the cardinality of $\mathcal{A}$. Just as in definition 5.2, let variables $s, j, a \in \mathcal{A}$ with $j \neq s$, and either $a = s$ or $a = j$, then equation (5.1) shows that the number of mental states n that are part of the mental state structure of agent s is a linear function of d and equals $34d - 31$. For example, in expression $B_sB_jD_s\widetilde{B}_s$ (in the last line of the equation), because variable j can take $d - 1$ different agent names, the number of different possible mental states equals $d - 1$. The expression $B_sB_jD_s\widetilde{B}_a$ equals $2(d - 1)$ because a is either equal to j or equal to s. For the other three mental states an

Number of agents d	1	2	3	10	1000
Number of mental states n	3	37	71	309	33969

Table 5.1: Number of mental states in the agent's cognitive state in relation to the number of agents in the multiagent system.

analogous situation holds, thus the total number in the last line equals $8(d-1)$.

$$\begin{array}{rcll}
n & = & 34d-31 & \\
 & = & 1 & (\text{for } \utilde{B_s}) \\
 & + & 4(d-1) & (\text{for } B_sB_j, B_s\widetilde{B}_j, \utilde{B_s}B_jB_s\utilde{B_j}, B_sB_j\utilde{B_s}\utilde{B_j}) \\
 & + & 4(d-1) & (\text{for } B_sB_jB_s, B_sB_jB_s, B_sB_jB_s, B_sB_j\utilde{B_s}) \\
 & + & 2d & (\text{for } D_sB_a, \utilde{D_s}\utilde{B_a}) \\
 & + & 8(d-1) & (\text{for } B_sD_{\utilde{j}}B_a, B_sD_j\widetilde{B}_a, \utilde{B_s}\widetilde{D}_jB_a, \utilde{B_s}\utilde{D_j}B_a) \\
 & + & 8(d-1) & (\text{for } B_sB_jD_s\utilde{B_a}, B_s\utilde{B_j}D_s\utilde{B_a}, B_sB_j\utilde{D_s}\utilde{B_a}, B_s\utilde{B_j}\utilde{D_s}\utilde{B_a}) \\
 & + & 8(d-1) & (\text{for } B_sB_jD_s\widetilde{B}_a, B_s\widetilde{B}_jD_s\widetilde{B}_a, B_sB_j\widetilde{D}_s\widetilde{B}_a, B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a)
\end{array} \tag{5.1}$$

The number of mental states that are part of an agent's mental state structure is reasonably small for a small number of agents. If only one agent s is present in the multiagent system, then three mental states are needed, viz., B_s, D_sB_s and $D_s\widetilde{B}_s$. If two agents are present, 37 mental states are needed per agent, and with 10 agents, 309 mental states are needed; see table 5.1.

Assume a bijective function exists from $\{i \in \mathbb{N} \mid i \leq n\}$ to the set of mental state names MSN_s. With such a function, instead of writing that a mental state with index i is part of Sarah's mental state structure $\mathcal{T}_i \in \mathcal{M}_s$, the index i is uniquely mapped to a mental state name, which is used to denote the mental state. For example, if we would map the mental state of Sarah's desire to believe, i.e. $D_sB_s \in MSN_s$, to integer 4, then instead of writing $\mathcal{T}_4 \in \mathcal{M}_s$, we can write $D_sB_s \in \mathcal{M}_s$, and instead of writing $\psi \in \mathcal{T}_4$, we can write $\psi \in D_sB_s$.

5.1.2 The language of mental state

We will define a language, that we call the language of mental state $\mathcal{L}_{MS}$, which can be used to formulate conditions on the agent's mental state structure. We call a sentence of $\mathcal{L}_{MS}$ a mental state sentence, which we will use to express the information that is represented in an agent's mental state structure. We will define two languages of mental state: a full language $\mathcal{L}_{MS}$, and a sublanguage $\mathcal{L}'_{MS} \subseteq \mathcal{L}_{MS}$ that consists of basic literals only. A subset of the sublanguage $\mathcal{L}'_{MS}$ is sufficiently expressive to uniquely identify mental state structures. In addition to the basic literals, the full language $\mathcal{L}_{MS}$ has the operators $\wedge$ and $\neg$ that denote conjunction and classical negation respectively, additionally, the operators $\in^{\in}$ and $\in^{\notin}$ denote types of entrenchment that will be defined in Section 6.1.5.

Definition 5.3 (the languages of mental state $\mathcal{L}_{MS}$ and $\mathcal{L}'_{MS}$)**.** *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, and the set of mental state names MSN, the* languages of mental state $\mathcal{L}'_{MS} \subseteq \mathcal{L}_{MS}$ *are the smallest sets closed under:*

- *if $\psi \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and $\mathcal{T} \in MSN$ then basic literals $\ulcorner \psi \in \mathcal{T} \urcorner, \ulcorner \psi \notin \mathcal{T} \urcorner \in \mathcal{L}'_{MS}$;*
- *if $\pi_1, \pi_2 \in \mathcal{L}'_{MS}$ then $\ulcorner \pi_1 \wedge \pi_2 \urcorner, \ulcorner \neg\pi_1 \urcorner \in \mathcal{L}_{MS}$;*
- *if $\pi_1, \pi_2 \in \mathcal{L}_{MS}$ then $\ulcorner \pi_1 \wedge \pi_2 \urcorner, \ulcorner \neg\pi_1 \urcorner \in \mathcal{L}_{MS}$;*
- *if $\psi \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and $\mathcal{T} \in MSN$ then $\ulcorner \psi \in^{\in} \mathcal{T} \urcorner, \ulcorner \psi \in^{\notin} \mathcal{T} \urcorner \in \mathcal{L}_{MS}$.*

Notation 5.2. *We introduce the following abbreviations: $\neg(\psi \in \mathcal{T}) \equiv \psi \notin \mathcal{T}$, $\psi \notin^{\in} \mathcal{T} \equiv \neg(\psi {\in^{\in}} \mathcal{T})$, $\psi {\notin^{\notin}} \mathcal{T} \equiv \neg(\psi {\in^{\notin}} \mathcal{T})$, $\pi_1 \vee \pi_2 \equiv \neg(\neg\pi_1 \wedge \neg\pi_2)$. Additionally, we have abbreviation $\{\pi_1, \pi_2, \ldots, \pi_n\} \equiv \pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_n$.*

Note the difference between $B_sB_j \in MSN_s$ and $B_sB_j \in \mathcal{M}_s$. The former B_sB_j is an *identifier* for Sarah's beliefs about John's beliefs, and the latter is a *theory* of multi-valued logic that represents Sarah's beliefs about John's beliefs.

Definition 5.4 (semantics of $\mathcal{L}'_{MS}$)**.** *Given a mental state structure $\mathcal{M}$, a mental state sentence $\pi \subseteq \mathcal{L}'_{MS}$ is satisfied in $\mathcal{M}$, denoted $\mathcal{M} \models \pi$, for $\mathcal{T} \in \mathcal{M}$ if:*

1. *for basic literal $\ulcorner \psi \in \mathcal{T} \urcorner \in \mathcal{L}'_{MS}$ holds $\mathcal{M} \models \ulcorner \psi \in \mathcal{T} \urcorner$ if $\psi \in \mathcal{T}$;*
2. *for basic literal $\ulcorner \psi \notin \mathcal{T} \urcorner \in \mathcal{L}'_{MS}$ holds $\mathcal{M} \models \ulcorner \psi \notin \mathcal{T} \urcorner$ if $\psi \notin \mathcal{T}$.*

Definition 5.5 (semantics of $\mathcal{L}_{MS}$)**.** *Given a mental state structure $\mathcal{M}$, a mental state sentence $\pi \subseteq \mathcal{L}_{MS}$ is satisfied in $\mathcal{M}$, denoted $\mathcal{M} \models \pi$, for $\mathcal{T} \in \mathcal{M}$ if:*

1. *for sentence $\ulcorner \neg\pi \urcorner \in \mathcal{L}_{MS}$ holds $\mathcal{M} \models \ulcorner \neg\pi \urcorner$ if $\mathcal{M} \not\models \pi$;*
2. *for sentence $\ulcorner \pi_1 \wedge \pi_2 \urcorner \in \mathcal{L}_{MS}$ holds $\mathcal{M} \models \ulcorner \pi_1 \wedge \pi_2 \urcorner$ if $\mathcal{M} \models \pi_1$ and $\mathcal{M} \models \pi_2$;*
3. *see definition 6.7, Section 6.1.5, for the semantics of $\ulcorner \psi \in^{\in} \mathcal{T} \urcorner, \ulcorner \psi \in^{\notin} \mathcal{T} \urcorner \in \mathcal{L}_{MS}$.*

Notation 5.3. *We adopt a number of notational conventions to make expressions less cumbersome. If confusion is unlikely, we write normal brackets instead or leave out the object language distinction '$\ulcorner\urcorner$' altogether. For example, instead of $\mathcal{M}_s \models \ulcorner \psi \in B_s \urcorner$, we may write $\mathcal{M}_s \models (\psi \in B_s)$ or $\mathcal{M}_s \models \psi \in B_s$. Instead of writing that several propositions are satisfied separately, sets of propositions can be satisfied. For example, instead of $\mathcal{M}_s \models \psi \in B_s$ and $\mathcal{M}_s \models \phi \in D_sB_s$, we may write $\mathcal{M}_s \models \{\psi \in B_s, \phi \in D_sB_s\}$ or $\mathcal{M}_s \models \{(\psi \in B_s), (\phi \in D_sB_s)\}$ or $\mathcal{M}_s \models (\psi \in B_s), (\phi \in D_sB_s)$.*

To express maximal and minimal elements in $\mathcal{L}_{MS}$, two syntactical substitutions are introduced for two finite conjunctions from $\mathcal{L}_{MS}$. The expression $p : \theta \in max_k(\mathcal{T})$ represents a finite conjunction of basic literals from $\mathcal{L}'_{MS}$, and states that $p : \theta$ is a maximal element of theory $\mathcal{T}$ in order $\leq_k$. Proposition $p : \theta$ is part of $max_k(\mathcal{T})$ if and only if $p : \theta \in \mathcal{T}$, and for all propositions $p : \theta'$ with a truth value with more information, i.e. $(\theta \leq_k \theta') \wedge (\theta \neq \theta')$, holds that $p : \theta'$ is not part of the theory.

$$p : \theta \in max_k(\mathcal{T}) \equiv (p : \theta \in \mathcal{T}) \wedge \{(p : \theta' \notin \mathcal{T}) \in \mathcal{L}'_{MS} \mid (\theta \leq_k \theta') \wedge (\theta \neq \theta')\}$$

Similarly, the expression $p : \theta \in min_k(\mathcal{T})$ represents a finite conjunction of basic literals from $\mathcal{L}'_{MS}$, and states that $p : \theta$ is a minimal element of theory $\mathcal{T}$ in order $\leq_k$. Proposition $p : \theta$ is part of $min_k(\mathcal{T})$ if and only if $p : \theta \in \mathcal{T}$, and for all propositions $p : \theta'$ with a truth value with less information, i.e. $(\theta' \leq_k \theta) \wedge (\theta \neq \theta')$, holds that $p : \theta'$ is not part of the theory.

$$p : \theta \in min_k(\mathcal{T}) \equiv (p : \theta \in \mathcal{T}) \wedge \{(p : \theta' \notin \mathcal{T}) \in \mathcal{L}'_{MS} \mid (\theta' \leq_k \theta) \wedge (\theta \neq \theta')\}$$

5.1.3 Descriptions of the mental states

In Section 4.3 on theories of multi-valued logic, three types of theories have been defined as sets of propositions that are closed under different sets of rules, viz., rules R1 up to R3 and rules R1d and R2d. We now explain which theories can describe which mental state.

An agent's belief state is the mental state with propositions for which the agent is justified to predicate *to believe*. Sarah's belief state is denoted $B_s \in MSN_s$, and is represented by a complete theory of multi-valued logic because the following three properties hold.

1. Sarah cannot be ignorant about propositions that represent no information. We assume that it is not possible for Sarah to be ignorant that a proposition has a truth value equal to u. Sarah always believes that a formula p has at least truth value u, i.e. $p : \mathsf{u} \in B_s$. This property is enforced by closure under rule R1.

2. If Sarah believes that a formula p has at least truth value θ_1, and truth value θ_2 has less information than θ_1, i.e. $\theta_2 \leq_k \theta_1$, then she also believes that p has at least truth value θ_2. For example, if Sarah believes that p has at least truth value i, i.e. $p : \mathsf{i} \in B_s$, then she necessarily also believes that p has at least truth value t and f. This property is enforced by closure under rule R2.

3. If Sarah believes that a formula p has at least truth value θ_1, and also that p has at least truth value θ_2, then because she has no choice, she necessarily has to believe that p has at least both truth values at the same time. That is to say, Sarah believes that p has at least the k-join of θ_1 and θ_2, i.e. $\theta_1 \oplus_k \theta_2$. For example, if Sarah believes that p has at least truth value t and at the same time that p has at least truth value f, then she necessarily also believes that p is inconsistent, i.e. $p : \mathsf{i} \in B_s$. This property is enforced by closure under rule R3.

Figure 5.1, on page 58, depicts Sarah's belief state that has been restricted to formula p; this bilattice is a generic bilattice with n truth values equal to figure 4.3. This set of truth values that Sarah believes at a given moment is graphically depicted by the lattice below truth value θ. The complement is the set of propositions that Sarah is said to be ignorant about at that given moment. Next to being in a state of believing a proposition, agents can be said to desire to believe certain propositions. We assume that an agent's desires originate from the expert that she represents.[1]

The mental state that will represent the set of propositions that Sarah desires to believe is denoted $D_sB_s \in MSN_s$, and is represented by a normal theory of multi-valued logic because the following three properties hold.

1. Sarah desires to believe that at least no information about a formula p holds; that is, it is not possible to have a desire to believe a proposition with less information than no information. Sarah desires to believe that a formula p has at least truth value u, i.e. $p{:}\mathsf{u} \in D_sB_s$. This property is enforced by closure under rule R1.

2. If Sarah desires to believe a proposition, then she also desires to believe propositions with less information. If she desires to believe that formula p has at least truth value θ_1, then she also desires to believe that p has at least truth value θ_2 with $\theta_2 \leq_k \theta_1$. This property is enforced by closure under rule R2.

3. Contrary to her beliefs, if Sarah desires to believe that formula p has a certain truth value, say t, then she can alternatively also desire that p has another truth value, say f, yet without desiring that p should be inconsistent. If Sarah desires that p has truth values t and f, then she desires to believe these in the future, but not necessarily at the same time in the future, because she may have a choice what to believe and when. Thus, Sarah may have two contradicting desires at a given moment, yet she need not desire to have an inconsistent belief somewhere in the future. However, she does *not* have a choice what to believe: if Sarah believes that p has truth values t and f, then she has the conjunctive belief that p is inconsistent; this because she believes, at this moment, that p has both truth values. Thus, if Sarah desires to believe that a formula has several truth values, she need not take the k-join of the truth values because she has a choice what to believe in the future. Hindriks observed a similar phenomenon in the agent language GOAL [Hin01, chap. 11] in which declarative goals, i.e. goals that specify the cognitive state an agent desires to reach, should be formalised with less stringent consistency requirements. The goals of a GOAL agent are not allowed to become inconsistent; a solution would be to enhance a declarative goal (or desire) with an explicit time frame when the goal is to be achieved. Two goals with different time frames would not be considered inconsistent. In our framework, if Sarah's desires were enhanced with an explicit time frame,

[1] In general, agents are often taken to have the tacit desire to believe as many truths as possible and to avoid believing falsehoods. Evolutionary biology may provide a justification for such an assumption for humans and animals: if they are to survive, they have to keep an accurate record of their environments to guide their behaviour to avoid lethal actions and situations. For software agents, however, such survival conditions may not exist, for they are unlikely to embark on perilous quests to discover Middle-Earth. Nevertheless, the two tacit desires are sensible: if agents are to achieve co-operative and competitive plans successfully, accurate beliefs about other agents' beliefs and desires are indispensable.

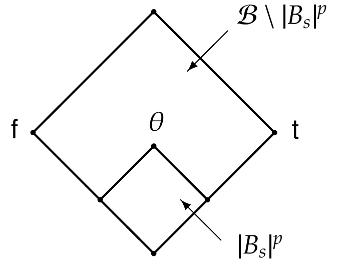

Figure 5.1: Sarah's belief state restricted to p (eq. (5.2)).

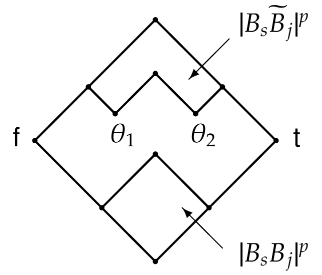

Figure 5.2: Mental states $B_s\widetilde{B}_j$ and B_sB_j are disjoint (eq. (5.3)).

then we could express that at this moment she can desire to believe that p has truth value t and f at the same future moment in time.[2] This desire would imply that she desires at that certain future moment in time to believe that p is inconsistent.

The mental state that will represent the set of propositions that Sarah desires to be ignorant about, i.e. the proposition she desires not to believe, is denoted $D_s\widetilde{B}_s \in MSN_s$, and is represented by an inverse theory of multi-valued logic because the following two properties hold.

1. Sarah should not desire to be ignorant about a proposition that represents no information. This is because she necessarily always believes that a formula has at least truth value u, and consequently, to desire to be ignorant about this, is not sensible. It is therefore never the case that Sarah desires to be ignorant about a formula p with no information, i.e. $p : \mathsf{u} \notin D_s\widetilde{B}_s$. This is enforced by closure under rule R1d.
2. If Sarah desires to be ignorant about a formula p with truth value θ_1, then she also desires to be ignorant about proposition $p : \theta_2$ that has more information, i.e. $\theta_1 \leq_k \theta_2$. This is enforced by closure under rule R2d.

For similar reasons why mental state $D_s\widetilde{B}_s$ is represented by an inverse theory, Sarah's beliefs about John's ignorance, denoted $B_s\widetilde{B}_j$, is represented by an inverse theory of multi-valued logic because the following two properties hold.

1. Because agents cannot be ignorant about propositions that represent no information, Sarah is not allowed to believe that John is ignorant that formula p has at least truth value u. i.e. $p{:}\mathsf{u} \notin B_s\widetilde{B}_j$. This is enforced by closure under rule R1d.
2. If Sarah believes that John is ignorant about a formula p with truth value θ_1, then she should also believe that John is at least ignorant that p has at least

[2] A mental state structure does not provide beliefs and desires with explicit time frames. We assume that time is explicitly represented within the formulae in $\mathcal{F}$.

normal theories	D_sB_a, $B_sD_jB_a$, $B_sB_jD_sB_a$, $B_s\widetilde{B}_jD_s\widetilde{B}_a$, $B_s\widetilde{B}_j\widetilde{B}_s$, $B_s\widetilde{D}_j\widetilde{B}_a$, $B_sB_j\widetilde{D}_s\widetilde{B}_a$, $B_s\widetilde{B}_j\widetilde{D}_sB_a$
complete theories	B_s, B_sB_j, $B_sB_jB_s$, $B_sB_jB_sB_j$
inverse theories	$B_s\widetilde{B}_j$, $B_s\widetilde{B}_jB_s$, $B_sB_j\widetilde{B}_s$, $B_sB_jB_s\widetilde{B}_j$, $D_s\widetilde{B}_a$, $B_sD_j\widetilde{B}_a$, $B_sB_jD_s\widetilde{B}_a$, $B_s\widetilde{B}_jD_sB_a$, $B_s\widetilde{D}_jB_a$, $B_sB_j\widetilde{D}_sB_a$, $B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a$

Table 5.2: The types of Sarah's mental states, with $s, j \in \mathcal{A}$ and $a = s$ or $a = j$.

truth value θ_2 that represents more information, i.e. $\theta_1 \leq_k \theta_2$. For example, if Sarah believes that John is ignorant about $p : \mathsf{t}$, then Sarah should also believe that John is ignorant about $p : \mathsf{i}$. This is enforced by closure under rule R2d.

Figure 5.2 depicts Sarah's belief about John's belief, i.e. B_sB_j, and Sarah's belief about John's ignorance, i.e. $B_s\widetilde{B}_j$. The truth values θ_1 and θ_2 are the truth values of the minimal elements of Sarah's beliefs about John's ignorance, i.e. $||B_s\widetilde{B}_j||^p = \{\theta_1, \theta_2\}$. For the set of truth values θ' with more information than θ_1 and θ_2, i.e. $\theta_1 \leq_k \theta'$ or $\theta_2 \leq_k \theta'$, Sarah believes that John is *at most* ignorant about. Sections 5.2.1 through 5.2.2 will provide comprehensive treatments of the relations between the different mental states.

As said before, an agent may also have beliefs regarding other agents' beliefs, their lack of beliefs, and what others may, or may not, desire to believe. Sarah's beliefs about John's desires what an agent $a \in \mathcal{A}$ is to believe is denoted $B_sD_jB_a \in MSN_s$. John can in his turn believe that Sarah believes this, or he can believe that Sarah is ignorant about this, which would be denoted by $B_jB_sD_jB_a \in MSN_j$ and $B_j\widetilde{B}_sD_jB_a \in MSN_j$ respectively. Sarah may also believe that John does not have a desire to believe a proposition ψ, this will be denoted by $\psi \in B_s\widetilde{D}_jB_j$. The most intricate mental state may be when Sarah believes that John is ignorant whether she desires him to be ignorant about ψ, i.e. $\psi \in B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_j$. See table 5.2 for a complete list of our agents' mental states and their corresponding type of theory.

5.1.4 The agent's knowledge base

As described in the introductory chapter, our agents represent one medical expert or one domain of relevant knowledge. The knowledge that our agents are assumed to have will be described in their knowledge bases, which we will define as sets of inference rules that allow our agents to infer beliefs from existing beliefs.

Syntactically, an inference rule consists of an antecedent and a consequent that both describe properties of an agent's mental state structure. Our inference rules will have a use similar to production rules (cf. Steffik [Ste95]); the formal definition of their usage will be provided in Section 6.3 where we define decision games to decide to believe propositions based on inference rules. Intuitively, if Sarah is entitled to know an inference rule, and the antecedent of the rule holds in her mental state structure, then she is taken to be correct to infer the consequent of the rule. A set of inference rules will be called a knowledge base.

Definition 5.6 (inference rule)**.** *Given languages of mental state* $\mathcal{L}'_{MS} \subseteq \mathcal{L}_{MS}$*, an* inference rule *is a pair* $\pi_1 \rightarrowtail \pi_2$ *with* $\pi_1 \in \mathcal{L}_{MS}$ *the* antecedent*, and* $\pi_2 \in \mathcal{L}'_{MS}$ *the* consequent*.*

The consequent of an inference rule is a basic literal from the language of mental state, while the antecedent can be any sentence from the language of mental states. We assume that inference rules are sufficiently expressive to represent our expert's domain knowledge.

Example 5.1 (inference rule)**.** *Assume formula* p_1 *reads as "the patient is suffering from disease X", and* q_1 *reads as "the patient shows symptom Y" The inference rule* $(q_1 : \mathsf{t} \in B_s) \rightarrowtail (p_1 : \mathsf{t} \in B_s)$ *has the reading that if Sarah is justified to believe that it is true that the patient shows symptom Y, then she is justified to believe that it is true that the patient suffers from disease X.*

Definition 5.7 (knowledge base $\mathcal{K}$)**.** *An agent's* knowledge base $\mathcal{K}$ *is a finite set of inference rules that the agent is entitled to predicate* to know*.*

Sarah's knowledge base $\mathcal{K}$ is a set of inference rules that is specified by the medical expert who Sarah represents. Because the expert has specified Sarah's knowledge base, Sarah is entitled to predicate to know the contents of her knowledge base. The inference rules from her knowledge base allow her to alter her mental state structure, she is however not allowed to alter her knowledge base.

5.1.5 The agent's ontology and epistemology

We will provide an agent with an explicit account of the meaning of having or lacking epistemic attitudes, such as believing and desiring, in what we call the agent's epistemology. This epistemology will define what it means, according to the agent's view of her community's agreements, to have beliefs, desires, and beliefs of other agents' beliefs and desires. We will abuse the term 'Epistemology', which is the branch of philosophy concerned with knowledge and justification, in a similar way as the term 'Ontology' is abused in Computer Science. We will first provide our interpretation of an ontology and an epistemology, we will then relate these interpretations to existing definitions.

The meaning of any sentence or statement, as described in Section 2.4.1, is that, according to our agents, they regard themselves entitled to predicate *to know* that the use of the sentence or statement is shared in their community, i.e. that the use is common knowledge. Stated differently, the agents know that a certain criterion rule is used in their community that describes the (correct) use of the sentence or statement.

- Informally, an *ontology* represents an agent's knowledge of the meaning of certain sentences. That is to say, in the terminology of the use semantics, an ontology represents the information that an agent knows that it is common knowledge that the use of the sentences is shared. Thus, according to the agent's cognitive state, the agent regards herself entitled to predicate *to know* that all agents in her community know that the use of the sentences is shared. Thus, an ontology would be the set of criterion rules that allow an agent to use sentences in accordance with the conventions in her community. If agents are allowed, based on their *private* cognitive states, to predicate that the rules of usage are shared, then the content of ontologies is agent specific. An ontology thus reflects an agent's personal view of the shared agreement on the sentences' meaning.

- Informally, an *epistemology* represents an agent's knowledge of what it means to have certain mental states, such as beliefs and desires. Given an epistemology, an agent regards herself entitled to predicate *to know* that all agents in her community know that the criterion for having (or lacking) certain mental states is shared. That is to say, an epistemology consists of the criterion rules that allow an agent to have mental states in accordance with the conventions in her community. Similar to ontologies, the contents of epistemologies are agent specific. An epistemology reflects our agent's personal view of the shared agreement on the meaning of believing and desiring.

We will implement criterion rules as inference rules, and an agent's epistemology as a finite set of inference rules. Similar to an agent's knowledge base, for which the agent is implicitly entitled to predicate that she knows the inference rules, for an agent's epistemology, the agent is implicitly entitled to predicate *to know* that all agents in her community know that the use of the inference rules is shared. Stated differently, for the inference rules part of an agent's epistemology, the agent knows that the rules are common knowledge.

Definition 5.8 (an epistemology $\mathcal{E}$). *An agent's* epistemology $\mathcal{E}$ *is a finite set of inference rules for which the agent regards herself entitled to predicate* to know *that the use of the inference rules is shared in her community.*

In general, agents can be a member in different divisions of communities within the multi-agent system. However, for our current purposes, we taken an agent's community to be all agents in the multi-agent system, i.e. the community equals $\mathcal{A}$.

We now give an example of an agent's epistemology. From MedicinNet.com, the medical online dictionary:

> "Fever: Although a fever technically is any body temperature above the normal of 98.6 degrees F. (37 degrees C.), in practice a person is usually not considered to have a significant fever until the temperature is above 100.4 degrees F (38 degrees C.)."

Sarah's epistemology may consist, among others, of the following two inference rules. Given four formulae with the following reading, q_2 reads "the patient has body temperature above 37 degrees C.", q_3 reads "the patient has a body temperature above 38 degrees C", p_2 reads "the patient has a fever", p_3 reads "the patient has a significant fever".

$$((q_2 : \mathsf{t} \in B_s) \rightarrowtail (p_2 : \mathsf{t} \in B_s)) \in \mathcal{E}_s$$
$$((q_3 : \mathsf{t} \in B_s) \rightarrowtail (p_3 : \mathsf{t} \in B_s)) \in \mathcal{E}_s$$

The second rule has the intuitive reading "if I, Sarah, am justified to believe that the patient has a body temperature above 38 degrees C, then I am justified to believe that the patient has a significant fever".

Next, we describe ontologies and their relation to our epistemologies. Ontology has a long-standing history in Philosophy, in which it provides a systematic account of existence. It is only since the 90s that the term Ontology, which is often used interchangeably with Metaphysics, came to be used for something different (Smith and Welty [SW01]). In Computer Science, according to Gruber [Gru93b, p. 907], "formal ontologies became an object for specifying content-specific agreement for the sharing and reuse of knowledge among software entities."

The view endorsed by the Semantic Web (Berners-Lee et al. [BLHL91]) is that ontologies are systematically ordered data structures that facilitate the exchange of information between computers. These ontologies define vocabularies, which are shared via web pages on the Internet. Web pages are annotated (Noy et al. [NSD$^+$01]) with specially designed declarative code from e.g. the Web Ontology Language (OWL) (Antoniou and van Harmelen [AvH03]) which functions as semantic content of a web page. The addition of meaning to web pages would allow users to search the Internet not only for syntactic, but also for semantic contents. By addressing the technical abilities of the Internet to distribute meaning, the World Wide Web is enhanced such that it can be searched for semantic content that is encoded in its ontologies.

In Computer Science, the purpose of ontologies is to facilitate effective communication between people, organizations, and software systems (Uschold [Usc96]; Uschold and Gruninger [UG96]). These ontologies intend to give rise to greater reuse, inter-operability of knowledge structures, and in general, more reliable and scalable software products. The software engineering discipline acknowledges the value of reusing knowledge structures; an ontology captures the explicit agreement that is needed to make co-operation between developers more effective.

Next, we contrast our definition of an epistemology with two definitions of an ontology that have been given in the literature.

- According to Gruber [Gru93a, p. 199], "an ontology is an explicit specification of a conceptualization". The explicitness of specifications allows designers to communicate effectively about properties of the objects they are developing, and because specifications can be communicated, in principle, the designers can explicitly agree on specifications. The software objects that are being developed need not be aware of an explicit specification. Gruber's ontologies are centred on use in software engineering: the ontology's goal is to facilitate inter-operability and reuse of knowledge structures.

- Sowa, in *Knowledge Representation* [Sow00] uses another definition:

  > "The subject of *ontology* is the study of the *categories* of things that exist or may exist in some domain. The product of such a study, called *an ontology*, is a catalog of the types of things that are assumed to exist in a domain of interest $\mathcal{D}$ from the perspective of a person who uses a language $\mathcal{L}$ for the purpose of talking about $\mathcal{D}$." [Sow00, p. 492]

  The artefacts Sowa calls ontologies have the purpose to enable persons to communicate about some domain. To allow communication, the concepts of the domain should have an agreed upon use in the language. Sowa acknowledges that the meaning of sentences can be personal and that persons can have a different perspective on the meaning.

Combining effective communication between software systems and scalable software provides a firm basis for the adoption of ontologies in multiagent systems. Agents communicate with other agents in conformance to some agreed upon language. The meaning of sentences and communicative acts that agents exchange, should be, by definition, explicitly or tacitly agreed upon by the agents (see Weiss [Wei99]). Agents could be built to automatically construct ontologies (e.g. Mena et al. [MIG00]) or agents may communicate *about* their private ontologies with the intent to align their ontologies (e.g. van Diggelen et al. [DBD^{+}06]). In the definitions by Gruber and Sowa, ontologies allow humans or software systems to communicate about words, but do not describe how the meaning of epistemic propositions can be defined. Although it is indispensable for agents to have a grasp of the meaning of propositions, with such an understanding of propositions, they still have no way of telling the epistemic status of the proposition. In addition to the meaning of propositions, agents in general need a definition when they are correct to believe and know propositions. An agent's epistemology will not only provide an agent with the meaning of propositions, but will also stipulate when propositions obtain a certain epistemic status.

Because an epistemology reflects an agent's *private* view on the agreement on the use of epistemic propositions, this private view does not enforce that every agent considers the alleged agreement shared. If two agents reveal a hitherto undiscovered difference in their use of a proposition, while both agents were perfectly entitled to predicate that both their usage was conventional, the agents become aware that their view on the agreements is not conventional within their community. This agent-specific notion of epistemology that represents the meaning of propositions, flies in the face of the mainstream thought that meaning conveying objects like ontologies and our epistemologies have to be conventional to be of any use. In Section 7.1 on speech act theory, we will describe that a strict convention is not necessary: a perceived convention would also suffice to make ontologies and epistemologies to be of use in decision making and communication.

5.1.6 Definition of the cognitive state

An agent's cognitive state is defined as the following structure.

Definition 5.9 (cognitive state *CS*)**.** *An agent's* cognitive state *CS is a tuple* $\langle \mathcal{M}, \mathcal{K}, \mathcal{E} \rangle$ *in which:* $\mathcal{M}$ *is the agent's mental states structure,* $\mathcal{K}$ *is the agent's knowledge base, and* $\mathcal{E}$ *is the agent's epistemology.*

Our agents will represent domain specific knowledge, which will be encoded in their inference rules. Predicated under the assumption that agents are not allowed to change the expert's knowledge, an agent need not be capable of changing her inference rules. We will not provide our agents with strategies to obtain knowledge themselves; agents are provided with knowledge from the domain experts. We thus assume that our agents are equipped with inference rules that they are said to know. Similarly, we will not provide strategies that allow agents to obtain common knowledge other than through the domain experts. However, in Chapter 6 on decision games, we will provide agents with rules to obtain beliefs based on other beliefs and beliefs about other agent's beliefs

5.2 Coherence of an Agent's Mental State Structure

In this section, we describe the different coherence relations that exist between the mental states that make up an agent's mental state structure.

5.2.1 Coherence of beliefs about beliefs

Some of the agent's epistemic modalities are not related with one another, such as agents' beliefs about other agents' beliefs and their own beliefs; however, other epistemic modalities are necessarily related. For example, if Sarah believes that John believes a proposition, then she is not allowed to believe also that John is ignorant about that proposition. These necessary relations between different belief and ignorance states are described by the inter-agent belief coherence principles.

The most primitive relation between epistemic states is that if an agent believes a proposition then she necessarily is not ignorant about the proposition. This allows Sarah's belief state and ignorance state to be represented with one mental state B_s. If Sarah believes proposition ψ, i.e. $\mathcal{M}_s \models \psi \in B_s$, it is then impossible for her to be ignorant about ψ at the same time, i.e. $\mathcal{M}_s \not\models \psi \notin B_s$. See also the previous figure 5.1 for a depiction. Because both belief and ignorance state are represented in one mental state, no effort has to be made to enforce this trivial relation. This trivial relation is provided because it may help to explain why other relations between epistemic modalities exist. The reverse implication also holds.

$$\mathcal{M}_s \models \psi \in B_s \quad \Leftrightarrow \quad \mathcal{M}_s \not\models \psi \notin B_s \tag{5.2}$$

If Sarah believes that John believes a proposition ψ, then she should not at the same time believe that John is ignorant about ψ. For sake of illustration, assume that if Sarah believes proposition ψ then it can be read as 'Sarah likes John', and if John believes ψ then it can be read as 'John likes Sarah'. If Sarah believes that John likes her, i.e. $\psi \in B_sB_j$, then Sarah cannot believe that John is ignorant about whether he likes her, i.e. $\psi \notin B_s\widetilde{B_j}$. In a proverbial sense, Sarah uses the relation expressed in equation (5.2) from left to right to believe something about John's ignorance state from the fact that he believes something. The other way around does not hold: if Sarah does not believe that John is ignorant about whether he likes her, then she is not allowed to believe that John likes her. See also the figure 5.2 on page 58 for a depiction.

$$\mathcal{M}_s \models \psi \in B_sB_j \quad \Rightarrow \quad \mathcal{M}_s \models \psi \notin B_s\widetilde{B_j} \tag{5.3}$$

By contraposition from equation (5.3), it follows the coherence relation that if Sarah believes that John is ignorant about ψ, i.e. $\psi \in B_s\widetilde{B_j}$, then she cannot also believe that John believes ψ, i.e. $\psi \notin B_sB_j$. Sarah has used equation (5.2) from right to left to believe something about John's beliefs from the fact that he does not believe something. The implication does not hold in the opposite direction. If Sarah does not believe that John believes he likes her, Sarah is not allowed to believe that John is ignorant whether he likes her; this because John may never have told her.

From equation (5.3), it follows that the intersection of Sarah's belief of John's belief with Sarah's belief of John's ignorance is empty, i.e. $B_sB_j \cap B_s\widetilde{B_j} = \varnothing$. The empty intersection reflects the intuition from equation (5.2) that an agent's beliefs and ignorance are complements. However, the 'inverse' that Sarah either believes that John believes ψ, or that she believes that he does not believe ψ, does not hold: it is possible that $\psi \notin B_sB_j \cup B_s\widetilde{B_j}$. Stated differently, it is possible that Sarah is not informed whether John believes ψ, and neither whether he does not believe ψ.

More intricate situations emerge when the epistemic modalities are nested one level deeper. If Sarah believes that John believes that she believes ψ, i.e. $\psi \in B_sB_jB_s$, then Sarah also has to believe that John is ignorant whether she is ignorant about ψ, i.e. $\psi \in B_s\widetilde{B_j}\widetilde{B_s}$. If Sarah believes that John believes that she believes she likes him, then Sarah also believes that John does not believe (i.e. is ignorant) that she is ignorant whether she likes him. This is because if John believes that Sarah believes she likes him, i.e. $\psi \in B_jB_s$, then, as expressed in equation (5.3), he does not believe that Sarah is ignorant whether she likes him, i.e. $\psi \notin B_j\widetilde{B_s}$. From this property, Sarah may believe that John does not believe that she is ignorant whether she likes him. The other way around does not hold: from Sarah's belief that John is ignorant whether she is ignorant whether she likes him, i.e. $\psi \in B_s\widetilde{B_j}\widetilde{B_s}$, Sarah is not allowed to believe that John believes that she likes him, while she may, but that is another story!

$$\mathcal{M}_s \models \psi \in B_sB_jB_s \quad \Rightarrow \quad \mathcal{M}_s \models \psi \in B_s\widetilde{B_j}\widetilde{B_s} \tag{5.4}$$

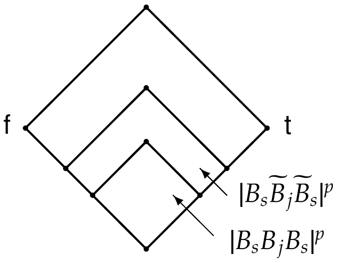

Figure 5.3: Mental state $B_sB_jB_s$ is a subset of $B_s\widetilde{B_j}\widetilde{B_s}$ (eq. (5.4)).

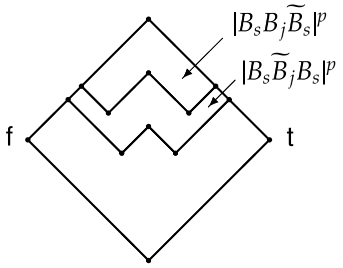

Figure 5.4: Mental state $B_sB_j\widetilde{B_s}$ is a subset of $B_s\widetilde{B_j}B_s$ (eq. (5.5)).

If Sarah believes that John believes that she is ignorant about ψ, i.e. $\psi \in B_sB_j\widetilde{B_s}$, then Sarah also believes that John is ignorant about whether she believes ψ, i.e. $\psi \in B_s\widetilde{B_j}B_s$. If Sarah believes that John believes that she is ignorant whether she likes him, then Sarah also believes that John does not believe that she believes she likes him. This is because if John believes that Sarah is ignorant whether she likes him, i.e. $\psi \in B_j\widetilde{B_s}$, then, as expressed by the contraposition of equation (5.3), he does not believe that Sarah believes she likes him, i.e. $\psi \notin B_jB_s$. From this property, Sarah may believe that John does not believe that she believes she likes him. Again, the opposite does not hold: from Sarah's belief that John does not believe that she likes him, Sarah cannot conclude that John believes that she is ignorant whether she likes him.

$$\mathcal{M}_s \models \psi \in B_sB_j\widetilde{B_s} \quad \Rightarrow \quad \mathcal{M}_s \models \psi \in B_s\widetilde{B_j}B_s \tag{5.5}$$

If Sarah believes that John does not believe (i.e. is ignorant) whether she is ignorant about ψ, i.e. $\psi \in B_s\widetilde{B_j}\widetilde{B_s}$, then Sarah does not believe that John believes her to be ignorant about ψ, i.e. $\psi \notin B_sB_j\widetilde{B_s}$. If Sarah believes that John does not believe that Sarah is ignorant whether she likes him, then Sarah does not believe that John believes that she is ignorant whether she likes him. This relation is similar to the contraposition of equation (5.3) in the sense that if Sarah believes something about the ignorance of John (regarding her), then Sarah should not believe John to believe something (about her). The opposite relation does not hold: from Sarah's lack of belief that John believes whether she likes him, Sarah cannot conclude that she believes that John is ignorant whether she believes she likes him. For example, John may believe that Sarah has not made up her mind whether she likes him or not, but he never told Sarah, i.e. $\psi \notin B_sB_j\widetilde{B_s}$, Sarah however, may believe that John is puzzled whether or not she has made up her mind about him.

$$\mathcal{M}_s \models \psi \in B_s\widetilde{B_j}\widetilde{B_s} \quad \Rightarrow \quad \mathcal{M}_s \models \psi \notin B_sB_j\widetilde{B_s} \tag{5.6}$$

By contraposition from equation (5.6), the coherence relation follows that if Sarah believes that John believes that she is ignorant about a proposition ψ, i.e. $\psi \in B_sB_j\widetilde{B_s}$, then Sarah cannot believe that John is ignorant that she is ignorant about ψ, i.e.

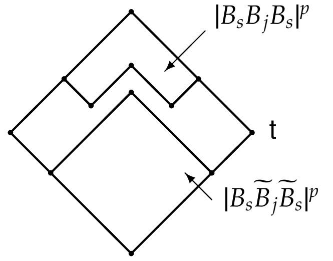
$|B_sB_j\widetilde{B}_s|^p$
f
t
$|B_s\widetilde{B}_j\widetilde{B}_s|^p$

Figure 5.5: Mental states $B_sB_j\widetilde{B}_s$ and $B_s\widetilde{B}_j\widetilde{B}_s$ are disjoint (eq. (5.5)).

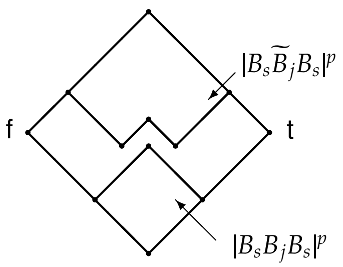

Figure 5.6: Mental states $B_s\widetilde{B}_jB_s$ and $B_sB_jB_s$ are disjoint (eq. (5.7)).

$\psi \notin B_sB_j\widetilde{B}_s$. From equation (5.6) it follows $B_sB_j\widetilde{B}_a \cap B_s\widetilde{B}_j\widetilde{B}_a = \emptyset$, see figure 5.5 for a depiction.

If Sarah believes that John is ignorant that she believes ψ, i.e. $\psi \in B_s\widetilde{B}_jB_s$, then Sarah does not believe that John believes that she believes ψ, i.e. $\psi \notin B_sB_jB_s$. If Sarah believes that John is ignorant whether she believes she likes him, she does not believe that John believes that she believes she likes him. This is because John may be ignorant whether she is ignorant whether she likes him. The implication the other way around does not hold: consider the following counterexample. It is very well possible for Sarah not to believe that John is ignorant whether she believes she likes him, and to be ignorant about whether John believes that she likes him, at the same time. Stated differently, Sarah neither believes that John is ignorant about whether she believes she likes him, and whether she does not believe that he believes that she believes she likes him, i.e. $\mathcal{M}_s \models (\psi \notin B_s\widetilde{B}_jB_s), (\psi \notin B_sB_jB_s)$.

$$\mathcal{M}_s \models \psi \in B_s\widetilde{B}_jB_s \quad \Rightarrow \quad \mathcal{M}_s \models \psi \notin B_sB_jB_s \tag{5.7}$$

By contraposition from equation (5.7), the coherence relation follows that if Sarah believes that John believes that she believes a proposition ψ, i.e. $\psi \in B_sB_jB_s$, then Sarah cannot believe that John is ignorant whether she believes ψ, i.e. $\psi \notin B_s\widetilde{B}_jB_s$. From equation (5.7), it follows that $B_sB_jB_s \cap B_s\widetilde{B}_jB_s = \emptyset$, see figure 5.6 for a depiction.

Two relations that may seem sensible are whether Sarah's beliefs about John's ignorance about whether Sarah is ignorant about ψ indicate that Sarah is ignorant about whether John is ignorant about whether Sarah believes ψ; i.e. whether $\psi \in B_s\widetilde{B}_j\widetilde{B}_s$ implies $\psi \notin B_s\widetilde{B}_jB_s$. In addition, the relation in which Sarah's beliefs about John's ignorance about whether Sarah believes ψ indicate that Sarah is ignorant about whether John is ignorant about whether Sarah is ignorant about ψ; i.e. whether $\psi \in B_s\widetilde{B}_jB_s$ implies $\psi \notin B_s\widetilde{B}_j\widetilde{B}_s$. Although they may seem sensible with the following analogy, they are not sensible in an epistemic sense. From equations (5.5) and (5.7) follows $\mathcal{M}_s \models \psi \in B_sB_j\widetilde{B}_s \Rightarrow \mathcal{M}_s \models \psi \notin B_sB_jB_s$, and from equations (5.4) and (5.6) follows $\mathcal{M}_s \models \psi \in B_sB_jB_s \Rightarrow \mathcal{M}_s \models \psi \notin B_sB_j\widetilde{B}_s$. If we were to replace 'B_j' by '$\widetilde{B}_j$' in the previous two derived coherence relations, we might be tempted to accept the two alleged relations. Consider the following counterexample. It is perfectly conceivable

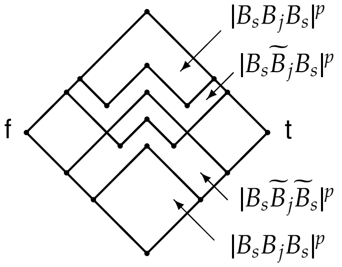
$|B_sB_j\widetilde{B}_s|^p$
$|B_s\widetilde{B}_jB_s|^p$
f
t
$|B_s\widetilde{B}_j\widetilde{B}_s|^p$
$|B_sB_jB_s|^p$

Figure 5.7: Overview of the relations from figures 5.3, 5.4, 5.5 and 5.6.

that John does not believe that Sarah believes she likes him, and that he does not believe that Sarah is ignorant whether she likes him, at the same time; as in figure 5.2: $\psi \notin B_jB_s$ and $\psi \notin B_j\widetilde{B}_s$. Sarah can come to believe the counterexample that John does not believe both beliefs, i.e. $\mathcal{M}_s \models (\psi \in B_s\widetilde{B}_jB_s), (\psi \in B_s\widetilde{B}_j\widetilde{B}_s)$, which provides a counterexample for both relations. See figure 5.7 for an overview of the relations between the four mental states $B_sB_j\widetilde{B}_s$, $B_s\widetilde{B}_jB_s$, $B_s\widetilde{B}_j\widetilde{B}_s$ and $B_sB_jB_s$.

5.2.2 Coherence of beliefs about desires

Analogous to Section 5.2.1 in which coherence relations between beliefs about beliefs and ignorance about beliefs are described, we next describe coherence relations between beliefs about desires and ignorance about desires. The modality that Sarah believes that John entertains beliefs will be broadened to the modality that Sarah believes that John entertains certain mental states. In the following discussion, if the mental state is instantiated with John's belief state, then the coherence relations, as described in Section 5.2.1, will result. In the following paragraphs, this mental state will be instantiated with desires to believe and desires to be ignorant.

Analogous to equation (5.3), if Sarah believes that John desires an agent a to believe ψ, i.e. $\psi \in B_sD_jB_a$, then she cannot believe that John does not desire a to believe ψ, i.e. $\psi \notin B_s\widetilde{D}_jB_a$. This is expressed in equation (5.8) if we substitute $\mathcal{T} := D_jB_a$ and $\widetilde{\mathcal{T}} := \widetilde{D}_jB_a$. Similar arguments hold for desires to be ignorant: we may substitute $\mathcal{T} := D_j\widetilde{B}_a$ and $\widetilde{\mathcal{T}} := \widetilde{D}_j\widetilde{B}_a$. If we substitute $\mathcal{T} := B_j$ and $\widetilde{\mathcal{T}} := \widetilde{B}_j$, then we have that equation (5.8) equals (5.3).

$$\mathcal{M}_s \models \psi \in B_s\mathcal{T} \quad \Rightarrow \quad \mathcal{M}_s \models \psi \notin B_s\widetilde{\mathcal{T}} \tag{5.8}$$

With substitutions $\mathcal{T} := B_s$ and $\widetilde{\mathcal{T}} := \widetilde{B}_s$ in equations (5.9) up to (5.12) we find the following equalities. Equation (5.9) equals (5.4), equation (5.10) equals (5.5), equation (5.11) equals (5.6), and equation (5.12) equals (5.7). Given agent $a \in \mathcal{A}$ and either $a = s$ and $a = j$, and the substitutions $\mathcal{T} := D_sB_a$ and $\widetilde{\mathcal{T}} := \widetilde{D}_sB_a$, or $\mathcal{T} := D_s\widetilde{B}_a$

and $\widetilde{\mathcal{T}} := \widetilde{D}_s\widetilde{B}_a$, then for a mental state structure we have the following equations.

$$\mathcal{M}_s \models \psi \in B_sB_j\mathcal{T} \quad \Rightarrow \quad \mathcal{M}_s \models \psi \in B_s\widetilde{B}_j\widetilde{\mathcal{T}} \tag{5.9}$$

$$\mathcal{M}_s \models \psi \in B_sB_j\widetilde{\mathcal{T}} \quad \Rightarrow \quad \mathcal{M}_s \models \psi \in B_s\widetilde{B}_j\mathcal{T} \tag{5.10}$$

$$\mathcal{M}_s \models \psi \in B_s\widetilde{B}_j\widetilde{\mathcal{T}} \quad \Rightarrow \quad \mathcal{M}_s \models \psi \notin B_sB_j\widetilde{\mathcal{T}} \tag{5.11}$$

$$\mathcal{M}_s \models \psi \in B_s\widetilde{B}_j\mathcal{T} \quad \Rightarrow \quad \mathcal{M}_s \models \psi \notin B_sB_j\mathcal{T} \tag{5.12}$$

5.2.3 Coherence of desires

Agents may desire that they themselves or others believe or are ignorant about propositions. However, under certain circumstances some combinations of desires can be assumed not to be sensible. We describe coherence principles that constrain the agent's mental state structure to avoid these combinations. The properties presented next are intra-agent desire coherence principles, that is, they hold for the cognitive state of a single agent.

In Section 5.1.3, we described that an agent's desire state is a normal theory of multi-valued logic, and, if an agent desires to believe two propositions, say, $p : \mathsf{t}$ and $p : \mathsf{f}$, then she need not desire to believe both propositions at the same time. A similar issue is that if Sarah desires to believe a proposition, she could, in principle, at the same time desire to be ignorant about the proposition. That is to say, Sarah could desire to believe a certain proposition at a given moment in the future, and desire to be ignorant about this proposition at *another* moment in the future. Sarah is considered irrational if she desires to believe a proposition at a certain moment in the future, and, *at the same time*, she desires to be ignorant about the proposition at the same moment in the future. We assume that such combinations of desires are irrational because we take it that agents cannot believe something and be ignorant about it at the same time. If Sarah's desires were enhanced with a time frame that states when her desires should hold in her mental state structure, then Sarah's desires can still be considered rational if she would desire to believe a proposition today and desire to be ignorant about it tomorrow. However, because we lack this ability to express when desires should hold, we assume that the desires to believe and the desires to be ignorant are disjoint.

Assumption 5.1 (disjoint belief and ignorance desires)**.** *If Sarah desires an agent $a \in \mathcal{A}$ to believe a proposition, then she does not desire a to be ignorant about the same proposition, i.e. $D_sB_a \cap D_s\widetilde{B}_a = \varnothing$.*

The desire to believe ψ and the desire to be ignorant about ϕ are coherent if and only if ψ and ϕ have different formulae, or ψ and ϕ are coherent if and only if their formulae are equal and their truth values are not comparable in order $\leq_k$. For example, because $\mathsf{t} \ngeqslant_k \mathsf{f}$, Sarah may desire to believe $p : \mathsf{t}$ and desire to be ignorant about $p : \mathsf{f}$. Sarah may, however, because $\mathsf{t} \geqslant_k \mathsf{i}$, i.e. $\mathsf{t} \leq_k \mathsf{i}$, not desire to believe $p : \mathsf{i}$, and desire to be ignorant about $p : \mathsf{t}$.

The 'inverse' of the assumption that Sarah desires John either to believe ψ or to be ignorant about ψ does not hold: it is possible that $\psi \notin D_sB_j \cup D_s\widetilde{B}_j$. Stated differently,

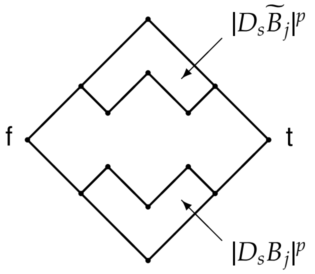

Figure 5.8: Mental states $D_s\widetilde{B}_j$ and D_sB_j are disjoint (assumption 5.1).

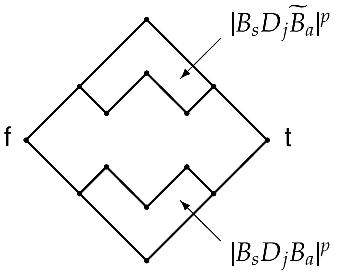

Figure 5.9: Mental states $B_sD_j\widetilde{B}_a$ and $B_sD_jB_a$ are disjoint (eq. (5.13)).

Sarah is indifferent whether John believes something. See figure 5.8 for a depiction of these disjoint desires. Notice that agents may desire themselves and others to believe inconsistent propositions, while it is not sensible for them to have conflicting desires. If agents do have conflicting desires, then they should be resolved before they may make decisions or utter communicative acts. Resolving conflicting desires is beyond the scope of this dissertation.

A desire to believe or to be ignorant about a proposition can be considered a declarative goal. A declarative goal is a description of the state of affairs sought by an agent (cf. Winikoff et al. [WPHT02]). If an agent desires to believe a specific proposition, she may undertake activity with the intent to come to believe the proposition. Procedural goals, in contrast, are activities that are performed in an attempt to achieve declarative goals. If an agent desires to believe a proposition, she may need to derive which activities given her current situation will achieve her declarative goal. Given that the situation remains unchanged, she may come to desire that the activities are performed, e.g. by other agents. Agents that can construct procedural goals from declarative goals will have to draw up a sound plan as to when to perform which activity. If we were to allow an agent both to desire to believe a proposition and to desire to be ignorant about the same proposition, then we have to presume an intricate mechanism to co-ordinate that the actions of achieving both desires are planned without becoming unachievable. A full-fletched theory of plans would allow all possible alternating sequences of actions of adding and retracting beliefs that would satisfy any complicated combination of enhanced desires. By assuming that an agent's desires are coherent, the agent need not be equipped with such advanced capabilities to plan decisions.

The desire coherence assumption, as given in assumption 5.1, does not exclude that an agent's desires regarding *other agents'* beliefs conflict.

Sarah may still desire that John believes a proposition and desire Fred to be ignorant about that same proposition. Similarly, John may desire Sarah to believe a proposition, while Fred may desire her to be ignorant about that proposition. However, Sarah may desire John (or herself) to believe a proposition $p : \theta_1$, and additionally, desire John to be ignorant about $p : \theta_2$, if and only if the truth values are

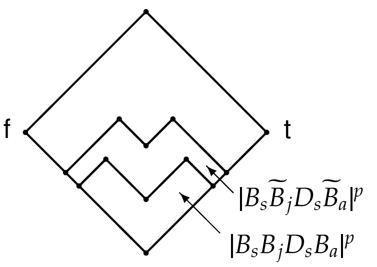

Figure 5.10: Mental state $B_sB_jD_sB_a$ is a subset of $B_s\widetilde{B}_jD_s\widetilde{B}_a$ (eq. (5.14)).

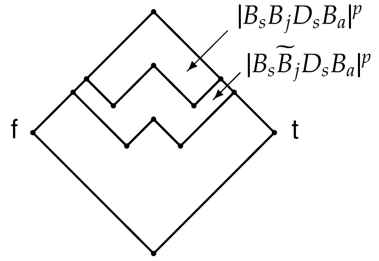
$|B_sB_jD_s\widetilde{B}_a|^p$
$|B_s\widetilde{B}_jD_sB_a|^p$
f
t

Figure 5.11: Mental state $B_sB_jD_s\widetilde{B}_a$ is a subset of $B_s\widetilde{B}_jD_sB_a$ (eq. (5.15)).

not comparable in order $\leq_k$, i.e. $\theta_1 \not\leq_k \theta_2$.

If Sarah believes that John desires an agent $a \in \mathcal{A}$ (with $a = s$ or $a = j$) to believe ψ, then, according to assumption 5.1, Sarah is not allowed to believe that John desires a to be ignorant about ψ. Analogously, if Sarah believes that John desires an agent a to be ignorant about ψ, then Sarah is not allowed to believe that John desires a to believe ψ.

$$\mathcal{M}_s \models \psi \in B_sD_jB_a \quad \Rightarrow \quad \mathcal{M}_s \models \psi \notin B_sD_j\widetilde{B}_a \tag{5.13}$$

By contraposition from equation (5.13), it follows that if Sarah believes that John desires an agent $a \in \mathcal{A}$ to be ignorant about a proposition ψ, i.e. $\psi \in B_sD_j\widetilde{B}_a$, then Sarah does not believe that John desires the agent a to believe ψ, i.e. $\psi \notin B_sD_jB_a$. In addition, it follows that the intersection of both mental states is empty, i.e. $B_sD_jB_a \cap B_sD_j\widetilde{B}_a = \emptyset$, see figure 5.9 for a depiction.

If the epistemic nesting is deepened another level, the following coherence relation can be expressed. If Sarah believes that John believes that she desires agent $a \in \mathcal{A}$ (with $a = s$ or $a = j$) to believe ψ, then Sarah believes that John is ignorant about whether she desires a to be ignorant about ψ. Sarah grounds her mental states on the assumption that John's mental state structure adheres to equation (5.13). Analogous to the previous relation, if Sarah believes that John believes that she desires a to be ignorant about ψ, then Sarah believes that John is ignorant about whether she desires a to believe ψ.

$$\mathcal{M}_s \models \psi \in B_sB_jD_sB_a \quad \Rightarrow \quad \mathcal{M}_s \models \psi \in B_s\widetilde{B}_jD_s\widetilde{B}_a \tag{5.14}$$

$$\mathcal{M}_s \models \psi \in B_sB_jD_s\widetilde{B}_a \quad \Rightarrow \quad \mathcal{M}_s \models \psi \in B_s\widetilde{B}_jD_sB_a \tag{5.15}$$

From equation (5.14) and equation (5.11), it follows that $\psi \in B_sB_jD_sB_a$ implies $\psi \notin B_sB_jD_s\widetilde{B}_a$, and from equation (5.15) and equation (5.12), it follows that $\psi \in B_sB_jD_s\widetilde{B}_a$ implies $\psi \notin B_sB_jD_sB_a$. From these two implications, it follows that $B_sB_jD_sB_a \cap B_sB_jD_s\widetilde{B}_a = \emptyset$; see figure 5.12 for an overview.

Similar to the possibility that agents have no beliefs about other agents' beliefs nor about their ignorance, as depicted in figure 5.2, or the possibility to have a non-empty intersection of $B_s\widetilde{B}_jB_s$ and $B_s\widetilde{B}_j\widetilde{B}_s$ as depicted in figure 5.7, beliefs about desires can

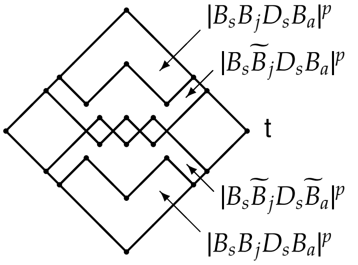
$|B_sB_jD_s\widetilde{B}_a|^p$
$|B_s\widetilde{B}_jD_sB_a|^p$
f
t
$|B_s\widetilde{B}_jD_s\widetilde{B}_a|^p$
$|B_sB_jD_sB_a|^p$

Figure 5.12: Overview of the relations from figures 5.10 and 5.11.

also have non-empty intersections. For example, if Sarah desires to believe ψ, then she may not desire to be ignorant about ψ; she may not have informed John about this, thus she could believe that John is ignorant what she desires, i.e. $\psi \in B_s\widetilde{B}_jD_sB_s$ and $\psi \in B_s\widetilde{B}_jD_s\widetilde{B}_s$. This is reflected, as depicted in figure 5.12, in the non-empty intersection of $B_s\widetilde{B}_jD_sB_a$ and $B_s\widetilde{B}_jD_s\widetilde{B}_a$.

Definition 5.10 (coherent mental state structure)**.** *An mental state structure $\mathcal{M}$ is coherent if and only if equations (5.8), (5.9), (5.10), (5.11), (5.12), assumption 5.1, equations (5.13), (5.14) and (5.15) hold.*

5.3 Agent System Dynamics

In Section 5.1 on the agent's cognitive state, we defined the static components that make up an agent's mental state structure. In Section 5.2, we described the coherence relations between different mental states that our agents have. Next, we will identify different processes that will take turns to change the agent's mental state structure. The order in which these processes are allowed to change the agents' mental state structures is defined in the agent's deliberation cycle (Section 5.3.4). However, first we will define actions that change the agent's mental states while respecting the coherence principles.

5.3.1 The basic actions of change

The action of adding propositions to theories of multi-valued logic (def. 4.11), yields theories in which the propositions are present. The addition of propositions to mental states that are part of an agent's mental state structure should respect the coherence principles. We will identify sequences of actions that change mental states while keeping the mental state structure that encompasses the mental state coherent. As defined in Section 5.2, numerous properties hold for an agent's mental state structure. These properties may be violated if a proposition is added to a mental state and subsequently, e.g., not retracted from other mental states. For example, if Sarah updates her belief that John believes ψ, i.e. she adds ψ to B_sB_j, then, as described by equation (5.3), she has to retract ψ from $B_s\widetilde{B}_j$. Methodologically, we view actions as functions on the mental states. First, we define basic actions of adding and retracting propositions from mental states.

Definition 5.11 (basic action of addition)**.** *The basic action of adding $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ to a mental state $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T} \in \mathcal{M}$, denoted $add^b_{ms}(\Psi, \mathcal{T})(\mathcal{M})$, is defined by lifting $add(\Psi, \mathcal{T})$ from definition 4.11 to a function of the mental state structure $add^b_{ms}(\Psi, \mathcal{T})(\mathcal{M})$ yielding $\mathcal{M}'$ with $\mathcal{T}' \in \mathcal{M}'$ and $\mathcal{T}' = add(\Psi, \mathcal{T})$.*

Definition 5.12 (basic action of retraction)**.** *The basic action of retracting $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ from a mental state $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T} \in \mathcal{M}$, denoted $retr^b_{ms}(\Psi, \mathcal{T})(\mathcal{M})$, is defined by lifting $retr(\Psi, \mathcal{T})$ from definition 4.12 to a function of the mental state structure $retr^b_{ms}(\Psi, \mathcal{T})(\mathcal{M})$ yielding $\mathcal{M}'$ with $\mathcal{T}' \in \mathcal{M}'$ and $\mathcal{T}' = retr(\Psi, \mathcal{T})$.*

Remark 5.1. *Because $add(\Psi, \mathcal{T})$ and $retr(\Psi, \mathcal{T})$ yield a theory of multi-valued logic, if $\mathcal{M}$ is a mental state structure, then $add^b_{ms}(\Psi, \mathcal{T})(\mathcal{M})$ and $retr^b_{ms}(\Psi, \mathcal{T})(\mathcal{M})$ yield a proper mental state structure.*

5.3.2 The complex actions of change

Functional composition of basic actions will be defined to form complex actions that will keep the mental state structure coherent. The order of the function composition of the basic actions that we define next will not matter, because the order of performing the actions will not influence the resulting mental state structure. Composition of functions $f(x)$ and $g(x)$ is denoted $(g \circ f)(x)$ and equals $(g \circ f)(x) = g(f(x))$.

Definition 5.13 (complex action of addition)**.** *The complex action of adding $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ to a mental state $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T} \in \mathcal{M}$, denoted $add_{ms}(\Psi, \mathcal{T})(\mathcal{M})$, yields a mental state structure $\mathcal{M}'$ equal to the following function compositions (with $s, j, a \in \mathcal{A}$, $s \neq j$ and either $a = s$ or $a = j$):*

1. *As described by equation (5.8):*
   - $add_{ms}(\Psi, B_sB_j)(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, B_sB_j) \circ retr_{ms}(\Psi, B_s\widetilde{B}_j)\big)(\mathcal{M}_s)$,
   - $add_{ms}(\Psi, B_s\widetilde{B}_j)(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, B_s\widetilde{B}_j) \circ retr_{ms}(\Psi, B_sB_j)\big)(\mathcal{M}_s)$.
2. *As described by equation (5.9), and the contraposition of equation (5.12):*
   - $add_{ms}(\Psi, B_sB_jB_s)(\mathcal{M}_s) =$ $\big(add^b_{ms}(\Psi, B_sB_jB_s) \circ add_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{B}_s) \circ retr_{ms}(\Psi, B_s\widetilde{B}_jB_s)\big)(\mathcal{M}_s)$.
3. *As described by equation (5.10), and the contraposition of equation (5.11):*
   - $add_{ms}(\Psi, B_sB_j\widetilde{B}_s)(\mathcal{M}_s) =$ $\big(add^b_{ms}(\Psi, B_sB_j\widetilde{B}_s) \circ add_{ms}(\Psi, B_s\widetilde{B}_jB_s) \circ retr_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{B}_s)\big)(\mathcal{M}_s)$.

4. *As described by assumption 5.1:*
   - $add_{ms}(\Psi, D_sB_a)(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, D_sB_a) \circ retr^b_{ms}(\Psi, D_s\widetilde{B}_a)\big)(\mathcal{M}_s)$,
   - $add_{ms}(\Psi, D_s\widetilde{B}_a)(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, D_s\widetilde{B}_a) \circ retr_{ms}(\Psi, D_sB_a)\big)(\mathcal{M}_s)$.
5. *As described by the contraposition of equation (5.13), and equation (5.8):*
   - $add_{ms}(\Psi, B_sD_jB_a)(\mathcal{M}_s) =$
     $\big(add^b_{ms}(\Psi, B_sD_jB_a) \circ retr_{ms}(\Psi, B_sD_j\widetilde{B}_a) \circ retr_{ms}(\Psi, B_s\widetilde{D}_jB_a)\big)(\mathcal{M}_s)$,
   - $add_{ms}(\Psi, B_sD_j\widetilde{B}_a)(\mathcal{M}_s) =$
     $\big(add^b_{ms}(\Psi, B_sD_j\widetilde{B}_a) \circ retr_{ms}(\Psi, B_sD_jB_a) \circ retr_{ms}(\Psi, B_s\widetilde{D}_j\widetilde{B}_a)\big)(\mathcal{M}_s)$.
6. *As described by the contraposition of equation (5.8):*
   - $add_{ms}(\Psi, B_s\widetilde{D}_jB_a)(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, B_s\widetilde{D}_jB_a) \circ retr_{ms}(\Psi, B_sD_jB_a)\big)(\mathcal{M}_s)$,
   - $add_{ms}(\Psi, B_s\widetilde{D}_j\widetilde{B}_a)(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, B_s\widetilde{D}_j\widetilde{B}_a) \circ retr_{ms}(\Psi, B_sD_j\widetilde{B}_a)\big)(\mathcal{M}_s)$.
7. *As described by equation (5.11), with substitutions* $\mathcal{T} := \widetilde{B}_s, \mathcal{T} := \widetilde{D}_sB_a, \mathcal{T} := \widetilde{D}_s\widetilde{B}_a$ *:*
   - $add_{ms}(\Psi, B_s\widetilde{B}_j\mathcal{T})(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, B_s\widetilde{B}_j\mathcal{T}) \circ retr_{ms}(\Psi, B_sB_j\mathcal{T})\big)(\mathcal{M}_s)$.
8. *As described by equation (5.12), with substitutions* $\mathcal{T} := B_s, \mathcal{T} := D_sB_a, \mathcal{T} := D_s\widetilde{B}_a$ *:*
   - $add_{ms}(\Psi, B_s\widetilde{B}_j\mathcal{T})(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, B_s\widetilde{B}_j\mathcal{T}) \circ retr_{ms}(\Psi, B_sB_j\mathcal{T})\big)(\mathcal{M}_s)$.
9. *As described by the contraposition of equation (5.12), and equations (5.9) and (5.14):*
   - $add_{ms}(\Psi, B_sB_jD_sB_a)(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, B_sB_jD_sB_a) \circ retr_{ms}(\Psi, B_s\widetilde{B}_jD_sB_a) \circ$
     $add_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{D}_sB_a) \circ add_{ms}(\Psi, B_s\widetilde{B}_jD_s\widetilde{B}_a)\big)(\mathcal{M}_s)$.
10. *As described by the contraposition of equation (5.12), and equations (5.9) and (5.15):*
    - $add_{ms}(\Psi, B_sB_jD_s\widetilde{B}_a)(\mathcal{M}_s) = \big(add^b_{ms}(\Psi, B_sB_jD_s\widetilde{B}_a) \circ retr_{ms}(\Psi, B_s\widetilde{B}_jD_s\widetilde{B}_a) \circ$
      $add_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a) \circ add_{ms}(\Psi, B_s\widetilde{B}_jD_sB_a)\big)(\mathcal{M}_s)$.
11. *As described by the contraposition of equation (5.11), and equation (5.10):*
    - $add_{ms}(\Psi, B_sB_j\widetilde{D}_sB_a)(\mathcal{M}_s) =$
      $\big(add^b_{ms}(\Psi, B_sB_j\widetilde{D}_sB_a) \circ retr_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{D}_sB_a) \circ add_{ms}(\Psi, B_s\widetilde{B}_jD_sB_a)\big)(\mathcal{M}_s)$,
    - $add_{ms}(\Psi, B_sB_j\widetilde{D}_s\widetilde{B}_a)(\mathcal{M}_s) =$
      $\big(add^b_{ms}(\Psi, B_sB_j\widetilde{D}_s\widetilde{B}_a) \circ retr_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a) \circ add_{ms}(\Psi, B_s\widetilde{B}_jD_s\widetilde{B}_a)\big)(\mathcal{M}_s)$.
12. *For all other theories* $\mathcal{T} \in MSN$*, such as Sarah's belief state* B_s*, we have:*

- $\mathit{add}_{ms}(\Psi, \mathcal{T})(\mathcal{M}_s) = \mathit{add}^b_{ms}(\Psi, \mathcal{T})(\mathcal{M}_s)$.

Proposition 5.1. *If $\mathcal{M}$ is a coherent mental state structure (def. 5.10), then for any $\psi \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and $\mathcal{T} \in MSN$, $\mathit{add}_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a coherent mental state structure. Proof in Appendix A.3.*

Definition 5.14 (complex action of retraction)**.** *The complex action of retracting $\Psi \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ from a mental state $\mathcal{T} \subseteq \mathcal{L}_{\mathcal{B},\mathcal{F}}$ with $\mathcal{T} \in \mathcal{M}$, denoted $\mathit{retr}_{ms}(\Psi, \mathcal{T})(\mathcal{M})$, yields a mental state structure $\mathcal{M}'$ equal to the following function:*

1. *As described by the contraposition of equation (5.9):*
   - $\mathit{retr}_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{B}_s)(\mathcal{M}) = \big(\mathit{retr}^b_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{B}_s) \circ \mathit{retr}^b_{ms}(\Psi, B_sB_jB_s)\big)(\mathcal{M})$,
   - $\mathit{retr}_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{D}_sB_a)(\mathcal{M}) = \big(\mathit{retr}^b_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{D}_sB_a) \circ \mathit{retr}^b_{ms}(\Psi, B_sB_jD_sB_a)\big)(\mathcal{M})$,
   - $\mathit{retr}_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a)(\mathcal{M}) = \big(\mathit{retr}^b_{ms}(\Psi, B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a) \circ \mathit{retr}^b_{ms}(\Psi, B_sB_jD_s\widetilde{B}_a)\big)(\mathcal{M})$.
2. *As described by the contraposition of equation (5.10):*
   - $\mathit{retr}_{ms}(\Psi, B_s\widetilde{B}_jB_s)(\mathcal{M}) = \big(\mathit{retr}^b_{ms}(\Psi, B_s\widetilde{B}_jB_s) \circ \mathit{retr}^b_{ms}(\Psi, B_sB_j\widetilde{B}_s)\big)(\mathcal{M})$.
3. *As described by the contraposition of equations (5.14) and (5.10):*
   - $\mathit{retr}_{ms}(\Psi, B_s\widetilde{B}_jD_s\widetilde{B}_a)(\mathcal{M}) =$
     $\big(\mathit{retr}^b_{ms}(\Psi, B_s\widetilde{B}_jD_s\widetilde{B}_a) \circ \mathit{retr}_{ms}(\Psi, B_sB_jD_sB_a) \circ \mathit{retr}^b_{ms}(\Psi, B_sB_j\widetilde{D}_s\widetilde{B}_a)\big)(\mathcal{M})$.
4. *As described by the contraposition of equations (5.15) and (5.10):*
   - $\mathit{retr}_{ms}(\Psi, B_s\widetilde{B}_jD_sB_a)(\mathcal{M}) =$
     $\big(\mathit{retr}^b_{ms}(\Psi, B_s\widetilde{B}_jD_sB_a) \circ \mathit{retr}^b_{ms}(\Psi, B_sB_jD_s\widetilde{B}_a) \circ \mathit{retr}^b_{ms}(\Psi, B_sB_j\widetilde{D}_sB_a)\big)(\mathcal{M})$.
5. *For all other theories $\mathcal{T} \in MSN$, such as Sarah's belief state B_s, we have:*
   - $\mathit{retr}_{ms}(\Psi, \mathcal{T})(\mathcal{M}_s) = \mathit{retr}^b_{ms}(\Psi, \mathcal{T})(\mathcal{M}_s)$.

Proposition 5.2. *If $\mathcal{M}$ is a coherent mental state structure (def. 5.10), then for any $\psi \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and $\mathcal{T} \in MSN$, $\mathit{retr}_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a coherent mental state structure. Proof in Appendix A.3.*

The complex actions that will keep the mental state structure coherent are update functions. For example, if Sarah decides to believe that John believes ψ, i.e. $\mathit{add}_{ms}(\Psi, B_sB_j)(\mathcal{M}_s)$, then Sarah updates her beliefs about his beliefs, and her beliefs about his ignorance. If Sarah later learns that John does not believe ψ, and she updates her mental state structure, i.e. $\mathit{retr}_{ms}(\Psi, B_sB_j)(\mathcal{M}_s)$, then she updates her beliefs about his beliefs only, and she does not add ψ back to his ignorance. If Sarah were to *revise* her decision to believe that John believes ψ, that is, she would 'undo' her decision and restore the original situation from before her decision. She would not only restore her beliefs about his beliefs, but also her beliefs about his ignorance. That is, our complex update actions do not perform revision of the agent's mental states, they update the agent's mental states.

5.3.3 Updating the agent's mental state structure

Definition 5.15 (update action). *The action of updating mental state structure $\mathcal{M}$ with $\pi \in \mathcal{L}'_{MS}$, denoted update$(\pi, \mathcal{M})$, yields a mental state structure $\mathcal{M}'$ with:*

- *if $\pi = \ulcorner \psi \in \mathcal{T} \urcorner$, then $\mathcal{M}' = add_{ms}(\psi, \mathcal{T})(\mathcal{M})$;*
- *if $\pi = \ulcorner \psi \notin \mathcal{T} \urcorner$, then $\mathcal{M}' = retr_{ms}(\psi, \mathcal{T})(\mathcal{M})$.*

We define a generalised update function that instead of basic literals from $\mathcal{L}'_{MS}$, updates a mental state structure with a coherent subset of $\mathcal{L}'_{MS}$.

Definition 5.16 (coherent subset of $\mathcal{L}'_{MS}$). *Set Π is a* coherent subset of $\mathcal{L}'_{MS}$ *if holds:*

- $\Pi \subseteq \mathcal{L}'_{MS}$ *;*
- *if $\ulcorner \psi \in \mathcal{T} \urcorner \in \Pi$, then $\ulcorner \psi \notin \mathcal{T} \urcorner \notin \Pi$;*
- *if $\ulcorner \psi \notin \mathcal{T} \urcorner \in \Pi$, then $\ulcorner \psi \in \mathcal{T} \urcorner \notin \Pi$.*

Definition 5.17 (update action on sets). *The action of updating mental state structure $\mathcal{M}$ with a coherent subset of $\mathcal{L}'_{MS}$ Π, denoted update$_s(\Pi, \mathcal{M})$, yields a mental state structure $\mathcal{M}'$ with:*

- *if $\Pi = \{\}$, then $\mathcal{M}' = \mathcal{M}$;*
- *if $\pi \in \Pi$, then $\mathcal{M}' = update_s(\Pi \setminus \pi, update(\pi, \mathcal{M}))$.*

Remark 5.2. *Because $add^b_{ms}(\Psi, \mathcal{T})(\mathcal{M})$ and $retr^b_{ms}(\Psi, \mathcal{T})(\mathcal{M})$ yield a mental state structure (remark 5.1), if $\mathcal{M}$ is a mental state structure, then update$_s(\Pi, \mathcal{M})$ yields a proper mental state structure.*

Proposition 5.3. *If $\mathcal{M}$ is a coherent mental state structure (def. 5.10), then update$_s(\Pi, \mathcal{M})$ yields a coherent mental state structure, for any coherent subset of $\mathcal{L}'_{MS}$ Π. Proof in Appendix A.3.*

5.3.4 The agent's deliberation cycle

In anticipation of the following two chapters, the changes on the agent's mental state structure are induced by decision-making and communication. An update of an agent's mental state structure can be either initiated by the agent's decision rules (Chapter 6), or by the agent's dialogue rules (Chapter 7). Both types of rules will have preconditions that specify the agent's mental state structure in which the rule may be applied. Analogously, both types of rules will have post-conditions that specify the properties of the agent's mental state structure directly after application of the rule.

The order in which the decision and dialogue rules are allowed to change the mental state structure is specified in the agent's deliberation cycle. Next, we define our deliberation cycle, which has eight steps, named 'a' up to 'h', including three tests, two actions, and three updates of the mental state structure. The deliberation

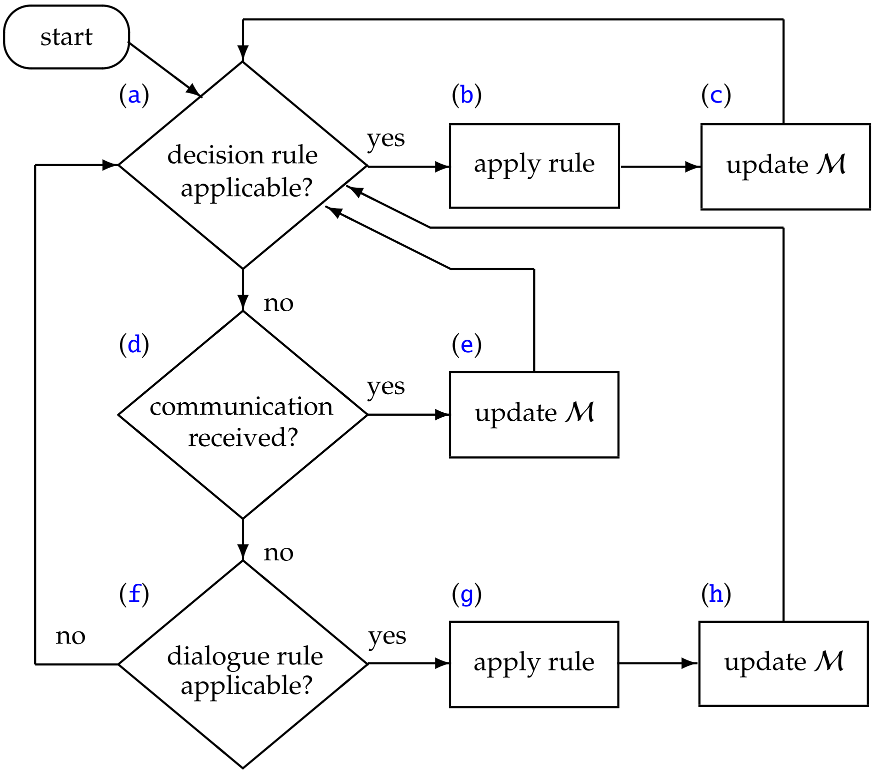

Figure 5.13: Our agent's deliberation cycle.

cycle starts at step a, has four embedded cycles, and no ending. The most embedded cycle a-b-c handles the agent's decision-making; the directly encompassing cycle a-d-e handles received communication, the first outer cycle a-d-f-g-h handles the outgoing communication, and the other outer cycle a-d-f describes that the agent is idle. See figure 5.13 for an illustration.

The deliberation cycle consists of three tests that steer the agent to enter the embedded cycles of the deliberation cycle. Once entered in an embedded cycle, the executions of dialogue and decision rules may change the agent's mental state structure.

a. Test whether decision rules are applicable, if such rules are applicable, then select a rule and go to step b to execute the rule, else go to step d to see whether communication has been received.

b. Execute the decision rule that has been selected in step a. The execution of a decision rule has no observable effects for other agents. Go to step c to update the mental state structure accordingly.

c. Update the mental state structure, that is, replace the current mental state structure by the result of the update function of the current mental state structure and the post-conditions of the decision rule from step b. Go to step a to check whether more decisions can be made.

d. Test whether communication has been received, that is, speech acts that are directed at the agent. Take the oldest act from the queue of received acts, and go to step e. If the queue is empty, go to step f to check whether the agents is allowed to utter communication.

e. Update the mental state structure, that is, replace the current mental state structure by the result of the update function of the current mental state structure and the post-conditions of the dialogue rule from step d. Go to step a to check whether decisions can be made.

f. Test whether dialogue rules are applicable, if such rules are applicable, select a rule and go to step g to execute the rule, else go to step a to check whether decisions can be made.

g. Execute the dialogue rule that has been selected in step f. The execution has the effect of uttering a speech act directed at another agent. Go to step h to update the mental state structure accordingly.

h. Update the mental state structure, that is, replace the mental state structure by the result of the update function of the current mental state structure and the post-conditions of the dialogue rule from step g. Go to step a to check whether decisions can be made.

Cycling through steps a-b-c enforces that decision-making is done before the agent engages in conversation. As we will define in Section 6.1.4 on the activity of decision-making, if the test whether the agents possesses an applicable decision rule is negative, then the agent's mental state structure is closed under decision-making. Thus, the agents will only engage in conversation if her mental state structure is closed under a set of decision rules.

Cycling through steps a-d-e enforces that all received communicative acts are processed before the agents may initiate conversation. Only after all received communication has been processed, that is, the mental state structure has been updated accordingly, then the agent may utter speech acts to other agents. After a speech act has been processed, step a allows the agent to make decisions that became applicable based on her changed mental state structure. That is to say, cycle a-b-c is embedded in cycle a-d-e with step a as the intermediating test giving control to either cycle.

The outer-cycle stepping through a-b-f-g-h handles outgoing communication. Only if all incoming communicative acts have been processed, then the agent will test whether communicative acts may be uttered. Step d is the intermediating test dividing control between the cycle for processing received communication and the cycle for outgoing communication.

If the agent neither possesses applicable decision rules, nor has unprocessed received communication, nor possesses applicable dialogue rules, then the agent is idle. The activity of an idle agent is that she cycles through steps `a-d-f` testing whether she may enter one of the three cycles. If the test in step `a` is answered negative, then Sarah's mental state structure is closed under decision-making. If the test in step `d` is answered negative, then the agent has no unprocessed received communication. If in addition to the previous two tests, the test in step `f` is answered negative, then the agent's mental state structure is closed under decision and dialogue rules and her mental state structure cannot change due to received communication, that is, the agent is idle. The steps from cycle `a-d-f` allow the agent to become active when decision rules become applicable again or a communication has been received, or when a dialogue rule has become applicable.

5.4 Concluding Remarks

In this chapter, we presented our agents' cognitive state as a structure that incorporates a mental state structure, the agent's knowledge base and her epistemology. A mental state structure represents an agent's mental states, such as her beliefs, desires, and beliefs about other agents' mental states. An agent's beliefs are represented by a complete theory of multi-valued logic, her desires to believe by a normal theory, and her ignorance and her desires to be ignorant about propositions by an inverse theory of multi-valued logic. We then defined two languages of mental state to allow us, and the agents themselves, to express properties of their mental state structure. This language allows experts to provide agents with inference rules that have preconditions that specify properties of an agent's mental state structure. In addition to the mental state structure, the agent's cognitive state consists of a knowledge base, which is a set of inference rules that, as we will define in the next chapter, the agent can use to derive new beliefs. This knowledge base will generally consist of domain specific knowledge that the agent is said to know. Analogous to the agent's knowledge base, is her epistemology, which is a set of inference rules that allows an agent not only to derive new beliefs but also to add new beliefs that other agents may derive. The epistemology represents the shared use of the propositions.

We defined several coherence principles on the agent's mental state structure. These principles resulted in sequences of actions that change the agent's mental state structure while keeping it coherent. The agent's deliberation cycle specifies when decision and dialogue rules, which we will present in the next two chapters, are allowed to change the agent's mental state structure.

Chapter **6**

Decision Games to Change Beliefs

Thinking is the talking of the soul with itself.
PLATO 427–347 BC

Agents are regarded as autonomous entities capable of performing activities independently from other agents. Their activities are actions intended to change parts of their environments, in particular for cognitive agents, as described in Chapter 5, these activities are mental exercises to change their mental states. The activity of communication, which will be elaborated in Chapter 7, is regarded as the action in which an agent utters sentences with the intent to change her environment, which will be the mental state structure of the agent that receives the uttered sentence. This chapter deals with the activity of decision-making, which will be regarded as the activity in which an agent intends to change her mental state structure. As Plato observed, thinking, and—what we will discuss shortly—decision-making, can be described as uttering sentences directed at oneself. Decision-making will be taken to be the 'uttering' of sentences with the intent to effect change in the 'speaker's' mental state structure. According to Ellenbogen [Ell03], Wittgenstein in *Philosophical Investigations* [Wit01], described that the meanings of statements are determined by the criteria for their correct use. In a similar fashion, we will describe that the criteria for correct use determine the meanings of decisions. A decision game defines the meaning of decisions with rules how to make decisions correctly.

This chapter will provide decision rules that define when agents are justified to decide to believe and disbelieve propositions. We will first deal with generic decision games. This first section will describe what decision-making with a 'meaning is use' interpretation looks like, and how a generic decision game is formalised. After a short formal description of cognitive preconditions and cognitively entrenched propositions, the generic preconditions of decisions to adopt and retract a belief will be discussed in Section 6.2. We then turn in Section 6.3 to the preconditions that explicitly justify agents to adopt and retract beliefs based on inference rules and meaning. In Section 6.4, we describe when our agents are justified to decide to change their beliefs based on the testimony of beliefs of other agents. In Section 6.5, the desirability of decisions will be discussed and the preconditions that, based on desirability, justify agents to decide to believe and be ignorant about propositions. We end with concluding remarks in Section 6.6.

6.1 Making Decisions

In general speech, agents can be said to have correctly drawn decisions if the necessary conditions of what they have decided upon have been met and the decisions are grounded in experience or knowledge (Baron [Bar00]). The situations in which a decision is correctly made can be specified as the decision's principal determinant. This section will provide how such determinants can be used in rules when to make decisions to provide the general structure of rule governed decision-making. Before we discuss our decision games, we first briefly review a few classic knowledge systems.

6.1.1 Symbolic AI

Symbolic AI, or Classical AI, is the branch of Artificial Intelligence research that concerns itself with attempting to represent human expertise explicitly. This research pursues to represent expertise in a declarative form, such as fact and rules that can be used by reasoning methods to produce human-like intelligence. If this approach is to be successful then it is necessary to translate often implicit or procedural knowledge possessed by humans into an explicit form using symbols and rules for their manipulation. Knowledge systems or experts systems are examples of success of Symbolic AI that have emerged in narrow but deep knowledge domains. Next, we review some implemented and well-known classic knowledge systems.

MYCIN (Shortliffe [Sho76]) is a knowledge system designed in the early 1970s at the Stanford University in the programming language Lisp. The system was designed to diagnose infectious diseases, especially bacterial infections of the blood, and recommend therapies and antibiotics (Buchanan [GS90]). Its diagnostic search is based on a heuristic classification model. Briefly, the specific data that has been entered by a user is abstracted to data about diseases, which is matched to abstract hypothesis. These hypotheses provide requirements for abstract therapies, which can be instantiated with specific therapies and antibiotics. The search is implemented with backward chaining production rules (cf. Steffik [Ste95, p. 134]).

MOLE (Eshelman [Esh88]), and its predecessor MORE (Kahn [Kah88]), are knowledge system shells for creating knowledge systems based on a classification model. MOLE and MORE have been mainly used for building knowledge systems that perform diagnosis. The initial design of MORE was based on experience with diagnostic problems in a drilling-fluids domain, and it was used for the diagnosis of epileptic seizures, and manufacturing defects in circuit boards. Unlike MYCIN, MOLE assumes that every finding has a cause and seeks a specific causal interpretation for the finding. It attempts to differentiate among competing hypotheses, favouring single, parsimonious explanations when possible.

PROSPECTOR (Duda [DR76]) is a knowledge system that classifies mineral exploration prospects and evaluates candidate ore deposits. It uses intermediate hypotheses to classify observational data, and unlike MYCIN, it employs an elaborate certainty calculus for combining and controlling the use of evidence. It implementation uses a probability-based inference network.

XCON, also called R1 (McDermott [McD82]), is another type of knowledge systems, unlike the previous described systems for classification, it is used for configuration. It is the first knowledge system for configuration and uses forward chaining (cf. Steffik [Ste95, p. 133]) to search a component database and sets of requirements to construct a configuration.

One major insight gained from early work in problem solving was the importance of domain-specific knowledge. Knowledge systems are constructed by obtaining knowledge from a human expert and coding it into a form that a computer may apply to similar problems. This reliance on the knowledge of a human domain expert for the system's problem solving strategies is a major feature of knowledge systems. These systems derive information from existing information. Next, we describe a similar process in which agents are said to decide to have a new information state.

6.1.2 Meaning and use of decisions

According to the dictum 'meaning is use', as described in Section 2.3.1, an agent is justified to believe a proposition, that is, justified to have a mental state in which she is said to believe a proposition, if, she regards herself entitled, according to her community, to predicate *to believe* the proposition. That is to say, according to the agent, the criterion to correctly predicate *to believe* the proposition has been met in her mental state structure. The meaning of believing a proposition is the circumstances in which an agent may come to entertain and maintain the mental state of believing. We will call the activity that changes an agent's mental state such that she can be said to entertain a mental state, decision-making. The activity of deciding to believe a proposition changes an agent's cognitive state to reflect that she believes the proposition. Informally, we say that an agent decides to add a belief to her mental state structure. Under the assumption that what an agent is going to decide upon is not already present in her mental state structure, a consequence of the decision is that the belief is added to her mental states. Analogously, it can be said that abandoning existing beliefs is the activity in which the agent's mental state structure changes to reflect that she does not believe the proposition anymore. To structure the processes of decision-making, we provide decision rules that will change an agent's mental state structure to reflect her perceived explicit justification for believing or disbelieving propositions.

The decision rules that we will define in subsequent sections implement that an agent regards herself entitled, according to her community, to know a conventional rule that describes the situations in which she is explicitly justified to make a certain decision. A decision rule consists of a set of preconditions that define in which situations an agent is explicitly justified to make a decision, and consists of a set of post-conditions that describe the properties that hold after the agent has made the decision. The set of preconditions reflect the criterion of a conventional rule, and the post-conditions provide the properties of an agent's mental state structure after the agent has changed her mental state structure to reflect that the decision has been made.

Two activities are involved in having beliefs: the agent's decision to adopt new beliefs, and the decision to retract existing beliefs. These decisions should implement that an agent may change her beliefs if she is justified to do so. That is to say, if an agent is justified to believe a proposition, then she may change her mental state structure to reflect this. We implement the change of the agent's mental state structure by adding a proposition to her belief state, and retracting propositions from her belief state. In subsequent sections, two decision games are defined that describe the use of decisions in which agents change their beliefs.

Our main objective in this chapter is to provide decision rules that *prescribe* in which situations our agents are explicitly justified to adopt and retract beliefs. The decision rules are not intended to *describe* the behaviour of some group of software agents, but the rules *prescribe* when our agents are correct to adopt and retract beliefs, according to *our* perceived conventions of the medical domain that our agents represent. Similar to the axioms of expected utility theory, which define economic rational behaviour (Neumann and Morgenstern [vNM44]), the decision rules can be regarded to define when agents are correct to change their beliefs. Our agents are not assumed to behave economically rational, as voiced by expected utility theory, but their decision behaviours should be conform our perceived conventions of the medial domain. Although our agents have not themselves agreed on the conventions when to make decisions, we assume that their responsible human experts have.

6.1.3 Decision games

A decision game is a finite number of decision rules that define the correct use of decisions. We will use decision games to explain what the effects of these decisions are, and why agents are allowed to make them. Intuitively, we view decision games as single-player games in which the moves in the game correspond with an agent's decisions. In two-player games, such as the game of chess, the players make moves and take turns with the aim of reaching a state in which one has won. In single-player games, such as the card game called patience, a player only makes moves, presumably also with the aim of reaching a state in which the player has won the game. Our agent's goal in decision games will be to reach a certain mental state that she may characterise as desirable. We will return to the issue of desirable states in Section 6.5 on desiring to change beliefs.

The different decisions that our agents can make are taken from a language of decisions $\mathcal{L}_D$ which we define next.

Definition 6.1 (decision language $\mathcal{L}_D$). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, a set of agent names $\mathcal{A}$, and the set of mental state identifiers MSN, the* decision language *$\mathcal{L}_D$ consists of the following decisions. If $s \in \mathcal{A}$, $\psi \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$ and $ms \in MSN$, then:*

1. *Agent s's decision to add proposition ψ to her mental state ms:*
   $d2a_i(s, \psi, ms) \in \mathcal{L}_D$ with $i \in \{1, 2, 3, 4\}$.

2. *Agent s's decision to retract proposition ψ from her mental state ms:*
   $d2r_i(s, \psi, ms) \in \mathcal{L}_D$ with $i \in \{1, 2, 3, 4\}$.

A decision rule for decision $\delta i \in \mathcal{L}_D$ consists of a precondition, which specifies a certain mental state structure, denoted $pre(\delta i) \in \mathcal{L}_{MS}$, and a consequent of δi, which is the post-condition on a mental state structure, denoted $post(\delta i) \subseteq \mathcal{L}'_{MS}$. The precondition specifies Sarah's mental state structure before she makes the decision; the post-condition specifies Sarah's mental state structure directly after she has made the decision. The following definition of a decision does not specify specific preconditions and post-conditions of decision. The three subsequent sections 6.3 up to 6.5 we define decisions $d2a_1$ and $d2r_1$ up to $d2a_4$ and $d2r_4$ by providing the specific sets of preconditions and post-conditions.

Definition 6.2 (decision rule). *Given the languages of the mental state $\mathcal{L}'_{MS} \subseteq \mathcal{L}_{MS}$, and the language of decisions $\mathcal{L}_D$, a* decision rule *for a decision $\delta i \in \mathcal{L}_D$ is a structure $(\delta i, pre(\delta i), post(\delta i))$ with $pre(\delta i) \in \mathcal{L}_{MS}$ the precondition of δi, and $post(\delta i) \subseteq \mathcal{L}'_{MS}$ the post-condition of δi.*

Instead of a precondition $pre(\delta i) \in \mathcal{L}_{MS}$, we will sometimes talk about a *set* of preconditions. Such a (finite) set is an abbreviation of a (finite) conjunction of sentences from $\mathcal{L}_{MS}$. Similarly, instead of a set of post-conditions $post(\delta i) \subseteq \mathcal{L}'_{MS}$, we will sometimes talk about the post-conditions as a conjunction from $\mathcal{L}_{MS}$.

We defined the structure of a decision rule; next, we will discuss the activity governed by decision rules called decision-making.

6.1.4 The activity of decision-making

Sarah is allowed to make decision δi if the preconditions of δi hold in her mental state structure, i.e. $\mathcal{M}_s \models pre(\delta i)$. Directly after Sarah has made decision δi, her mental state structure will have been updated with the post-conditions of the decision, i.e. $\mathcal{M}'_s \models post(\delta i)$. We will not define the actual activity of updating an agent's mental state structure such that the post-conditions of a decision will become to hold. We only provide the preconditions and post-conditions. The activity in which Sarah is said to make decision will be governed by forward chaining production rules (cf. Section 6.1.1).

The activity of making a decision should change the agent's mental state structure. If a decision does not have observable effects on an agent's mental state structure, then whether or not the agent makes the decision would not make an observable difference. Such a decision could not serve a purpose. If for all decision rules holds that if the post-conditions are valid in an agent's mental state structure, i.e. $\mathcal{M} \models post(\delta i)$, and the preconditions are not valid in that mental state structure, i.e. $\mathcal{M} \not\models pre(\delta i)$, then the decision changes the agent's mental state structure. To enforce that decisions do change an agent's mental state structure we make the following assumption.

Assumption 6.1. *For all decision rules, the negation of the post-condition of the decision is part of the precondition of the decision.*

Example 6.1 (decision rule)**.** *If set* $\{(\psi_2 \in \mathcal{T}_2), (\psi_3 \notin \mathcal{T}_3)\} \subseteq \mathcal{L}_{MS}$ *is the post-condition of decision* δi*, then with assumption 6.1 we have* $\neg((\psi_2 \in \mathcal{T}_2) \wedge (\psi_3 \notin \mathcal{T}_3)) \in pre(\delta i)$*. If we extend the preconditions of* δi *exclusively with* $(\psi_1 \in \mathcal{T}_1) \in \mathcal{L}_{MS}$*, then we have decision rule:* $(\delta i, (\psi_1 \in \mathcal{T}_1) \wedge \neg((\psi_2 \in \mathcal{T}_2) \wedge (\psi_3 \notin \mathcal{T}_3)), \{(\psi_2 \in \mathcal{T}_2), (\psi_3 \notin \mathcal{T}_3)\})$*. Directly after Sarah has made decision* δi *holds that* $\mathcal{M}_s \models post(\delta i)$ *and* $\mathcal{M}_s \not\models pre(\delta i)$*.*

In formalising decision-making using the application of decision rules, the focus is on functional input-output behaviour of the agent: the agent starts with a certain initial mental state structure, and after decision-making, she has a different one. In many decision processes, several possible sets of conclusions, that is, different specifications of a mental state structure, can be reached. This phenomenon occurs in different formalisations of reasoning. In belief revision for example, retracting beliefs in a minimal fashion can be done in several ways. In practical decision-making, as exercised by our agents, different allowed decisions may lead to different sets of conclusions and different mental state structures. We will regard these multiple end states, often called extensions or expansions, a feature of the specific domain that the agents represent.

Next to adding new beliefs, our agents may decide to retract their beliefs. Retracting beliefs or non-monotonic reasoning in general is a subfield of Artificial Intelligence trying to find more realistic formal models of reasoning than classical logic. In common sense reasoning, one often draws conclusions that have to be withdrawn when further information is obtained. The set of conclusions thus does not grow monotonically with the given information. It is this phenomenon that nonmonotonic reasoning methods try to formalize. Many systems that exhibit non-monotonic behaviour have been described and studied in the literature. Examples are e.g. Reiter [Rei78] who set forth the rule of negation called the closed world assumption.

With this rule, we can assume the negation of a proposition if we cannot prove the proposition (in Horn logic theories). Clark [Cla78] described the rule of negation as failure. Informally, negation as failure is related to the common default assumption that what is not known to be true is false, negation as failure is an interpretation of logical negation according to which the negation of a formula is true if and only if the formula cannot be proved true. A non-monotonic logic is a formal logic whose consequence relation is not monotonic (cf. McDermott and Doyle [MD80]). Circumscription (McCarthy [McC80]) is a non-monotonic logic that deals with the minimization of predicates subject to restrictions expressed by predicate formulas. Informally, circumscription formalizes the common sense assumption that things are as expected unless otherwise specified. Default logic (Reiter [Rei80, Rei87]) is a non-monotonic logic that formalizes reasoning with default assumptions. Default logic aims at formalizing inference rules without explicitly mentioning all their exceptions. In autoepistemic logic (McDermott and Doyle [MD80], Marek and Truszczyski [MT91]) a modal operator is provided to reconstruct a non-monotonic logic as a model of an ideally rational agent's reasoning about its own beliefs. These non-monotonic logics do not model the reasoning of an agent but only provide static systems. We will not provide a logic for our agents' reasoning, we will only assume that reasoning ends.

Another salient aspect of decision-making, or reasoning in general, is that the process of making decisions may never end. We assume that our agent's decision processes do end. This assumption may seem strong; however, as viewed from a programming perspective, this is a quite normal precondition. A much stronger precondition is that in order to enforce termination of the decision process, the rules from the agent's knowledge base and her epistemology are taken to terminate with the agents' decision rules. The agent cannot herself determine that her decision process will not terminate, we take it to be the responsibility of the medical expert to provide the agent with rules such that endless decision-making does not occur.

Definition 6.3 (closure of mental state structures)**.** *The closure of a mental state structure $\mathcal{M}$ under a set of decision rules Δ, denoted $Cl(\mathcal{M}, \Delta)$, has the idempotency property: $Cl(Cl(\mathcal{M}, \Delta)) = Cl(\mathcal{M}, \Delta)$. In general, the closure of mental state structures is not increasing and not monotone.*

If Sarah cannot make decisions anymore, then her mental state structure is closed under the process of decision-making.

Definition 6.4 (closed under decision-making *Cl*)**.** *Given a set of decision rules Δ, a mental state structure $\mathcal{M}$ is* closed under decision-making *with decisions from Δ, if and only if for all $(\delta i, pre(\delta i), post(\delta i)) \in \Delta$ holds $\mathcal{M} \not\models pre(\delta i)$.*

We defined the activity of decision-making; next, we define the consequences this activity has on the agent's mental states.

6.1.5 Cognitive entrenchment

Sarah's belief in ψ is entrenched in her cognitive state in the following situation. If Sarah decides to retract her belief in ψ, then the closure of her mental state structure under decision-making will make her believe ψ again. If Sarah's belief in ψ is *not* cognitively entrenched, then retracting her belief in ψ and closing the resulting mental state structure under decision-making will not make her believe ψ again.

Definition 6.5 (cognitively entrenched presence $\in^{\in}$). *Given a mental state structure $\mathcal{M}$, a mental state $\mathcal{T} \in \mathcal{M}$, and a set of decision rules Δ, the* presence *of ψ in $\mathcal{T}$ is* cognitively entrenched *in $\mathcal{T}$ for $\mathcal{M}$ and Δ, denoted $\mathcal{M}, \Delta \models \psi \in^{\in} \mathcal{T}$, if and only if $Cl(retr_{ms}(\psi, \mathcal{T})(\mathcal{M}), \Delta) \models \psi \in \mathcal{T}$.*

Similar to Sarah's beliefs that can be entrenched in her cognitive state are her ignorance about propositions that can also be entrenched in her cognitive state. Sarah's lack of belief is entrenched in the following situation. If Sarah decides to believe ψ, then the closure of her mental state structure under decision-making will make her ignorant about ψ again. The absence of a proposition from her belief state is entrenched in her cognitive state. If Sarah's ignorance about ψ is *not* cognitively entrenched, then adding her belief in ψ and closing the resulting mental state structure under decision-making will not make her ignorant about ψ again.

Definition 6.6 (cognitively entrenched absence $\in^{\notin}$). *Given a mental state structure $\mathcal{M}$, a mental state $\mathcal{T} \in \mathcal{M}$, and a set of decision rules Δ, the* absence *of ψ in $\mathcal{T}$ is* cognitively entrenched *in $\mathcal{T}$ for $\mathcal{M}$ and Δ, denoted $\mathcal{M}, \Delta \models \psi \in^{\notin} \mathcal{T}$, if and only if $Cl(add_{ms}(\psi, \mathcal{T})(\mathcal{M}), \Delta) \models \psi \notin \mathcal{T}$.*

Definition 6.7 (semantics of entrenchment in $\mathcal{L}_{MS}$). *In addition to definition 5.4, given a cognitive state $CS = \langle \mathcal{M}, \mathcal{K}, \mathcal{E} \rangle$, sentences $\ulcorner \psi \in^{\in} \mathcal{T} \urcorner, \ulcorner \psi \in^{\notin} \mathcal{T} \urcorner \in \mathcal{L}_{MS}$ are satisfied in $\mathcal{M}$, alongside the $\mathcal{K}$ and $\mathcal{E}$, together with a set of decision rules Δ if:*

- *for $\ulcorner \psi \in^{\in} \mathcal{T} \urcorner \in \mathcal{L}_{MS}$ holds $\mathcal{M}, \Delta \models \ulcorner \psi \in^{\in} \mathcal{T} \urcorner$ if $\mathcal{M}, \Delta \models \psi \in^{\in} \mathcal{T}$;*
- *for $\ulcorner \psi \in^{\notin} \mathcal{T} \urcorner \in \mathcal{L}_{MS}$ holds $\mathcal{M}, \Delta \models \ulcorner \psi \in^{\notin} \mathcal{T} \urcorner$ if $\mathcal{M}, \Delta \models \psi \in^{\notin} \mathcal{T}$.*

6.2 Generic Criteria to Change Beliefs

In this section, we will describe the common denominator of the agent's decisions to believe propositions, and the diametrically opposed decisions to become ignorant about propositions.

6.2.1 Generic criteria for decisions to believe

Sarah's decision to believe proposition ψ, i.e. the decision to add ψ to her belief state, $d2a(s, \psi, B_s) \in \mathcal{L}_D$, has *at least* the following preconditions: Sarah is ignorant about ψ, and her ignorance about ψ is not cognitively entrenched in her cognitive state. This description is formalised in the following equation by substitution $\mathcal{T} := B_s$.

$$(\psi \notin \mathcal{T}), (\psi \notin^{\notin} \mathcal{T}) \;\in\; \mathit{pre}(\mathit{d2a}(s, \psi, \mathcal{T})) \tag{6.1}$$

The set of post-conditions of the decision to believe a proposition is equal for all different uses, which we will present in subsequent sections. If an agent decides to believe ψ, she will believe ψ.

$$(\psi \in \mathcal{T}) \;\in\; \mathit{post}(\mathit{d2a}(s, \psi, \mathcal{T})) \tag{6.2}$$

The post-conditions can be enforced on Sarah's mental state structure, as defined in Section 5.3.1, by the action $\mathit{add}_{ms}(\psi, \mathcal{T})(\mathcal{M}_s)$.

6.2.2 Generic criteria for decisions to be ignorant

Sarah's decision to be ignorant about proposition ψ, i.e. the decision to retract ψ from her belief state, $\mathit{d2r}(s, \psi, B_s) \in \mathcal{L}_D$, has *at least* the following preconditions. Sarah believes ψ, and her belief in ψ is not cognitively entrenched in her cognitive state. This description is formalised in the following equation by substitution $\mathcal{T} := B_s$.

$$(\psi \in \mathcal{T}), (\psi \notin^{\in} \mathcal{T}) \;\in\; \mathit{pre}(\mathit{d2r}(s, \psi, \mathcal{T})) \tag{6.3}$$

Irrespective of the precise set of preconditions to decide to be ignorant about a proposition ψ, if an agent decides to be ignorant about ψ, she will not believe ψ.

$$(\psi \notin \mathcal{T}) \;\in\; \mathit{post}(\mathit{d2r}(s, \psi, \mathcal{T})) \tag{6.4}$$

The post-conditions can be enforced on Sarah's mental state structure, as defined in Section 5.3.1, by the action $\mathit{retr}_{ms}(\psi, \mathcal{T})(\mathcal{M}_s)$.

As defined in Section 2.3.1, the meaning for an agent to have a mental state structure in which she believes ψ is that the preconditions to decide to believe ψ have been met in the past, and that ever since the preconditions to be ignorant about this ψ have not been met, and that if these preconditions had been met, the agent has not had the opportunity to retract her belief in ψ. Analogously, the meaning for an agent to be ignorant about ψ, i.e. not to believe ψ, is that the preconditions to decide not to believe ψ have been met, and that ever since the agent was not justified to decide to believe ψ, and that if she had been justified to do so, she has not had the opportunity to add ψ to her belief state. Thus, to remain in a state of believing ψ, or being ignorant about ψ, presupposes that the decisions to retract ψ, or to add ψ, respectively, have not been allowed.

The truism that should be reflected in an agent's decision rules is that if the agent believes a proposition, then she cannot be ignorant about this proposition (cf. equation (5.2)). It is of vital importance that the use of the decision to believe a proposition does not interfere with the use of the decision to be ignorant about that same proposition. The meaning of the decision to believe and the meaning of the decision to be ignorant have to be chosen (i.e. agreed upon) such that after an agent has decided to believe a proposition ψ, a decision rule to decide to be ignorant about ψ should *not* directly be applicable. The interference of both rules may lead to oscillatory behaviour in which an agent decides to believe a proposition ψ, after which she directly decides to be ignorant about ψ, after which she decides to believe ψ again, *ad infinitum*. Such oscillatory, or even chaotic, behaviour is a challenging problem for multiagent systems and distributed artificial intelligence in general since it is not obvious how to avoid or mitigate such harmful overall system behaviour (see Moulin and Chaib-Draa [MCD96]).

6.3 Inference Rules and Meaning to Change Beliefs

In this section, we discuss when our agents are justified to infer new beliefs from existing beliefs, that is, when agents may decide to adopt beliefs, and when agents may decide to retract beliefs. Additionally, we discuss when our agents are justified to decide to change their beliefs based on the meaning of propositions.

6.3.1 Deciding to add beliefs using inference rules

In the following paragraphs, we provide a semantic for the inference rules that have been defined in Section 5.1.4 and that make up the agent knowledge base. Similar to Modus Ponens, if the antecedent of an inference rule holds for the agent's mental state structure but not the consequent, then by application of the inference rule, the agent will decide that the consequent holds in her mental state structure. Under the assumption that Sarah is justified to know the inference rule and that the antecedent holds in her mental state structure, then Sarah is justified to change her mental state structure to reflect the consequent.

Sarah is explicitly justified to decide to believe ψ in the following situation. First, in addition to the generic preconditions $d2a(s,\psi,B_s)$ from Section 6.2.1, she knows an inference rule $\pi \rightarrowtail (\psi \in B_s) \in \mathcal{K}_s$. Second, as required by assumption 6.1, the consequent of the inference rule does not hold for her mental state structure, i.e. she does not believe ψ. Third, the antecedent of the inference rule holds in her mental state structure, i.e. $\mathcal{M}_s \models \pi$. This precondition is generalised by replacing Sarah's belief B_s by a general mental state name $\mathcal{T} \in MSN_s$. Thus, Sarah is justified to have $\psi \in \mathcal{T}$ if she knows an inference rule $\pi \rightarrowtail (\psi \in \mathcal{T}) \in \mathcal{K}_s$, and $\mathcal{M}_s \models \pi$. This precondition on Sarah mental state structure $\mathcal{M}_s$ is formalised in the following equation.

$$\begin{aligned} pre(d2a(s,\psi,\mathcal{T})) &\subset pre(d2a_1(s,\psi,\mathcal{T})) \\ \pi &\in pre(d2a_1(s,\psi,\mathcal{T})) \end{aligned} \tag{6.5}$$

Example 6.2. *If Sarah regards herself entitled to know the inference rule $(\phi \in B_s) \rightarrowtail (\psi \in B_s) \in \mathcal{K}_s$ with ψ reads as "the patient is suffering from disease X", and ϕ reads as "the patient shows symptom Y", then, alongside the decision rule from equation* (6.5), *the inference rule* *has the following reading. "If I (Sarah) believe that symptom Y has been observed by the patient, then I am justified to believe that the patient suffers from disease X."*

6.3.2 Deciding to retract beliefs using inference rules

Analogous to the previous paragraph, Sarah is justified to decide to be ignorant about ψ in the following situation. First, in addition to the generic preconditions $d2r(s, \psi, B_s)$ from Section 6.2.2, she knows an inference rule $\pi \rightarrowtail (\psi \notin B_s) \in \mathcal{K}_s$. Second, the consequent does not hold in her mental state structure, i.e. Sarah believes ψ. Third, the antecedent holds in her mental state structure, i.e. $\mathcal{M}_s \models \pi$. This precondition is generalised by replacing Sarah's belief B_s by a general mental state structure $\mathcal{T} \in MSN_s$. Thus, Sarah is justified to have $\psi \notin \mathcal{T}$ if she knows an inference rule $\pi \rightarrowtail (\psi \notin \mathcal{T}) \in \mathcal{K}_s$, and $\mathcal{M}_s \models \pi$. This precondition on Sarah mental state structure $\mathcal{M}_s$ is formalised in the following equation, and is part of the precondition of Sarah's decision to have $\psi \notin \mathcal{T}$.

$$\begin{aligned} pre(d2r(s, \psi, \mathcal{T})) &\subset pre(d2r_1(s, \psi, \mathcal{T})) \\ \pi &\in pre(d2r_1(s, \psi, \mathcal{T})) \end{aligned} \tag{6.6}$$

The generalisations (of the decision to retract and the decision to add) to include not only to change the agent's beliefs, but to allow inference rules that can change any mental state, provide a practical mechanism for designer to program the agents to change their mental state structures with inference rules. We provide this mechanism in anticipation of Chapter 8 where we will provide inference rules that program the agents to change their mental state such that they will resolve their disagreements or agree to disagree.

6.3.3 Deciding to add and retract beliefs using meaning

Informally, an agent is justified to decide to believe a proposition if she regards herself entitled to know the meaning of the proposition, i.e. she knows the criterion for the correct use of the proposition, and the agent is aware that the criterion has been met in her mental state structure.

The decision rules that agents may use to deduce beliefs based on meaning are similar to the decision rules for deducing beliefs with inference rules. The difference with decision rules from the previous paragraph is that instead of knowing an inference rule, an agent regards herself entitled to know a conventional rule that conveys the meaning of being in some mental state, such as believing. If an agent knows a conventional rule, then, according to the agent, the *use* of the criterion to establish the consequent of the rule is common knowledge. Because, according to the agent, the use is common knowledge, the agent can be said to know that other agents use the same criterion to establish some mental state. While an inference rule that is part of an agent's knowledge base provides an agent with a relation that transfers justification from the antecedent to the consequent, an inference rule that is part of an agent's epistemology provides an agent with a relation that can also be used to describe that another agent transfers justification from the antecedent to the consequent.

Sarah is justified to decide to believe a proposition ψ if, in addition to the general preconditions $d2r(s, \psi, B_s)$, Sarah knows the meaning of believing ψ, i.e. she knows an inference rule is common knowledge $\pi \rightarrowtail (\psi \in B_s) \in \mathcal{E}_s$, and the antecedent π holds in her mental state structure. This precondition on $\mathcal{M}_s$ is generalised by replacing Sarah's belief B_s by a general mental state structure $\mathcal{T} \in MSN_s$.

$$\begin{aligned} pre(d2a(s, \psi, \mathcal{T})) &\subset pre(d2a_2(s, \psi, \mathcal{T})) \\ \pi &\in pre(d2a_2(s, \psi, \mathcal{T})) \end{aligned} \tag{6.7}$$

Sarah is justified to decide to be ignorant about ψ if, in addition to the general preconditions $d2r(s, \psi, B_s)$, she knows the meaning of being ignorant about ψ, i.e. she knows an inference rule is common knowledge $\pi \rightarrowtail (\psi \notin B_s) \in \mathcal{E}_s$, and the antecedent π holds in her mental state structure. This precondition on $\mathcal{M}_s$ is generalised by replacing Sarah's belief B_s by a general mental state structure $\mathcal{T} \in MSN_s$.

$$\begin{aligned} pre(d2r(s, \psi, \mathcal{T})) &\subset pre(d2r_2(s, \psi, \mathcal{T})) \\ \pi &\in pre(d2r_2(s, \psi, \mathcal{T})) \end{aligned} \tag{6.8}$$

6.3.4 Deciding that other agents add and retract beliefs

If an agent regards herself entitled to know the meaning of a proposition, then the agent knows the criterion that has been agreed upon by her community. Agents are thus said to know that other agents in their community use the same criterion as they do. Thus, Sarah may infer that the consequences of a conventional rule hold in John's mental state structure if the preconditions of the rule hold according to Sarah for John's mental state structure. We need a function that translates a set of conditions on Sarah's mental state structure to the set of conditions that, as seen from Sarah's perspective, hold for John's mental state structure. For example $\psi \in B_s$ has to be translated to $\psi \in B_sB_j$. First, we define a function that translates sentences (i.e. basic literals) from the $\mathcal{L}'_{MS}$, then we define a function that translates sentences from the full language of $\mathcal{L}_{MS}$.

Definition 6.8 (translation function $tran_1$). *Given a $\mathcal{L}'_{MS}$, and a set of agent names $\mathcal{A}$,* translation function *$tran_1(x, \pi)$ translates sentence $\pi \in \mathcal{L}'_{MS}$ to the perspective of agent x yielding $\pi' \in \mathcal{L}'_{MS}$.*

- *If $x \neq s$, then $tran_1(x, \psi \in B_s) = \psi \in B_sB_x$, $tran_1(x, \psi \notin B_s) = \psi \in B_s\widetilde{B}_x$, $tran_1(x, \psi \in D_sB_s) = \psi \in B_sD_xB_x$, and $tran_1(x, \psi \in D_s\widetilde{B}_s) = \psi \in B_sD_x\widetilde{B}_x$.*
- *If $x \neq s$ and $a \neq s$, then $tran_1(x, \psi \in D_sB_a) = \psi \in B_sD_xB_a$, and $tran_1(x, \psi \in D_s\widetilde{B}_a) = \psi \in B_sD_x\widetilde{B}_a$.*
- *If $x \neq s$, then $tran_1(x, \psi \in B_s\mathcal{T}_j) = \psi \in B_sB_x\mathcal{T}_j$, and $tran_1(x, \psi \notin B_s\mathcal{T}_j) = \psi \in B_s\widetilde{B}_x\mathcal{T}_j$, with $\mathcal{T}_j \in \{B_j, \widetilde{B}_j\}$.*

- *If $x \neq s$, then $tran_1(x, \psi \in B_s D_j B_a) = \psi \in B_s B_x D_j B_a$, and $tran_1(x, \psi \notin B_s \mathcal{T}_j) = \psi \in B_s \widetilde{B}_x \mathcal{T}_j$, with $\mathcal{T}_j \in \{B_j, \widetilde{B}_j\}$.*

- *In all other cases, $tran_1(s, \pi) = \top$.*

Definition 6.9 (translation function *tran*$_2$)**.** *Given the languages of mental state $\mathcal{L}'_{MS} \subseteq \mathcal{L}_{MS}$, and a set of agent names $\mathcal{A}$,* translation function *$tran_2(x, \pi)$ translates sentence $\pi \in \mathcal{L}_{MS}$ to the perspective of agent $x \in \mathcal{A}$ yielding $\pi' \in \mathcal{L}_{MS}$ with:*

- *If $\pi \in \mathcal{L}'_{MS}$, then $\pi' = tran_1(x, \pi)$.*

- *If $\pi = \neg\pi_1$, then $\pi' = \neg tran_2(x, \pi_1)$.*

- *If $\pi = \pi_1 \wedge \pi_2$, then $\pi' = tran_2(x, \pi_1) \wedge tran_2(x, \pi_2)$.*

Example 6.3. *Sarah knows that in her community, a criterion to believe ϕ is that ϕ is believed by another agent and that she does not believe ψ, i.e. $(((\phi \in B_s B_j) \wedge (\phi \notin B_s)) \rightarrowtail (\psi \in B_s)) \in \mathcal{E}_s$. Because Sarah knows the rule is used by all agents in her community, she may want to test whether the antecedent holds for another agent $x \in \mathcal{A}$, so we have $tran_2(x, (\phi \in B_s B_j) \wedge (\phi \notin B_s)) = (\phi \in B_s B_x B_j) \wedge (\phi \in B_s \widetilde{B}_x)$.*

Sarah may decide to believe that John believes ψ, i.e. $\psi \in B_s B_j$, if she does not believe that John believes ψ, i.e. $\psi \notin B_s B_j$, she knows that π is a criterion to believe ψ, i.e. $\pi \rightarrowtail (\psi \in B_s) \in \mathcal{E}_s$, and according to her mental state structure, π holds in John's mental state structure, i.e. $\mathcal{M}_s \models tran_2(j, \pi)$. In such a situation, Sarah may deduce that John believes ψ, i.e. $\psi \in B_s B_j$. Instead of defining a decision rule that implements this behaviour, we will define a different decision rule that, alongside the coherence relations from Section 5.2.1, will make Sarah perform the same deductions.

Sarah may decide to be ignorant that John is ignorant about ψ, i.e. $\psi \notin B_s \widetilde{B}_j$, if she believes that John is ignorant about ψ, i.e. $\psi \in B_s \widetilde{B}_j$, she knows that π is a criterion to believe ψ, i.e. $\pi \rightarrowtail (\psi \in B_s) \in \mathcal{E}_s$, and according to her mental state structure, π holds in John's mental state structure, i.e. $\mathcal{M}_s \models tran_2(j, \pi)$. The post-condition of this decision will result in a mental state structure, due to coherence relation $\mathcal{M}_s \models \psi \in B_s B_j \Rightarrow \mathcal{M}_s \models \psi \notin B_s \widetilde{B}_j$ (eq. (5.3)), in which the proposed post-condition of the decision of the previous paragraph holds. The precondition of the decision of the previous paragraph, $\psi \notin B_s B_j$, will hold in the mental state structure, due to the contraposition of equation (5.3), if the precondition of the following decision holds. Sarah's decision to retract ψ from her belief about John's ignorance is denoted $d2r(s, \psi, B_s \widetilde{B}_j)$.

$$(\psi \notin B_s \widetilde{B}_j),\ tran_2(j, \pi)\ \in\ pre(d2r_2(s, \psi, B_s \widetilde{B}_j)) \tag{6.9}$$

Sarah may decide to believe that John is ignorant about ψ, i.e. $\psi \in B_s \widetilde{B}_j$, if she is ignorant whether John is ignorant about ψ, i.e. $\psi \notin B_s \widetilde{B}_j$, she knows that π is the criterion to be ignorant about ψ, i.e. $\pi \rightarrowtail (\psi \notin B_s) \in \mathcal{E}_s$, and according to her mental state structure, π holds in John's mental state structure, i.e. $\mathcal{M}_s \models tran_2(j, \pi)$. In such situation, Sarah may deduce that John is ignorant about ψ, i.e. $\psi \in B_s\widetilde{B_j}$. This decision behaviour will be subsumed under the following decision rule.

Sarah may decide to be ignorant that John believes ψ, i.e. $\psi \notin B_sB_j$, if she believes that John believes ψ, i.e. $\psi \in B_sB_j$, she knows that π is the criterion to be ignorant about ψ, i.e. $\pi \rightarrowtail (\psi \notin B_s) \in \mathcal{E}_s$, and $\mathcal{M}_s \models tran_2(j,\pi)$ holds in John's mental state structure. Due to equation (5.3) and its contraposition, the proposed decision rule from the previous paragraph is subsumed under the following decision rule.

$$(\psi \in B_sB_j),\ tran_2(s,\pi)\ \in\ pre(d2r_2(s,\psi,B_sB_j)) \tag{6.10}$$

Example 6.4. *If Sarah regards herself entitled to know the conventional rule* $(\phi \in B_s) \rightarrowtail (\psi \in B_s) \in \mathcal{E}_s$ *with* ψ *reads as "the patient is suffering from disease X", and* ϕ *reads as "the patient shows symptom Y", then, with decision rule from equation (6.9), the conventional rule has the following reading. "Anybody who believes that symptom Y has been observed by the patient, is justified to believe that the patient suffers from disease X." If Sarah believes that John believes* ϕ*, then Sarah may decide to believe that John believes* ψ *with a* $d2a_2(s,\psi,B_sB_j)$*.*

It is possible that Sarah is justified to believe that John believes ψ while in fact he does not believe ψ. Based on the meaning of believing ψ, Sarah may be justified to believe that John believes ψ. Even though Sarah's beliefs about John's mental state structure are not incorrect, Sarah can be justified to believe that John believes ψ, while, John may not have had the opportunity to decide to believe ψ. Thus, Sarah can come to have beliefs about John that he does not have. However, if both Sarah's and John's mental state structure are closed under decision-making, this situation will be resolved.

6.4 Conformism to Change Beliefs

6.4.1 Conformism to believe

Another ground for agents to be justified to believe a proposition is by conforming to other agents' beliefs. If Sarah believes that John believes a proposition that she does not believe herself, then Sarah may decide to believe the proposition on the basis that John has been justified to decide to believe the proposition in the past. If Sarah has no reason to suspect that her view of the meaning of the predicate *to believe* is different from John's, then Sarah may continue to regard herself entitled to know that her perceived criterion of *to believe* is conventional. If the criterion to predicate *to believe* is shared by Sarah and John, then Sarah can use her belief about John's belief as a substitute for other preconditions, like those presented in Section 6.3; this saves her the effort to check whether other preconditions have been met.

Sarah is justified to decide to believe a proposition if she believes that John believes the proposition. Sarah's belief that John believes a proposition is justified either in testimony, which will be the result of communication that is to be discussed in Chapter 7, or justified in the meaning of propositions, as described in Section 6.3.3.

In anticipation of the decision rules that allow agents to retract beliefs, which we will discuss in Section 6.5.6, Sarah needs to be assured that the source in which she justifies her belief is aware of the dependence of her newly accepted belief. If Sarah believes that John believes that she believes that he believes ψ, i.e. $\psi \in B_sB_jB_sB_j$, then Sarah is justified to believe that if John retracts his belief in ψ, then he will consult her.

Sarah is justified to decide to believe ψ, if, in addition to the general preconditions of $d2a(s, \psi, B_s)$, she believes that John believes ψ, i.e. $\psi \in B_sB_j$, and she believes that John believes this, i.e. $\psi \in B_sB_jB_sB_j$.

$$\begin{array}{rcl} pre(d2a(s,\psi,B_s)) & \subset & pre(d2a_3(s,\psi,B_s)) \\ (\psi \in B_sB_j),(\psi \in B_sB_jB_sB_j) & \in & pre(d2a_3(s,\psi,B_s)) \end{array} \tag{6.11}$$

6.4.2 Conformism to be ignorant

Inverse to the decisions in which agents conform to the beliefs of other agents, as described in Section 6.4.1, we now describe the decisions that allow agents to conform to other agents' ignorance. If Sarah adopts a belief grounded in the beliefs about John's beliefs, she does so without verifying the preconditions John may have used to adopt the belief. If an agent adopts an ignorance state, that is to say, she retracts a belief based on the beliefs about the ignorance of John; she does so without checking whether John has cogent reasons not to believe the proposition.

Sarah is justified to decide to be ignorant about ψ if Sarah believes a proposition ψ, in addition to the general preconditions of $d2r(s, \psi, B_s)$, if she believes that John is ignorant about ψ, i.e. $\psi \in B_s\widetilde{B}_j$, and she believes that John believes this, i.e. $\psi \in B_sB_jB_s\widetilde{B}_j$.

$$\begin{array}{rcl} pre(d2r(s,\psi,B_s)) & \subset & pre(d2r_3(s,\psi,B_s)) \\ (\psi \in B_s\widetilde{B}_j),(\psi \in B_sB_jB_s\widetilde{B}_j) & \in & pre(d2r_3(s,\psi,B_s)) \end{array} \tag{6.12}$$

Decisions may cancel out each other's post-conditions, resulting in deviant behaviour. This behaviour occurs when Sarah's decision to conform to John's ignorance (eq. (6.12)) cancels out the effects of her decision to conform to Fred's beliefs (eq. (6.11)). The truism that if a community comes to believe a proposition then one agent has to be first to do so, provides the following problem. Assume Fred and John both do not believe ψ, and Sarah is the first to believe ψ. If Sarah has used a decision rule we have presented so far, then, because of this decision rule, her belief in ψ is entrenched in her mental state structure. Because Sarah's belief in ψ is entrenched, the decision to conform to either Fred's or John's ignorance will not make her retract her belief in ψ. Fred cannot change his beliefs because of a similar reason: the decision rule that allows Fred to conform to John's ignorance entrenches his ignorance to conform to Sarah's belief, which means that Fred cannot conform to Sarah's beliefs, and can not decide to believe ψ. The same argument holds for John: he cannot conform to Sarah's belief in ψ, because his ignorance is entrenched in his belief that Fred is ignorant about ψ. For the sake of argument, assume that Fred flouts the rules once, and does conform to Sarah's belief and decides to believe

ψ. John may now conform to either John or Sarah's belief and decide to believe ψ because he is the last to be ignorant about ψ; however, this is only possible because Fred has changed the meaning of believing ψ. The question remains how Sarah first became justified to believe the proposition. Assume Sarah knows an inference rule that justifies her to believe ψ. If this rule is applicable, she may still not decide to believe ψ with $d2a_1$ (eq. (6.5)) because her ignorance about ψ is entrenched with $d2r_3$ (eq. (6.12)). Thus, the only way that Sarah could have become the first to believe ψ is if she, just like Fred, has flouted the rules. In the following section, we resolve this problem by augmenting the preconditions to conform to other agents' beliefs such that agents may change their beliefs only if they desire to do so.

6.5 Desiring to Change Beliefs

We assume that an agent's desires originate from the human expert she is said to represent, and that ultimately, the agent's beliefs will have purpose for this expert. In this section, we will restrict the activity of conforming to other agents' beliefs, as described in Section 6.4, to those beliefs that will serve a purpose. Our agents will use desires as motivational aspects such that she makes decisions with the intention to balance her mental state structure.

6.5.1 The motivations to make decisions

We will restrict the use of decisions to those situations that an agent considers to have purpose. When a decision's post-conditions are desired or indirectly desired, as will be defined in Section 6.5.4, then our agents will consider the decision to have a purpose. In *The Fixation of Belief* [Men97, p. 13], Peirce goes even further when he states that beliefs that are not desired should be rejected.

> "The irritation of doubt causes a struggle to attain a state of belief. . . . It is certainly best for us that our beliefs should be such as may truly guide our actions so as to satisfy our desires; and this reflection will make us reject any belief which does not seem to have been so formed as to insure this result."

Our agents may have desires for different reasons. We assume that Sarah's desires regarding an agent $a \in \mathcal{A}$, as specified with the mental states D_sB_a and $D_s\widetilde{B_a}$, originate from the human experts the agents represent. We assume that if agents desire to believe or desire to be ignorant about propositions, then believing or disbelieving these propositions serves a purpose for the experts. A desire alone is not considered a sufficient condition to come to believe propositions, neither for humans nor for agents. As Peirce said, desires will guide an agent's actions so as to satisfy her desires, and thus guide the adoption of beliefs. If our agents are constrained to conform to other agents' beliefs that they desire to believe, then our agents do not have to conform to all other beliefs others may have.

If an agent desires her mental state structure to possess a certain property, which it does not possess, then the agent has an unbalanced mental state structure. For example, if Sarah desires to believe ψ, and she does not believe ψ, then she has an unbalanced mental state structure. We will use these unbalanced mental states as motivational aspects to direct the agents' activities. If an agent has an unbalanced mental state structure, she is motivated to act with the intention of balancing her mental state structure. Beun [Beu01] used a similar account of unbalanced mental states to motivate communication between agents. In Chapter 7, we will use unbalanced mental state structures to describe when agents are motivated to communicate with the intent to become justified to make decisions that balance their mental states. Sarah can balance her mental state structure by either changing her beliefs or by changing her desires. Because an agent's desires originate from a human expert, agents are not allowed to change those; what agents are allowed to change are their beliefs.

1. Sarah has an unbalanced mental state structure if she desires to believe a proposition ψ, and, at the same time, she is ignorant about ψ, i.e. she does not believe ψ. See equation (6.13). If Sarah has such an unbalanced mental state structure, she is motivated to decide to add ψ to her beliefs, because this would balance her mental state structure.

$$\mathcal{M}_s \models (\psi \in D_sB_s),\ (\psi \notin B_s) \tag{6.13}$$

2. An analogous unbalanced mental state structure is when Sarah desires to be ignorant about ψ, and, at the same time, she believes ψ, i.e. she is not ignorant about ψ. See equation (6.14). If Sarah has such an unbalanced mental state structure, she is motivated to decide to retract ψ, because this would balance her mental state structure.

$$\mathcal{M}_s \models (\psi \in D_s\widetilde{B_s}),\ (\psi \in B_s) \tag{6.14}$$

6.5.2 Cognitive preconditions

An agent's unbalanced mental state structure can be balanced by updating the mental state structure with certain basic literals $\pi \in \mathcal{L}'_{MS}$. To establish which basic literals are needed to balance a mental state structure, not only are the agent's current mental state structure and her desires needed, such as D_sB_a and $D_s\widetilde{B_a}$, but also the decision rules that allow her to change her beliefs. Informally, if Sarah's mental state structure will be closed under her decision rules, then updating her mental states with sets of basic literals $\pi \in \mathcal{L}'_{MS}$ may make her justified to make a decision which has as a consequence that she balances her mental state structure. We now define these sets of preconditions. In Section 6.5.4 we will use these cognitive preconditions to define how Sarah can update her mental states to balance it.

Given that Π_1 is a coherent subset of $\mathcal{L}'_{MS}$ (def. 5.16) and $\Pi_2 \subseteq \mathcal{L}'_{MS}$ is a set of conditions on a mental state structure, we define when Π_1 is a cognitive precondition of Π_2. Informally, Π_1 is a cognitive precondition of Π_2 if the mental state structure is updated with Π_1, and closed under decision-making, then Π_2 will hold in the resulting mental state structure. Additionally, if the mental state structure is updated with Π_1 and not closed under decision-making, then Π_2 does not hold in the resulting mental state structure. That is to say, updating the mental state structure with Π_1 indeed changed the mental state structure such that Π_2 becomes valid.

Definition 6.10 (cognitive precondition ///)**.** *Given a $\mathcal{L}'_{MS}$, a coherent subset of $\mathcal{L}'_{MS}$ Π_1 (def. 5.16) is a* cognitive precondition *of* $\Pi_2 \subseteq \mathcal{L}'_{MS}$ *for mental state structure $\mathcal{M}$ and a set of decision rules Δ, denoted $\mathcal{M}, \Delta \models \Pi_1 \mathbin{///} \Pi_2$, if and only if:*

- *$Cl(update_s(\Pi_1, \mathcal{M}), \Delta) \models \Pi_2$;*
- *$update_s(\Pi_1, \mathcal{M}) \not\models \Pi_2$.*

Set of conditions Π_1 is minimal cognitive precondition of Π_2, informally, if Π_1 is a cognitive precondition of Π_2 and all proper subsets of Π_1 are not a cognitive precondition of Π_2.

Definition 6.11 (minimal cognitive precondition //)**.** *Given a $\mathcal{L}'_{MS}$, a coherent subset of $\mathcal{L}'_{MS}$ Π_1 (def. 5.16) is a* minimal cognitive precondition *of* $\Pi_2 \subseteq \mathcal{L}'_{MS}$ *for $\mathcal{M}$ and Δ, denoted $\mathcal{M}, \Delta \models \Pi_1 \mathbin{//} \Pi_2$, if and only if:*

- *$\Pi_1 \mathbin{///} \Pi_2$;*
- *$\forall \Pi'_1 \subset \Pi_1 \; \neg(\Pi'_1 \mathbin{///} \Pi_2)$.*

Notation 6.1. *If sets $\Pi_1, \Pi_2 \subseteq \mathcal{L}'_{MS}$ equal a singleton set, we substitute the element for the sets, e.g. instead of $\{\psi_1 \in \mathcal{T}_1\} \mathbin{//} \{\psi_2 \in \mathcal{T}_2\}$ we write $\psi_1 \in \mathcal{T}_1 \mathbin{//} \psi_2 \in \mathcal{T}_2$.*

Example 6.5. *Given a set of decision rules Δ with $\delta r_1, \delta r_2 \in \Delta$:*

- *$\psi_1, \psi_2, \psi_3, \psi_4, \psi_5$ are pair wise unequal; $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_5$ are pair wise unequal;*
- *$\delta r_1 = (\delta i_1, \{\psi_1 \in \mathcal{T}_1, \psi_2 \in \mathcal{T}_2\}, \{\psi_3 \in \mathcal{T}_3\})$;*
- *$\delta r_2 = (\delta i_2, \{\psi_3 \in \mathcal{T}_3, \psi_4 \in \mathcal{T}_4\}, \{\psi_5 \in \mathcal{T}_5\})$;*
- *$\mathcal{M}$ is a mental state structure.*

Then:

- *We have $\mathcal{M}, \Delta \models \{\psi_1 \in \mathcal{T}_1, \psi_2 \in \mathcal{T}_2\} \mathbin{//} \psi_3 \in \mathcal{T}_3$ because with δr_1 we have $Cl(update_s(\{\psi_1 \in \mathcal{T}_1, \psi_2 \in \mathcal{T}_2\}, \mathcal{M}), \Delta) \models \psi_3 \in \mathcal{T}_3$.*
- *We have $\mathcal{M}, \Delta \models \{\psi_3 \in \mathcal{T}_3, \psi_4 \in \mathcal{T}_4\} \mathbin{//} \psi_5 \in \mathcal{T}_5$ because with δr_2 we have $Cl(update(\{\psi_3 \in \mathcal{T}_3, \psi_4 \in \mathcal{T}_4\}, \mathcal{M}), \Delta) \models \psi_5 \in \mathcal{T}_5$.*
- *We have $\mathcal{M}, \Delta \not\models \{\psi_1 \in \mathcal{T}_1, \psi_2 \in \mathcal{T}_2, \psi_3 \in \mathcal{T}_3, \psi_4 \in \mathcal{T}_4\} \mathbin{//} \psi_5 \in \mathcal{T}_5$ because with δr_1 and δr_2 we have $\mathcal{M}, \Delta \models \{\psi_1 \in \mathcal{T}_1, \psi_2 \in \mathcal{T}_2, \psi_4 \in \mathcal{T}_4\} \mathbin{//} \psi_5 \in \mathcal{T}_5$.*

A condition π_1 is partial minimal cognitive precondition of Π_2, informally, if π_1 is an element of a minimal cognitive precondition of Π_2.

Definition 6.12 (partial cognitive precondition /)**.** *Given a* $\mathcal{L}'_{MS}$*, basic literal* $\pi \in \mathcal{L}'_{MS}$ *is a partial minimal cognitive precondition or* partial cognitive precondition*, of* $\Pi \subseteq \mathcal{L}'_{MS}$ *for* $\mathcal{M}$ *and* Δ*, denoted* $\mathcal{M}, \Delta \models \pi \,/\, \Pi$*, if and only if:*

- $\exists \Pi' \subseteq \mathcal{L}'_{MS}\ (\mathcal{M}, \Delta \models \Pi' \,/\!/\, \Pi)$ *;*
- $\pi \in \Pi'$ *.*

6.5.3 Coherent desires

Based on the assumption that an agent's desires are coherent, as described in Section 5.2.3, we may be tempted to conclude that these coherent desires only motivate an agent to make decisions with post-conditions that do not conflict with her desires. For example, if Sarah desires to believe ψ, then under assumption 5.1, she is not allowed to desire to be ignorant about ψ. Thus, using her desire to motivate her decision to come to believe ψ cannot be in conflict with her desire to be ignorant about ψ. However, as we will describe next, mental state structures closed under decision-making may motivate agents to make decisions that have consequences that do conflict with their desires.

If Sarah decides to believe a proposition, then in the resulting mental state structure she may use inference rules (Section 6.3) to decide to believe additional propositions. These inferred propositions may conflict with the agent's desires. For example, assume Sarah desires to believe she is a millionaire, $\psi \in D_s B_s$, and that she desires to believe she is a professor, $\phi \in D_s B_s$. Also, assume that Sarah knows an inference rule that states that if she believes she is a professor, then she cannot believe that she is a millionaire, i.e. $(\phi \in B_s) \rightarrowtail (\psi \notin B_s) \in \mathcal{K}_s$. Sarah's decision to believe that she is a professor does not conflict with her desires, because ϕ is not part of her desires to be ignorant, i.e. $\phi \notin D_s \widetilde{B}_s$. However, if Sarah closes her mental state structure under decision-making, a consequence of her decision to believe ϕ is that with decision rule $d2r_1$ (eq. (6.6)), she will become ignorant that she is a millionaire, i.e. $\psi \notin B_s$. Although the direct consequence of Sarah's decision to believe ϕ does not conflict with her desires, because $\psi \in D_s B_s$, the indirect consequence that Sarah becomes ignorant about ψ does conflict with her desires to believe.

- Sarah's hypothetical belief in ψ does not conflict with her desires, denoted *coherent*$(\psi \in B_s, \Delta_s)$, if and only if the following set of conditions hold for her mental state structure. If Sarah's belief in ψ is a partial cognitive precondition to believe ϕ, i.e. $\psi \in B_s \,/\, \phi \in B_s$, then she should not desire to be ignorant about ϕ. Analogously, if Sarah's belief in ψ is a partial cognitive precondition to be ignorant about ϕ, i.e. $\psi \in B_s \,/\, \phi \notin B_s$, then she should not desire to believe ϕ. The set *coherent*$(\psi \in B_s, \Delta_s)$ is a conjunction of the propositions that should not be present in her desires if Sarah's belief in ψ is to be coherent with her mental state structure. Thus, if holds $\mathcal{M}_s \models \mathit{coherent}(\psi \in B_s, \Delta_s)$, then Sarah can decide to believe ψ without going against her desires.

$$\mathit{coherent}(\psi \in B_s, \Delta_s) = \begin{array}{l} \{(\phi \notin D_s\widetilde{B}_s) \in \mathcal{L}'_{MS} \mid \psi \in B_s \ / \ \phi \in B_s\} \cup \\ \{(\phi \notin D_sB_s) \in \mathcal{L}'_{MS} \mid \psi \in B_s \ / \ \phi \notin B_s\} \end{array} \tag{6.15}$$

- A similar condition states that Sarah's hypothetical ignorance about ψ does not conflict with her desires, denoted $\mathit{coherent}(\psi \notin B_s, \Delta_s)$, if and only if the following set of conditions hold for her mental state structure. If Sarah's ignorance about ψ is a partial cognitive precondition to believe ϕ, i.e. $\psi \notin B_s \ / \ \phi \in B_s$, then she may not desire to be ignorant about ϕ. Analogously, if Sarah's ignorance about ψ is a partial cognitive precondition to be ignorant about ϕ, i.e. $\psi \notin B_s \ / \ \phi \notin B_s$, then she should not desire to believe ϕ. Thus, if holds $\mathcal{M}_s \models \mathit{coherent}(\psi \notin B_s, \Delta_s)$, then Sarah can decide to be ignorant about ψ without going against her desires.

$$\mathit{coherent}(\psi \notin B_s, \Delta_s) = \begin{array}{l} \{(\phi \notin D_s\widetilde{B}_s) \in \mathcal{L}'_{MS} \mid \psi \notin B_s \ / \ \phi \in B_s\} \cup \\ \{(\phi \notin D_sB_s) \in \mathcal{L}'_{MS} \mid \psi \notin B_s \ / \ \phi \notin B_s\} \end{array} \tag{6.16}$$

The previous two conditions state that Sarah's decisions to change her belief are coherent with her desires. Next, we provide conditions in which John's decision to change his beliefs are coherent with Sarah's desires.

- John's decision to believe ψ does not conflict with Sarah's desires, denoted $\mathit{coherent}(\psi \in B_sB_j, \Delta_s)$, if and only if the following conditions hold. If, according to Sarah, John's belief in ψ is a partial cognitive precondition for him to believe ϕ, i.e. $\psi \in B_sB_j \ / \ \phi \in B_sB_j$, then Sarah should not desire that John believes ϕ. Analogously, if, according to Sarah, John's belief in ψ is a partial cognitive precondition for him to be ignorant about ϕ, i.e. $\psi \in B_sB_j \ / \ \phi \in B_s\widetilde{B}_j$, then Sarah should not desire John to be ignorant about ϕ. Thus, if holds $\mathcal{M}_s \models \mathit{coherent}(\psi \in B_sB_j, \Delta_s)$, then, according to Sarah, John can decide to believe ψ without going against Sarah's desires.

$$\mathit{coherent}(\psi \in B_sB_j, \Delta_s) = \begin{array}{l} \{(\phi \notin D_s\widetilde{B}_j) \in \mathcal{L}'_{MS} \mid \psi \in B_sB_j \ / \ \phi \in B_sB_j\} \cup \\ \{(\phi \notin D_sB_j) \in \mathcal{L}'_{MS} \mid \psi \in B_sB_j \ / \ \phi \in B_s\widetilde{B}_j\} \end{array} \tag{6.17}$$

- Analogous arguments state that John's decision to be ignorant about ψ does not conflict with Sarah's desires if and only if the following conditions hold. If, according to Sarah, John's ignorance about ψ is a partial cognitive precondition for him to believe ϕ, i.e. $\psi \in B_s\widetilde{B}_j \ / \ \phi \in B_sB_j$, then Sarah should not desire that John is ignorant about ϕ. Analogously, if, according to Sarah, John's ignorance about ψ is a partial cognitive precondition for him to be ignorant about ϕ, i.e. $\psi \in B_s\widetilde{B}_j \ / \ \phi \in B_s\widetilde{B}_j$, then Sarah should not desire John to believe ϕ. Thus, if holds $\mathcal{M}_s \models \mathit{coherent}(\psi \notin B_sB_j, \Delta_s)$, then, according to Sarah, John can decide to be ignorant about ψ without going against Sarah's desires.

$$\mathit{coherent}(\psi \in B_s\widetilde{B}_j, \Delta_s) = \begin{array}{l} \{(\phi \notin D_s\widetilde{B}_j) \in \mathcal{L}'_{MS} \mid \psi \in B_s\widetilde{B}_j \ / \ \phi \in B_sB_j\} \cup \\ \{(\phi \notin D_sB_j) \in \mathcal{L}'_{MS} \mid \psi \in B_s\widetilde{B}_j \ / \ \phi \in B_s\widetilde{B}_j\} \end{array} \tag{6.18}$$

6.5.4 Indirect coherent desires

In subsequent sections and chapters, our agents will motivate their decisions and communication with their unbalanced mental state structures. We now define indirect desires that are derived from Sarah's direct desires based on her decision-making. Informally, indirect desires are states of Sarah's mental state structure in which, if closed under decision-making, Sarah may be allowed to balance her mental state structure. Additionally, indirect desires do not conflict with Sarah's direct desires.

Sarah will be said to desire indirectly to believe a proposition if the decision to believe the proposition balances her mental state structure, or the decision will balance a mental state structure of another agent.

- The set $ind(D_sB_s)$ is the set of propositions that Sarah indirectly desires to believe. It equals her direct desires to believe, i.e. D_sB_s, combined with the set of propositions ψ that are coherent with her desires, i.e. $coherent(\psi \in B_s, \Delta_s)$, and that contains the following propositions. As usual, take $s, j, a \in \mathcal{A}$ with $j \neq s$ and either $a = s$ or $a = j$.
  - Proposition ψ is an element of $ind(D_sB_s)$ if Sarah's belief in ψ is a partial cognitive precondition for her belief in ϕ, i.e. $\psi \in B_s \ / \ \phi \in B_s$, and either Sarah desires to believe ϕ herself, or she believes that John desires a to believe ϕ, i.e. $(\phi \in D_sB_s) \vee (\phi \in B_sD_jB_a)$.
  - Proposition ψ is an element of $ind(D_sB_s)$ if $\psi \in B_s$ is a partial cognitive precondition for Sarah to be ignorant about ϕ, i.e. $\psi \in B_s \ / \ \phi \notin B_s$, and either Sarah desires to be ignorant about ϕ, or she believes that John desires a to be ignorant about ϕ, i.e. $(\phi \in D_s\widetilde{B_s}) \vee (\phi \in B_sD_j\widetilde{B_a})$.
  - Proposition ψ is an element of $ind(D_sB_s)$ if Sarah's ignorance about ψ is a partial cognitive precondition for her to be ignorant about ϕ, i.e. $\psi \notin B_s \,/\, \phi \notin B_s$, and either Sarah desires to believe ϕ herself, or she believes that John desires a to believe ϕ, i.e. $(\phi \in D_sB_s) \vee (\phi \in B_sD_jB_a)$. By deciding to believe ψ, Sarah's ignorance about ϕ may become un-entrenched, making it possible to balance her mental state structure.
  - Proposition ψ is an element of $ind(D_sB_s)$ if $\psi \notin B_s$ is a partial cognitive precondition for $\psi \in B_s$, i.e. $\psi \notin B_s \ / \ \phi \in B_s$, and either Sarah desires to be ignorant about ϕ, or she believes that John desires a to be ignorant about ϕ, i.e. $(\phi \in D_s\widetilde{B_s}) \vee (\phi \in B_sD_j\widetilde{B_a})$. By deciding to believe ψ, Sarah's belief in ϕ may become un-entrenched, making it possible to balance her mental state structure.

Formally, we have:

$$\begin{aligned} ind(D_sB_s) \;=\; D_sB_s \;\cup\; \Big\{\psi \mid \mathcal{M}_s, \Delta_s \models\, & coherent(\psi \in B_s, \Delta_s) \wedge \\ & \Big(\Big((\psi \in B_s \;/\; \phi \in B_s) \wedge ((\phi \in D_sB_s) \vee (\phi \in B_sD_jB_a))\Big) \\ & \vee \Big((\psi \in B_s \;/\; \phi \notin B_s) \wedge ((\phi \in D_s\widetilde{B_s}) \vee (\phi \in B_sD_j\widetilde{B_a}))\Big) \\ & \vee \Big((\psi \notin B_s \;/\; \phi \notin B_s) \wedge ((\phi \in D_sB_s) \vee (\phi \in B_sD_jB_a))\Big) \\ & \vee \Big((\psi \notin B_s \;/\; \phi \in B_s) \wedge ((\phi \in D_s\widetilde{B_s}) \vee (\phi \in B_sD_j\widetilde{B_a}))\Big)\Big)\Big\} \end{aligned}$$

- The set $ind(D_s\widetilde{B_s})$ is the set of propositions that Sarah indirectly desires to be ignorant about. It equals her direct desires to be ignorant about, i.e. $D_s\widetilde{B_s}$, combined with the set of propositions ψ that are coherent with her desires, $coherent(\psi \notin B_s, \Delta_s)$, and that contains the following propositions.
  - Proposition ψ is an element of $ind(D_s\widetilde{B_s})$ if Sarah's ignorance about ψ is a partial cognitive precondition for her belief in ϕ, i.e. $\psi \notin B_s \;/\; \phi \in B_s$, and either Sarah desires to believe ϕ herself, or she believes that John desires a to believe ϕ, i.e. $(\phi \in D_sB_s) \vee (\phi \in B_sD_jB_a)$.
  - Proposition ψ is an element of $ind(D_s\widetilde{B_s})$ if $\psi \notin B_s$ is a partial cognitive precondition for Sarah to be ignorant about ϕ, i.e. $\psi \notin B_s \;/\; \phi \notin B_s$, and either Sarah desires to be ignorant about ϕ herself, or she believes that John desires a to be ignorant about ϕ, i.e. $(\phi \in D_s\widetilde{B_s}) \vee (\phi \in B_sD_j\widetilde{B_a})$.
  - Proposition ψ is an element of $ind(D_s\widetilde{B_s})$ if $\psi \in B_s$ is a partial cognitive precondition for $\phi \in B_s$, i.e. $\psi \in B_s \;/\; \phi \in B_s$, and either Sarah desires to believe ϕ herself, or she believes that John desires a to believe ϕ, i.e. $(\phi \in D_s\widetilde{B_s}) \vee (\phi \in B_sD_j\widetilde{B_a})$. By deciding to be ignorant about ψ, Sarah's belief in ϕ may become un-entrenched.
  - Proposition ψ is an element of $ind(D_s\widetilde{B_s})$ if $\psi \notin B_s$ is a partial cognitive precondition for $\psi \in B_s$, i.e. $\psi \in B_s \;/\; \phi \notin B_s$, and either Sarah desires to be ignorant about ϕ herself, or she believes that John desires a to be ignorant about ϕ, i.e. $(\phi \in D_sB_s) \vee (\phi \in B_sD_jB_a)$. By deciding to be ignorant about ψ, Sarah's ignorance about ϕ may become un-entrenched.

Formally, we have:

$$\begin{aligned} ind(D_s\widetilde{B_s}) \;=\; D_s\widetilde{B_s} \;\cup\; \Big\{\psi \mid \mathcal{M}_s, \Delta_s \models\, & coherent(\psi \notin B_s, \Delta_s) \wedge \\ & \Big(\Big((\psi \notin B_s \;/\; \phi \in B_s) \wedge ((\phi \in D_sB_s) \vee (\phi \in B_sD_jB_a))\Big) \\ & \vee \Big((\psi \notin B_s \;/\; \phi \notin B_s) \wedge ((\phi \in D_s\widetilde{B_s}) \vee (\phi \in B_sD_j\widetilde{B_a}))\Big) \\ & \vee \Big((\psi \in B_s \;/\; \phi \in B_s) \wedge ((\phi \in D_s\widetilde{B_s}) \vee (\phi \in B_sD_j\widetilde{B_a}))\Big) \\ & \vee \Big((\psi \in B_s \;/\; \phi \notin B_s) \wedge ((\phi \in D_sB_s) \vee (\phi \in B_sD_jB_a))\Big)\Big)\Big\} \end{aligned}$$

In subsequent sections and chapters, we will abstract from our design decisions how our agents can reach certain mental states. Instead of specifying which decisions an agent has to make to reach a certain mental state, e.g. to decide to believe a proposition, we will use indirect desires $ind(D_sB_s)$ and $ind(D_s\widetilde{B}_s)$ in the preconditions of decisions. The sets of indirect desires provide an abstraction over the precise decisions and desires that are involved to guide our agent's activities, such as decision-making and communication.

Analogous to Sarah's indirect desires about her own beliefs and ignorance, she will be said to have indirect desires regarding other agents. Indirect desires regarding other agents should not contradict Sarah's direct desires.

- The set $ind(D_sB_j)$ is the set of propositions that Sarah indirectly desires John to believe. It equals Sarah direct desires about John's beliefs, i.e. D_sB_j, combined with the set of propositions ψ that are coherent with Sarah's desires about John's beliefs, $coherent(\psi \in B_sB_j, \Delta_s)$, and that contains the following properties.
  - Proposition ψ is an element of $ind(D_sB_j)$ if Sarah believes that John's belief in ψ is a partial cognitive precondition for her belief in ϕ, i.e. $\psi \in B_sB_j \;/\; \phi \in B_s$, and Sarah has the indirect desire to believe ϕ, i.e. $\phi \in ind(D_sB_s)$.
  - Proposition ψ is an element of $ind(D_sB_j)$ if $\psi \in B_sB_j$ is a partial cognitive precondition for $\phi \in B_sB_j$, i.e. $\psi \in B_sB_j \;/\; \phi \in B_sB_j$, and Sarah desires John to believe ϕ, i.e. $\phi \in D_sB_j$.
  - Proposition ψ is an element of $ind(D_sB_j)$ if $\psi \in B_sB_j$ is a partial cognitive precondition for Sarah to be ignorant about ϕ, i.e. $\psi \in B_sB_j \;/\; \phi \notin B_s$, and Sarah indirectly desires to be ignorant about ϕ, i.e. $\phi \in ind(D_s\widetilde{B}_s)$.
  - Proposition ψ is an element of $ind(D_sB_j)$ if $\psi \in B_sB_j$ is a partial cognitive precondition for $\psi \in B_s\widetilde{B}_j$, i.e. $\psi \in B_sB_j \;/\; \phi \in B_s\widetilde{B}_j$, and Sarah desires John to be ignorant about ϕ, i.e. $\phi \in D_s\widetilde{B}_j$.

  Formally, we have:

$$
\begin{array}{rcl}
ind(D_sB_j) & = & D_sB_j \;\cup\; \Big\{\psi \mid \mathcal{M}_s, \Delta_s \models coherent(\psi \in B_sB_j, \Delta_s) \;\wedge \\
 & & \quad \Big(\big((\psi \in B_sB_j \;/\; \phi \in B_s) \wedge (\phi \in ind(D_sB_s))\big) \\
 & & \quad \vee \big((\psi \in B_sB_j \;/\; \phi \in B_sB_j) \wedge (\phi \in D_sB_j)\big) \\
 & & \quad \vee \big((\psi \in B_sB_j \;/\; \phi \notin B_s) \wedge (\phi \in ind(D_s\widetilde{B}_s))\big) \\
 & & \quad \vee \big((\psi \in B_sB_j \;/\; \phi \in B_s\widetilde{B}_j) \wedge (\phi \in D_s\widetilde{B}_j)\big)\Big)\Big\}
\end{array}
$$

- The set $ind(D_s\widetilde{B}_j)$ is the set of propositions that Sarah indirectly desires John to be ignorant about. It equals Sarah direct desires about John's ignorance, i.e. $D_s\widetilde{B}_j$, combined with the set of propositions ψ that are coherent with Sarah's desires about John's ignorance, $coherent(\psi \in B_s\widetilde{B}_j, \Delta_s)$, and that contains the following properties.

- Proposition ψ is an element of $ind(D_s\widetilde{B}_j)$ if Sarah believes that John's ignorance about ψ is a partial cognitive precondition for her belief in ϕ, i.e. $\psi \in B_s\widetilde{B}_j \;/\; \phi \in B_s$, and Sarah has the indirect desire to believe ϕ, i.e. $\phi \in ind(D_sB_s)$.
- Proposition ψ is an element of $ind(D_s\widetilde{B}_j)$ if $\psi \in B_s\widetilde{B}_j$ is a partial cognitive precondition for $\phi \in B_sB_j$, i.e. $\psi \in B_s\widetilde{B}_j \;/\; \phi \in B_sB_j$, and Sarah desires John to believe ϕ, i.e. $\phi \in D_sB_j$.
- Proposition ψ is an element of $ind(D_s\widetilde{B}_j)$ if $\psi \in B_s\widetilde{B}_j$ is a partial cognitive precondition for Sarah to be ignorant about ϕ, i.e. $\psi \in B_s\widetilde{B}_j \;/\; \phi \notin B_s$, and Sarah indirectly desires to be ignorant about ϕ, i.e. $\phi \in ind(D_s\widetilde{B}_s)$.
- Proposition ψ is an element of $ind(D_s\widetilde{B}_j)$ if $\psi \in B_s\widetilde{B}_j$ is a partial cognitive precondition for $\phi \in B_s\widetilde{B}_j$, i.e. $\psi \in B_s\widetilde{B}_j \;/\; \phi \in B_s\widetilde{B}_j$, and Sarah desires John to be ignorant about ϕ, i.e. $\phi \in D_s\widetilde{B}_j$.

Formally, we have:

$$
\begin{array}{rcl}
ind(D_s\widetilde{B}_j) & = & D_s\widetilde{B}_j \;\cup\; \Big\{\psi \,|\mathcal{M}_s, \Delta_s \models coherent(\psi \in B_s\widetilde{B}_j, \Delta_s) \wedge \\
 & & \quad \Big(\Big((\psi \in B_s\widetilde{B}_j \;/\; \phi \in B_s) \wedge (\phi \in ind(D_sB_s))\Big) \\
 & & \quad \vee \Big((\psi \in B_s\widetilde{B}_j \;/\; \phi \in B_sB_j) \wedge (\phi \in D_sB_j)\Big) \\
 & & \quad \vee \Big((\psi \in B_s\widetilde{B}_j \;/\; \phi \notin B_s) \wedge (\phi \in ind(D_s\widetilde{B}_s))\Big) \\
 & & \quad \vee \Big((\psi \in B_s\widetilde{B}_j \;/\; \phi \in B_s\widetilde{B}_j) \wedge (\phi \in D_s\widetilde{B}_j)\Big)\Big)\Big\}
\end{array}
$$

6.5.5 Desiring to believe

As said, our agents will use desires as motivational aspects such that they will make decisions with the intention to balance her mental state structure. Sarah's decision to conform to desired belief ψ is denoted $d2a_4(s, \psi, B_s)$.

Sarah will be justified to decide to believe ψ if the following preconditions hold. First, the preconditions of the decision to conform to John's beliefs, $d2a_3(s, \psi, B_s)$ from equation (6.11), hold in her mental state structure. That is, the set of preconditions $pre(d2a_3(s, \psi, B_s))$ is a subset of $pre(d2a_4(s, \psi, B_s))$. Second, Sarah will be justified to decide to believe ψ if she indirectly desires to believe ψ. That is, condition $\psi \in ind(D_sB_s)$ is an element of $pre(d2a_4(s, \psi, B_s))$.

$$
\begin{array}{rcl}
pre(d2a_3(s, \psi, B_s)) & \subset & pre(d2a_4(s, \psi, B_s)) \\
(\psi \in ind(D_sB_s)) & \in & pre(d2a_4(s, \psi, B_s))
\end{array}
\tag{6.19}
$$

Third, if Sarah believes that John believes that she is ignorant about ψ, i.e. $\psi \in B_sB_j\widetilde{B}_s$, then she believes that John does not desire her to be ignorant about ψ, i.e. $\psi \in B_s\widetilde{D}_j\widetilde{B}_s$. That is to say, a property of the set of preconditions is that if precondition $\psi \in B_sB_j\widetilde{B}_s$ holds in Sarah mental state structure, then precondition $\psi \in B_s\widetilde{D}_j\widetilde{B}_s$ holds also in her mental state structure. If this property holds, then Sarah is aware that John does not use her ignorance to justify his beliefs or ignorance. See Section 7.3.2 for the reasons for asking approval for decisions to add and retract beliefs.

$$\forall j \in \mathcal{A} \setminus s \Big((\psi \in B_s B_j \widetilde{B}_s) \;\in\; pre(d2a_4(s, \psi, B_s)) \;\Rightarrow \\ (\psi \in B_s \widetilde{D}_j \widetilde{B}_s) \;\in\; pre(d2a_4(s, \psi, B_s)) \Big) \quad (6.20)$$

6.5.6 Desiring to be ignorant

Sarah will be justified to decide to ignorant about ψ if the following preconditions hold. First, the preconditions of the decision to conform to John's ignorance, $pre(d2r_3(s, \psi, B_s))$ from equation (6.12), hold in her mental state structure. Second, Sarah will be justified to decide to be ignorant about ψ if she indirectly desires to be ignorant about ψ.

$$\begin{array}{rcl} pre(d2r_3(s, \psi, \utilde{B}_s)) & \subset & pre(d2r_4(s, \psi, B_s)) \\ (\psi \in ind(D_s B_s)) & \in & pre(d2r_4(s, \psi, B_s)) \end{array} \quad (6.21)$$

Third, if Sarah believes that John believes that she believes ψ, i.e. $\utilde{\psi} \in B_s B_j B_s$, then she believes that John does not desire her to believe ψ, i.e. $\psi \in B_s D_j B_s$. Thus, the set of preconditions has the property that if precondition $\psi \in B_s B_j B_s$ holds in Sarah mental state structure, then precondition $\psi \in B_s \widetilde{D}_j B_s$ holds also in her mental state structure. Informally, in addition to the other preconditions, Sarah is justified to retract her belief in ψ if she is not aware that another agent, say John, has used her ignorance to justify his beliefs or ignorance.

$$\forall j \in \mathcal{A} \setminus s \Big((\psi \in B_s B_j B_s) \;\in\; pre(d2a_4(s, \psi, \widetilde{B}_s)) \;\Rightarrow \\ (\psi \in B_s \widetilde{D}_j B_s) \;\in\; pre(d2a_4(s, \psi, \widetilde{B}_s)) \Big) \quad (6.22)$$

6.6 Concluding Remarks

In this chapter, we presented decision games that provide the meaning of decisions to add beliefs and decisions to retract beliefs. These decision games consist of decision rules that reflect conventional rules that are taken to be agreed upon by the community of agents. Our agents have not themselves agreed upon the criteria to make decisions correctly, but we assume that the experts who the agents represent have agreed on the criteria of the decisions. Because agents represent a domain of knowledge or a medical expert, and the meaning of believing and being ignorant, i.e. the use of the decision to believe and to be ignorant, have been agreed upon by the experts, the agents are not allowed to alter the meaning of decisions. It is the responsibility of the expert that the agents are equipped with adequate knowledge that either reflects their personal knowledge or that reflects conventional knowledge. Agents are thus not allowed to create knowledge.

Our agents can decide to believe and decide to be ignorant about propositions based on inference rules that the agents know. That is, agents use decision $d2a_1$ (eq. (6.5)) and decision $d2r_1$ (eq. (6.6)) in combination with their private knowledge of inference rules to change their mental states. Second, our agents may decide to believe and decide to be ignorant about propositions based on the meaning of the propositions that the agent knows. That is, the agents use decision $d2a_2$ (eqs. (6.7) and (6.9)) and decision $d2r_2$ (eqs. (6.8) and (6.10)) to change their beliefs, but also their beliefs about other agents' beliefs. Lastly, our agents may decide to believe and decide to be ignorant about propositions based on conformism and desires. That is, agents use decision $d2a_4$ (eq. (6.19)) and decision $d2r_4$ (eq. (6.21)) to ground their own beliefs on their beliefs about other agents' beliefs. In subsequent chapters, we assume that our agents have these decisions with the semantics as defined at their disposal. The decisions $d2a_3$ and $d2r_3$ to conform to other agent's beliefs will not be used directly; the preconditions of these decisions have been used in the decisions $d2a_4$ and $d2r_4$. The set of decision rulesΔ is the following set.

$$\begin{aligned}\Delta = \{&(d2a_1, pre(d2a_1), post(d2a_1)), (d2a_2, pre(d2a_2), post(d2a_2)),\\ &(d2a_4, pre(d2a_4), post(d2a_4)), (d2r_1, pre(d2r_1), post(d2r_1)),\\ &(d2r_2, pre(d2r_2), post(d2r_2)), (d2r_4, pre(d2r_4), post(d2r_4))\}\end{aligned}$$

The decision rules from Δ respect the coherence principles of the agent's mental state structure as provided in Section 5.2. That is to say, if the agent's mental state structure is coherent, application of the decision rules will retain a coherent mental state structure.

In our multiagent system, agents should not become explicitly justified to believe propositions that their responsible human experts should not also be justified to believe. However, as has been said before, being justified to believe something does not mean that it will be believed. The software agents will assist physicians in their judgements: either a doctor considers that a certain diagnosis, i.e. decision, can be made conform to the conventions of the medical community, and the agent may agree, or the agent may provide a diagnosis that the physician failed to consider.

Chapter 7

Dialogue Games to Communicate Beliefs and Desires

To allow our agents to communicate beliefs and desires, we will define speech acts that enable them to exchange information about their mental states. The later Wittgenstein described with his language games that the meaning of statements is determined by a community's agreement on the criteria for their correct use. We will describe, in a similar fashion, that the meaning of speech acts is also determined by a community's agreement on the criteria for their correct use. We will call the sets of rules that prescribe in which situations our agents are correct to use speech acts dialogue games. These dialogue games define the meaning of speech acts. Stated differently, information exchange between our agents about their mental states is the result of moves in a dialogue game that our agents make and interpret as speech acts.

First, in Section 7.1 we provide a background to dialogue games in which we will describe speech act theory, agent communication, and the relation with Wittgenstein's language games. In the subsequent five sections, we will define dialogue games that allow our agents to communicate beliefs and desires. Section 7.2 provides the meaning of speech acts in which agents ask others whether they may come to believe propositions. The diametrically opposed speech act in which agents ask others whether they may become ignorant about propositions will be described in Section 7.3. Section 7.4 provides the meaning of speech acts that allow agents to requests others to come to believe propositions, and in Section 7.5, we define speech acts that allow agents to request others to be ignorant about propositions. In Section 7.6, a dialogue game is provided in which agents inform others about their beliefs and desires that have changed due to decision processes. A special section is devoted to the situation in which an agent has run out of speech acts that may make her justified to decide to believe and decide to retract propositions (Section 7.7). We conclude in Section 7.8.

7.1 A Background to Dialogue Games

Before we present how our agents can convey information to each other, we provide a brief background to the dialogue games of subsequent sections. In order to communicate information, as we will discuss in Section 7.1.1, our agents' utterances will need to have certain properties as described by speech act theory. In Section 7.1.2, we turn to agent communication issues, and relate these to an agent's view of the meaning of, in this chapter, speech acts. In Section 7.1.3, we introduce dialogue games as sets of usage rules.

7.1.1 Communications in speech act theory

Philosophy of language provides the basis of communications in speech act theory (Austin [Aus62]), which is founded on the idea that with the utterance of words we can not only make statements (Searle [Sea69]), but also perform actions, and even change our social world (Habermas [Hab84]). Three main aspects of a speech act are identified: the *locution* refers to the lowest level of the speech act that only exhibits syntactic properties that are transmitted. The *illocutionary* act refers to the meaning, i.e. the use, of the speech act. The *perlocutionary* act refers to the changes that the locution induces on its receiver. Different illocutions can have an equal locution, making it troublesome for the receiver to interpret and reveal the illocution. The perlocution may depend on the receiver, because the receiver may interpret locutions differently.

Grice, in his article *Meaning* in 1957, drew a distinction between what he called natural meaning, and what he called non-natural meaning [Gri57, Gri89, chap. 14]. In general, natural meaning concerns some state of affairs, while non-natural meaning of words is related to a speaker's intentions in communicating something to a listener. Examples of natural meaning might be "those spots mean measles" and "that drop in barometric pressure means bad weather". Not all bearers of natural meaning are natural events or facts. In the example "high secretion levels of cortisol mean depression", the amount of secretion is used to describe behaviour of some brain tissue, and the word "depression" is used to describe the social and physiological behaviour of a person.

Grice offered the following analysis of a type of non-natural meaning called speaker's meaning or intentional meaning. The speaker Sarah means something non-naturally by uttering a locution if, and only if, she intends the locution to produce some effect in the receiver John, by means of his recognition of her intention. In *Utterer's meaning, sentence-meaning and word-meaning*, Grice offers the following account of the meaning of a speech act according to the receiver of the act [Gri68, Gri89, chap. 6]. Sarah's utterance of a locution that she believes proposition ψ, means for the receiver that she believes ψ if, and only if, according to the receiver John, Sarah uttered a locution such:

1. that John should believe that Sarah believes ψ,

2. that John should believe that Sarah intended 1, and
3. that 1 should be achieved by means of achieving 2.

A distinction can be made between the natural meaning of a locution, which we may take to be an agent's shared agreement on its general use, and what an agent means by the locutions *over and above* their natural meaning. Grice describes the latter with what he called conversational implicature [Gri75, Gri89, chap. 2]. He explains conversational implicature with the cooperative principle, which calls on a speaker to "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." [Gri75, Gri89, p. 26]. This cooperative principle is intended as a description of a speaker's normal behaviour in conversations—not as a prescriptive command. From the cooperative principle, the maxims of conversation are derived.

- *Maxim of quantity: information.* Make your contribution as informative as is required for the current purposes of the exchange, and do not make your contribution more informative than is required.
- *Maxim of quality: truth.* Do not say what you believe to be false, and do not say that for which you lack adequate evidence.
- *Maxim of relation: relevance.* Be relevant.
- *Maxim of manner: clarity.* Avoid obscurity of expression, avoid ambiguity, be brief (avoid unnecessary prolixity), and be orderly.

The maxims of conversation should be understood as describing the assumptions listeners normally make about the way speakers will communicate. Rather than moral prescriptions for how speakers ought to communicate, the maxims are presumptions that listeners rely on to interpret locutions. Research into this extra meaning of locutions over and above their conventional meaning is called the study of implicature. By anticipating the presumed adherence to the maxims of conversation, a speaker exploits this presumed adherence such that the receiver can retrieve the speaker's intentions even by using speech acts in other ways than their community has agreed upon. If a locution seems to flout the maxims of conversation, while nonetheless the receiver judges the speaker to obey the cooperative principle, the receiver has to look for other interpretations such that the locution does obey the maxims of conversation. Pragmatics is generally conceived as the study of understanding language and specifically as the study of how context influences the interpretation of meaning. The study of implicature is included in pragmatics, but not in the study of semantics.

Roughly, if a speaker Sarah thinks a listener John will interpret that she believes ψ through his recognition that she uses a certain locution with the intent to convey her belief ψ, then, by uttering the locution, the speaker conversationally implicates that she believes ψ (cf. Grice [Gri89, pp. 30–31]). For example, if Sarah says to John, "I am out of petrol", and John replies, "there is a garage round the corner", then John may be taken to have conversationally implicated that the garage is open and has gas to sell. "[T]he speaker implicates that which he must be assumed to believe in order to preserve the assumption that he is observing the maxim of relation." Grice [Gri89, p. 32]. Conveying information from John to Sarah is possible under the assumption that he did not flout the maxim of relation. It is only then that Sarah can interpret and retrieve the information that (John believes that) the garage is open and has gas to sell; otherwise John's response would have been irrelevant.

7.1.2 Agent communication

Similar to their human counterparts, software agents will communicate with the intent to transfer information to an intended receiver. As Weiss [Wei99, p. 361] observes, "Communications are a natural way in which the agents ... may interact with one another other than through incidental interactions through the environment." The structures that agents use to transfer information are locutions; the structure that identifies a locution with a certain content that originates from an uttering agent and is directed to an intended receiver, we call a communicative act.

The multiagent system community has been concentrating for some years on building standard interaction frameworks. Two examples of agent communication languages that both rely on speech act semantics are the Knowledge Query and Manipulation Language (KQML, Finin et al. [FFMM94]) and the standards body Foundation for Intelligent Physical Agents (FIPA) agent communication language (FIPA-ACL, [FIP02]).

KQML is a language and protocol for exchanging information and knowledge. Programs use KQML to communicate attitudes about information, such as querying, stating, believing, requiring and subscribing. It is both a message format and a message-handling protocol that supports run-time knowledge sharing among agents.

FIPA-ACL provides speech act semantics, expressed semi-formally using epistemic modal logic. FIPA defined utterances in ACL in terms of the beliefs, desires and intentions of the speaker. The ACL intends to enable agents using FIPA-ACL to be certain that other agents will understand the meaning of utterances in the same way as the speaker, cf. Labrou [Lab01].

In this mentalistic approach, a receiver should be able to verify that a speaker's criterion for making an utterance is equal to the receiver's criterion for interpreting the utterance. This ability to verify is questionable, and as a result, the mentalistic approach has been criticised (Moulin [Mou97], Singh [Sin03, Sin98]). This critique is based on what is called the semantic verification problem [Woo00]: an agent cannot verify that other agents act according to held dialogues. This problem is analogous to the verification problem for truths in Correspondence theory: because propositions' truth-conditions are facts in reality, and agents may not have access to reality, agents may have no method of knowing the truth (cf. Section 2.2.1). By definition, agents do not have access to the cognitive state of other agents, and only have indirect access to these states through communication. If the truth-conditions whether Sarah is correct to predicate that John believes a proposition are part of John's cognitive state, then Sarah has no method of knowing whether John believes the proposition. That is to say, with the mentalistic approach a receiver can never verify the speaker's mentalistic criteria of an utterance.

The FIPA-ACL specification [FIP02] has been provided a formal semantics for utterances expressed in BDI logic (Rao and Georgeff [RG98]). This specification assumes that agents do have access to the truth-conditions for predicating whether agents have, for example, beliefs and desires. With such mentalistic semantics, it has been difficult to provide a basis for inter-operable agents between different platforms. A related problem with agents from different platforms and the verification problem is the sincerity assumption, which is considered too restrictive to be of use in, for example, electronic business (see Dignum [Dig00]).

In the same line of thought of the dictum 'meaning is use' (Section 2.3.1), we take it that the members of a community have agreed upon the criteria for the correct use of communicative acts. If an agent regards herself entitled to know that she uses a communicative act in accordance with the use that her community has agreed upon, then the agent is justified to believe that she can use the communicative act to transfer information to a receiver of the act. Our agents may not themselves have agreed upon what the criteria of the communicative act, but we take it that the human counterparts the agents represent have. The speaker is justified to believe that she can transfer information because, according to the speaker, the intended receiver will be aware of the shared use of the communicative act.

The dialogue rules that we will define in subsequent sections implement that an agent regards herself entitled, according to her community, to know a conventional rule that describes the situations in which she is explicitly justified to utter a certain communicative act. Similar to a decision rule, as described in Section 6.1.2, a dialogue rule provides a set of preconditions and a set of post-conditions. The criterion of the conventional rule is the set of preconditions that define in which situations an agent is explicitly justified to utter a communicative act. The post-conditions describe the properties that hold after uttering the act. Thus, if the preconditions of a dialogue rule have been met for an agent, then the agent may utter the communicative act associated with that rule.

As said, the set of preconditions of a dialogue rule reflect the criterion of a conventional rule. That is to say, an agent is entitled to know that the use of these preconditions is shared, and that this use is common knowledge in her community. The preconditions specify properties of an uttering agent's mental state structure that, according to the agent, her community has agreed upon to be the circumstances in which a communicative act is used correctly. Stated differently, the preconditions for the correct use of a locution provide the meaning of the locution, i.e. the illocution of the speech act. The post-conditions of the dialogue rule provide the properties that hold in the agent's mental state structure after the utterance of the locution, i.e. the perlocution of the speech act. If an agent regards herself entitled, according to her community, to know that the use of an illocution is shared, then she knows that all agents in her community use the same criterion for that illocution. Thus, because the speaker may assume that an illocution's use is common knowledge, she may derive that the properties that have to hold if she were to utter a communicative act must hold for all speakers of that act. Thus, if an agent receives a locution that she interprets to be a certain illocution, then she can derive properties of the speaker's mental states. These properties are described in the post-conditions of the communicative act. If Sarah utters a locution directed at John, John can deduce properties of Sarah's mental states that have held at the time she uttered the act. Because John knows the situations in which he is justified to utter the locution, he knows that the same preconditions have held for Sarah. Thus, agents can come to believe properties of their communication partners by uttering and receiving communicative acts.

Agents may be justified to predicate *to believe* propositions, but fail to believe them, simply because they have not had the opportunity to decide to believe the propositions and change their belief state accordingly (see Section 2.3.1). A similar argument holds for communicative acts. Agents may be explicitly justified to utter communicative acts, but fail to utter them because other activities, such as decision-making, have taken precedence. Stated differently, the dialogue rules do not provide a strategy when to communicate, or what will be communicated, the dialogue rules define in which situations agents are explicitly justified to communicate.

7.1.3 Dialogue games

In computer science, and especially in the development of multiagent systems, communication between autonomous agent systems has been used to allow agents to provide arguments (see Parsons et al. [PWA03] and Reed [Ree98]) to convince other agents to perform actions to reach private and collective goals. The dialogue typology by Walton and Krabbe [WK95] identifies different categories of dialogues by specifying the agent's initial situations and goals.

1. Persuasion dialogues are dialogues in which a speaker seeks to convince a listener to accept to believe a particular proposition. McBurney and Parsons [MP01a] provide a dialectical argumentation framework for qualitative representation of epistemic uncertainty in scientific domains. Prakken [Pra00] provides a formal framework for argumentative dialogue systems with the possibility of counterarguments. Dialogue games in Law have similar objectives, for example, Prakken and Sartor [PS98, PS96] discuss disputes in which participants have conflicting arguments.

2. Negotiation dialogues are dialogues in which participants seek to agree on how to divide a resource (see McBurney et al. [MvEPA03] and Sadri et al. [STT01]).

3. Deliberation dialogues are dialogues in which participants make plans by discussing which actions to perform in which situations (see Hitchcock et al.

[HMP01]) or deliberate which uncertain belief should be accepted (see McBurney and Parsons [MP01b]). These dialogues result in collective intentions (see Dignum et al. [DDKV01]) or group plans (see Dignum et al. [DDKV00]).

4. Information seeking dialogues are dialogues in which a speaker seeks to find a truth value of a proposition by asking a listener who may have the answer (see Hulstijn [Hul00] and Beun [Beu01]). Inquiries are information seeking dialogues in which the participants co-operate to answer a question for which the participants have no individual answer.

The dialogue games that we will define in subsequent sections will allow agents to persuade others to add or retract beliefs. Agents will be persuaded to change their mental states when the preconditions of decision rules have been met. Thus, by providing other agents with information through communication, the following dialogues contribute to the category of persuasion dialogues. Our dialogue games allow our agents to seek for information. Specifically, our agents seek information that may make them explicitly justified to make decisions that will balance their mental state structure. The current work also contributed to the category of information seeking dialogues.

We use unbalanced mental state structures to provide agents with motives to communicate. If Sarah desires to believe a proposition ψ that she does not believe, and, at the same time, her decision rules have not made her come to believe ψ, then she is motivated to communicate with the intent to balance her mental state structure. The agent's deliberation cycle, as described in Section 5.3.4, defines when agent may make decisions and engage in conversation. It is only if Sarah's mental state structure is closed under decision-making that she may utter communicative acts. Consequently, if Sarah's mental state structure is closed under decision-making, and her mental state structure remains unbalanced, then her decision rules could not balance her mental states.

Just as decision games are sets of rules that define when agents are justified to make decisions (see Section 6.1.3), dialogue games are finite sets of dialogue rules that define in which situations agents are justified to utter communicative acts. Unlike the game of chess, a dialogue game does not define the goal of winning. In a dialogue game, an agent's goal is to balance her mental state structure. Just as in the game of chess, participating agents in a dialogue game make moves and take turns. They make moves with the aim to reach a state in which their mental state structures are balanced.

The communication language that our agents use consists of communicative acts which are structures with the following syntax: $\lambda(s, j, \psi)$ where λ is a locution, s is the name for the agent uttering the locution, j is the name of the intended receiver, and ψ is the content of the locution. The communication language is defined as the set of all possible communicative acts between agents in the multiagent system. In subsequent sections, we will present five dialogue games in which each dialogue game provides the semantics of the communicative acts.

Definition 7.1 (communication language $\mathcal{L}_C$). *Given a $\mathcal{L}_{\mathcal{B},\mathcal{F}}$, a set of agent names $\mathcal{A}$, the* communication language *$\mathcal{L}_C$ consists of the following communicative acts. If $s, j \in \mathcal{A}$, $s \neq j$, and $\psi \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$, then:*

1. *Section 7.2 provides the dialogue game with questions for belief additions: $qba_i(s, j, \psi)$, $gqba_i(s, j, \psi)$, $dqba_i(s, j, \psi) \in \mathcal{L}_C$, with $i = 1$ or $i = 2$.*
2. *Section 7.3 provides the dialogue game with questions for belief retractions: $qbr_i(s, j, \psi)$, $gqbr_i(s, j, \psi)$, $dqbr_i(s, j, \psi) \in \mathcal{L}_C$, with $i = 1$ or $i = 2$.*
3. *Section 7.4 provides the dialogue game with requests for belief additions: $rba(s, j, \psi)$, $grba_i(s, j, \psi)$, $drba_i(s, j, \psi) \in \mathcal{L}_C$, with $i = 1$ or $i = 2$.*
4. *Section 7.5 provides the dialogue game with request for belief retractions: $rbr(s, j, \psi)$, $grbr_i(s, j, \psi)$, $drbr_i(s, j, \psi) \in \mathcal{L}_C$, with $i = 1$ or $i = 2$.*
5. *Section 7.6 provides the dialogue game with inform statements about belief and desire changes: $is(s, j, \psi, ms) \in \mathcal{L}_C$, with $ms \in MSN_s$.*

Definition 7.2 (dialogue rule). *Given the languages of mental state $\mathcal{L}'_{MS} \subseteq \mathcal{L}_{MS}$, and the language of communication $\mathcal{L}_C$, a* dialogue rule *for a communicative act $\mu \in \mathcal{L}_C$ is a structure $(\mu, pre(\mu), post(\mu))$ with $pre(\mu) \in \mathcal{L}_{MS}$ the set of preconditions of μ, and $post(\mu) \subseteq \mathcal{L}'_{MS}$ the set of post-conditions of μ.*

7.1.4 The activity of communicating

Sarah is allowed to utter a communicative act $\lambda(s, j, \psi)$ if the preconditions of $\lambda(s, j, \psi)$ hold her mental state structure, i.e. $\mathcal{M}_s \models pre(\lambda(s, j, \psi))$. Directly after Sarah has uttered μ, her mental state structure will have been updated with the post-conditions of μ, i.e. $\mathcal{M}'_s \models post(\lambda(s, j, \psi))$. We will not define the actual activity of updating an agent's mental state structure such that the post-conditions of a communicative act will become to hold. We only provide the preconditions and post-conditions. The mental state structure of a receiving agent John need not be updated directly. This is because another agent Fred may also have directed a communicative act to John, and John is still processing Fred's communicative (cf. the agent's deliberation cycle in Section 5.3.4).

For the same reason that the activity of making a decision (Section 6.1.4) should to change the agent's mental state structure, the activity of uttering a communicative act should change the agent's mental state structure. If a speaker utters a communicative act, then the act should be informative according to the speaker of the act. If an utterance of a communicative act $\lambda(s, j, \psi)$ does not change the speaker's mental state structure, then whether or not the speaker utters the act does not make a difference.[1]

[1] Under the assumption of an ideal communication channel and an ideal cognitive state that does not 'forget' unintentionally, every uttered communicative act will arrive at the intended receiver. Consequently, once a speaker has uttered a communicative act directed at some receiver, the second time the speaker utters the same act directed at that receiver, the receiver cannot retrieve new information from that

If the utterance of $\lambda(s, j, \psi)$ has not made a difference from a speaker's perspective, then the act is not informative. To enforce that communicative acts are informative, the negation of the speaker's post-conditions of the acts are taken to be part of the preconditions.

Assumption 7.1. *For all dialogue rules, the negation of the speaker's post-condition of a communicative act is part of the precondition of the communicative act (cf. assumption 6.1).*

Example 7.1 (dialogue rule). *If set $\{(\psi \in B_s B_j B_s), (\psi \in B_j B_s)\} \subseteq \mathcal{L}_{MS}$ is the post-condition of communicative act $\lambda(s, j, \psi)$, then with assumption 7.1 we have $\neg(\psi \in B_s B_j B_s) \in pre(\lambda(s, j, \psi))$. Note that $\neg(\psi \in B_j B_s)$ is not part of the precondition, this because it is not part of the speaker's post-condition. If we extend the preconditions of $\lambda(s, j, \psi)$ exclusively with $(\psi \in B_s) \in \mathcal{L}_{MS}$, then we have dialogue rule: $(\lambda(s, j, \psi), (\psi \in B_s) \wedge \neg(\psi \in B_s B_j B_s), \{(\psi \in B_s B_j B_s), (\psi \in B_j B_s)\})$. Directly after Sarah has uttered $\lambda(s, j, \psi)$, her part of the post-condition holds in her mental state structure, i.e. $\mathcal{M}_s \models (\psi \in B_s B_j B_s)$. Note that $\mathcal{M}_s \not\models \lambda(s, j, \psi)$. Directly after John has processed $\lambda(s, j, \psi)$, his part of the post-condition holds in his mental state structure, i.e. $\mathcal{M}_j \models (\psi \in B_j B_s)$.*

7.2 Questions for a Belief Addition

We will present a dialogue game that provides the rules of usage for agents to pose and answer questions whether they may decide to be believe propositions. This set of dialogue rules is the dialogue game with questions for a belief addition. First, we provide an intuitive reading of the game's communicative acts (Section 7.2.1) and an example dialogue (Section 7.2.2), after which we define the communicative act in which agents ask whether they may decide to believe propositions (Section 7.2.3). A receiving agent must interpret these communicative acts (Section 7.2.4), and update her mental states accordingly, after which she may respond. An agent may respond with a positive answer (Section 7.2.5) or with a negative answer (Section 7.2.6). The effects of these communications are that a chain of justifications is created distributed over the mental state structures of the dialogue participants (Section 7.2.7). This section is partly based on published work by Lebbink et al. [LWM03a, LWM04c, LWM04a, LWM04b].

7.2.1 Intuitive readings

With an utterance of a question for a belief addition, or *qba* for short, agents ask for information that they may need for deciding to believe propositions. With a *qba*, an agent communicates her desire to believe a certain proposition with the intent to receive information such that she may become justified to believe the proposition. It is with respect to the normative agreement of the speaker and receiver's community on

act. If the assumption of an ideal communication channel would be waived, uttering a communicative act more than once can make sense: a speaker would repeat her communication if she assumes that it did not arrive at the intended receiver.

the meaning of the predicate *to believe* a proposition, that the speaker asks permission to believe a proposition. In this question, an agent does not ask for permission from the intended receiver to believe a proposition, but asks for information that, according to their community, provides a justification for believing the proposition. A similar, but different utterance is when an agent asks for information whether she may decide, according to the desires of the intended receiver, to believe a proposition. An agent may need to check whether her decisions to believe propositions will agree with other agents' desires. On the surface, this latter question may seem equal to the first. "May I believe the proposition that p?" or similarly, "Does p?" in which p reads "this fifteen-amp fuse fits in our fuse box" asked by Sarah can be interpreted as that she desires to believe that this is physically possible according to the manufacturing specifications of the box and the fuse. A different question is whether the fifteen-amp agrees with John's desire to use an electric circular saw needing more current. To distinguish between the two different illocutions, qba_1 is the questions asked against the background of the communities' agreed upon meaning of the predicate *to believe*. The illocution, in which an agent asks whether her belief agrees with the intended receiver, is denoted by qba_2. Agents can ask both questions, it will be up to the receiver to interpret the question, and answer accordingly.

An agent may utter a locution that is intended to be a response to a qba. A receiver may interpret this locution as the illocution of granting a question for a belief addition, denoted $gqba$. In a $gqba_1$, a speaker provides a receiver with a testimony that she judges that a specific proposition may be believed according to the proposition's meaning. In a $gqba_2$, a speaker provides a receiver with a testimony that the receiver may believe a specific proposition according to the speaker's desires. An agent may also utter a locution that is interpreted as denying a question for a belief addition, denoted $dqba$. The illocution of a $dqba_1$ is uttered with the intent to convey that the speaker has insufficient information whether the receiver may decide to believe the proposition. The utterance does not connote that the receiver may, according to their community, be ignorant about the proposition; this will be dealt with in Section 7.3 on questions for belief retraction. In a $dqba_2$, the speaker intends to convey that the receiver learns that a proposition does not agree with the speaker's desires. If Sarah asks whether she may believe that she can use a fifteen-amp fuse, i.e. a qba_1, and whether he agrees she will use a fifteen-amp fuse, i.e. a qba_2, John could answer her with "yes, you may believe that you could", i.e. $gqba_1$, "but I believe I need a bigger one", i.e. $dqba_2$.

In short, the communicative acts have the following reading.

- $qba_1(s, j, \psi)$ reads "May I according to the use of ψ believe ψ?"
- $qba_2(s, j, \psi)$ reads "May I according to your desires believe ψ?"
- $gqba_1(s, j, \psi)$ reads "According to the use of ψ, you may believe ψ."
- $gqba_2(s, j, \psi)$ reads "According to my desires, you may believe ψ."
- $dqba_1(s, j, \psi)$ reads "I do not have justifications for you to believe ψ."

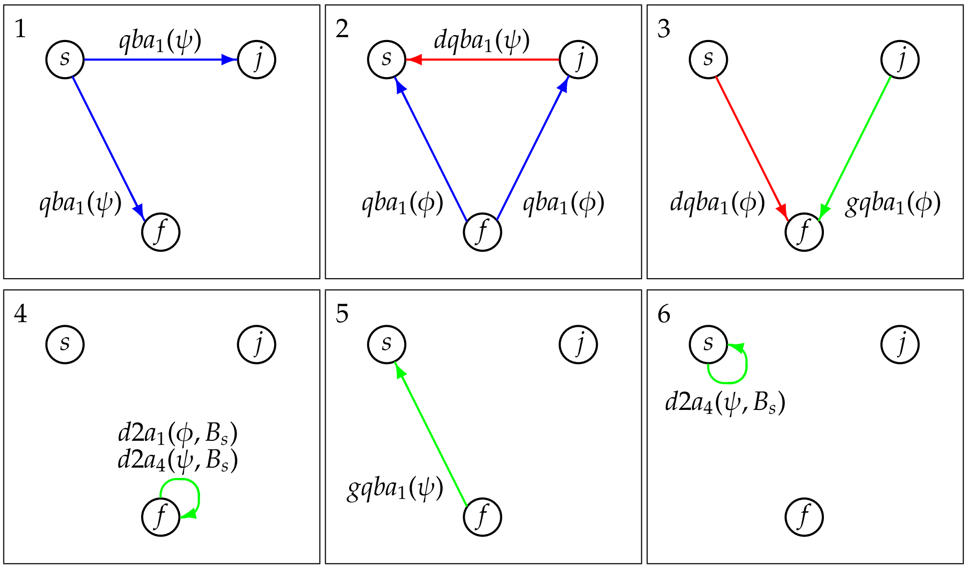

Figure 7.1: Dialogues about questions for belief addition (example 7.2).

- $dqba_2(s, j, \psi)$ reads "According to my desires, you may not believe ψ."

Sarah is motivated to decide to believe ψ, as described in Section 6.5.1, if she has an indirect desire to believe ψ and she is ignorant about ψ, that is, if she has an unbalanced mental state structure, i.e. $\mathcal{M}_s \models (\psi \in ind(D_s B_s)), (\psi \notin B_s)$. If Sarah has such an unbalanced mental state structure and the decision rules that enable her to change her beliefs have not balanced it, then Sarah is motivated to utter questions with the intention of getting answers in which she will be provided with information such that she becomes explicitly justified to believe ψ. Thus, if Sarah's mental state structure has been closed under decision-making, she desires to believe ψ, and does not believe ψ, then she is motivated to ask other agents whether she may come to believe propositions such that her decision rules will balance her mental state structure.

7.2.2 An example dialogue

We will present an example dialogue between Sarah, John and Fred, who consult each other about a patient's situation.

Example 7.2. *Assume proposition ψ reads as "it is true that the patient is suffering from disease X", and ϕ reads as "it is true that the patient shows symptom Y" and Fred knows the inference rule from example 5.1. See figure 7.1 for a depiction.*

1. *Assume Sarah indirectly desires to believe ψ, i.e. $\psi \in ind(D_sB_s)$, and she does not believe ψ, i.e. $\psi \notin B_s$, then she has an unbalanced mental state structure regarding ψ. Sarah asks Fred and John whether she may believe ψ.*

2. - *Assume John does not believe ψ. John answers negatively with "I do not have information from which you would be allowed to predicate to believe that the patient is suffering from X".*
   - *Assume that Fred knows an inference rule from which he can derive that if he would believe ϕ, he would be entitled to believe ψ, i.e. $(\phi \in B_f) \rightarrowtail (\psi \in B_f) \in \mathcal{K}_f$. Assume also that Fred does not believe ϕ, i.e. $\phi \notin B_f$. Thus, due to this inference rule and his belief that Sarah desires to believe ψ, he has an indirect desire to believe ϕ, i.e. $\phi \in ind(D_fB_f)$. He asks both Sarah and John whether he may believe ϕ.*

3. - *Assume Sarah does not believe ϕ, i.e. $\phi \notin B_s$. Sarah answers negatively.*
   - *Assume John believes ϕ, i.e. $\phi \in B_j$. John answers positively.*

4. *Fred decides to believe ϕ with decision $d2a_1$ on the grounds of his belief that John believes ϕ. Fred then decides to believe ψ with decision $d2a_1$, based on his belief in ϕ.*

5. *Fred answers Sarah's question regarding ψ positively.*

6. *Sarah decides to believe ψ with decision $d2a_1$ grounded on Fred's belief in ψ, and thus balancing her mental state structure regarding her desire to believe ψ.*

7.2.3 Posing questions for belief addition

If Sarah has an unbalanced mental state structure regarding some proposition ψ and she is explicitly justified to believe ψ, then she is motivated to decide to believe ψ. If Sarah's mental state structure is closed under decision-making and she still has an unbalanced mental state structure, then the preconditions to decide to believe ψ have not been met in her mental state structure. That is to say, Sarah does not believe that John believes ψ which would make her justified to believe ψ. If Sarah does not believe that John believes ψ, then she is justified to ask whether he believes ψ. Thus, if Sarah is motivated to decide to believe ψ, but the decision ($d2a_4$ from equation (6.19)) to conform to John's belief is not allowed because Sarah is ignorant whether John believes ψ, then a qba_1 may resolve Sarah's ignorance about whether John believes ψ, and possibly balance her mental state structure.

If Sarah indirectly desires to believe ψ and she does not believe ψ, i.e. $\psi \in ind(D_sB_s)$ and $\psi \notin B_s$, then she is motivated to utter a *qba*. An indirect desire (Section 6.5.4) restricts the use of a *qba* to those situations in which the speaker's belief in ψ balances her mental state structure. The communicative acts qba_1 and qba_2 will be derived from the meaning of this *qba*. The common precondition of all *qba* is that the speaker has an unbalanced mental state structure.

$$(\psi \in ind(D_sB_s)), (\psi \notin B_s) \ \in \ pre(qba(s, j, \psi)) \tag{7.1}$$

In addition to the common precondition, Sarah is justified to pose a qba_1 regarding ψ, if she does not believe that John believes ψ, and she does not believe that John is ignorant about ψ, i.e. $\psi \notin B_sB_j$ and $\psi \notin B_s\widetilde{B}_j$. If these preconditions hold, Sarah is justified to direct a qba_1 to John. This part of the meaning is replaced by different preconditions to provide alternative meanings, i.e. usage, of the communicative act qba_1. The preconditions of qba are a proper subset of the preconditions of qba_1, i.e. $pre(qba(s, j, \psi)) \subset pre(qba_1(s, j, \psi))$.

$$\begin{array}{rcl} pre(qba(s, j, \psi)) & \subset & pre(qba_1(s, j, \psi)) \\ (\psi \notin B_sB_j), (\psi \notin B_s\widetilde{B}_j) & \in & pre(qba_1(s, j, \psi)) \end{array} \tag{7.2}$$

Sarah may pose a qba_2 regarding ψ if she does not believe that John does not desire her to be ignorant about ψ, i.e. $\psi \notin B_s\widetilde{D}_j\widetilde{B}_s$. That is, according to Sarah, it is possible that John desires her to be ignorant about ψ. If this is the case, she should ask him whether her belief in ψ would agree with his desires. However, Sarah need not check with agents that according to her beliefs do not believe that she is ignorant about ψ, that is, Sarah only has to check with those agents that according to her beliefs believe that she is ignorant about ψ, i.e. $\psi \in B_sB_j\widetilde{B}_s$. Sarah has to check this, because, as described in Section 6.5.5, the preconditions for Sarah to decide to believe ψ, i.e. $d2a(s, \psi, B_s)$, is that if Sarah believes that John believes that she is ignorant about ψ, i.e. $\psi \in B_sB_j\widetilde{B}_s$, then she should believe that John does not desire her to be ignorant about ψ, i.e. $\psi \in B_s\widetilde{D}_j\widetilde{B}_s$. As we will describe in Section 7.2.5 in equation 7.11, the precondition $\psi \in B_s\widetilde{D}_j\widetilde{B}_s$ will follow as a post-condition of communication. The agents who believe that Sarah is ignorant about ψ could have used their beliefs about her ignorance to justify their decisions; Sarah only has to check with these agents whether her adoption of a belief interferes with their desires.

$$\begin{array}{rcl} pre(qba(s, j, \psi)) & \subset & pre(qba_2(s, j, \psi)) \\ (\psi \in B_sB_j\widetilde{B}_s), (\psi \notin B_s\widetilde{D}_j\widetilde{B}_s) & \in & pre(qba_2(s, j, \psi)) \end{array} \tag{7.3}$$

Given the common preconditions to utter a qba from equation (7.1), the receiver can derive properties of the speaker's mental state structure. The receiver John may derive that speaker Sarah desires to believe ψ, i.e. $\psi \in B_jD_sB_s$, and that she does not believe ψ, i.e. $\psi \in B_j\widetilde{B}_s$, that is, John is aware that Sarah has an unbalanced mental state structure regarding ψ. Additionally, John may become to believe that Sarah believes that he believes that she is ignorant about ψ, i.e. $\psi \in B_jB_sB_j\widetilde{B}_s$. As described in Section 6.4.2, $\psi \in B_jB_sB_j\widetilde{B}_s$ is a precondition for John to decide to retract his belief in ψ. If John interprets a received locution as a qba_1, the following additional properties of Sarah's mental states may be derived. John may derive that Sarah does not believe that he believes ψ, and that she believes that he does not believe ψ; that is, John may believe that Sarah has no clue whether he believes ψ, i.e. $\psi \in B_j\widetilde{B}_sB_j$ and $\psi \in B_j\widetilde{B}_s\widetilde{B}_j$. If John interprets a locutions from Sarah as a qba_2, then John is justified to believe that she does not believe that he desires her to be ignorant about ψ. John's mental state structure should change according to the following post-conditions.

$$\begin{array}{r} (\psi \in B_j\widetilde{B_s}B_j), (\psi \in B_j\widetilde{B_s}\widetilde{B_j}), (\psi \in B_jD_sB_s), \\ (\psi \in B_j\widetilde{B_s}), (\psi \in B_jB_sB_j\widetilde{B_s}) \end{array} \in \; post(qba_1(s, j, \psi)) \tag{7.4}$$

$$\begin{array}{r} (\psi \in B_j\widetilde{B_s}\widetilde{D_j}\widetilde{B_s}), (\psi \in B_jD_sB_s), \\ (\psi \in B_j\widetilde{B_s}), (\psi \in B_jB_sB_j\widetilde{B_s}) \end{array} \in \; post(qba_2(s, j, \psi)) \tag{7.5}$$

Because the meaning of the communicative acts is taken to be agreed upon by the human experts that the agents represent, a speaker of a communicative act is justified to believe what properties the listener will derive after receiving and interpreting the communicative act. After uttering a $qba(s, j, \psi)$, Sarah is justified to believe that John believes that she desires to believe ψ, i.e. $\psi \in B_sB_jD_sB_s$. Additionally, after a $qba_1(s, j, \psi)$, Sarah is justified to believe that John believes that she is ignorant about ψ. Sarah's mental state structure has changed according to the following post-conditions.

$$(\psi \in B_sB_j\widetilde{B_s}), (\psi \in B_sB_jD_sB_s) \; \in \; post(qba_1(s, j, \psi)) \tag{7.6}$$

$$(\psi \in B_sB_jD_sB_s) \; \in \; post(qba_2(s, j, \psi)) \tag{7.7}$$

7.2.4 Interpreting communicative acts

As said, an agent is explicitly justified to utter a communicative act if she regards herself entitled, according to her community, to use the communicative act. Illocutions have readings, which we use to provide intuitive meanings of communicative acts; in contrast, the set of preconditions that reflect a community's agreement on the correct use of the illocutions provide the agents with the conventional meaning. It is the agent's awareness of the agreement on the correct use of illocutions that agents use to justify their utterances. More important, because an agent regards herself entitled to know that the use of a communicative act is shared in her community, she can use the act to convey information. Because a speaker can assume her utterance to be recognised by the receiver to be in accord with their community's agreement, a speaker is justified to believe that the receiver will interpret the speaker's communicative act as she has intended. This anticipated recognition makes the speaker justified to believe that she conveys information to a receiver. For example, the $qba_1(s, j, \psi)$ with reading "May I according to the use of ψ believe ψ?" reflects Sarah's assumption that John recognises that, in accord with their communities' agreement on the use of the illocution, she desires to believe ψ. Thus, according to Sarah's view of the use of the $qba_1(s, j, \psi)$, she has to desire to believe ψ; because she regards that the use is shared, she is justified to believe that John can derive her desire to believe ψ. Additionally, Sarah intends to convey that she is not justified, conform to the conventions of her community, to believe ψ, and because she assumes John assumes she communicates in compliance with the maxims of conversation, she thinks her utterance will be interpreted (by John) to intend to be responded such that she may become justified to believe ψ.

Communicative acts can be interpreted to belong to different types, such as assertives, commissives, directives, declaratives (see Searle [Sea79]). In a commissive, a speaker commits herself to an action; in a directive, a speaker tries to get the hearer to carry out a certain action. We like to distinguish our communicative acts similarly. In a question, the speaker tries to make the receiver utter a speech act that will provide the speaker with certain information. In an answer, an agent commits herself to the information that she intends to convey. A useful interpretation of our communicative acts is to form question-answer pairs. A communicative act is called an answer in response to another communicative act, which is called a question, if the answer conveys information that is requested in the question.[2]

Based on the agreed upon meaning of the two communicative acts, an agent may interpret that an act is an answer in response to another act. Sarah may interpret that her uttering $\lambda_1(s, j, \psi)$ is in response to $\lambda_2(j, s, \phi)$ if, according to her view of the use of the two locutions (i.e. provided by her dialogue rules) the preconditions of $\lambda_1(s, j, \psi)$ are a subset of the post-conditions of the $\lambda_2(j, s, \phi)$. According to Sarah, John will also interpret that $\lambda_1(s, j, \psi)$ is an answer to $\lambda_2(j, s, \phi)$; because the use of the communicative acts is shared.

If an agent is to use a communicative act to exchange information with another agent, she has to be entitled to know that the intended receiver of the communicative act uses the same meaning (i.e. the agreed upon use) of the act. By definition of meaning, an agent assumes that other agents of her community use the same preconditions and post-conditions for communicative acts. In principle, a speaker in a dialogue can know when the intended receiver would interpret the speaker's communicative act as an answer in response to some communicative act. A formal analysis of agents who can become aware of question-answer pairs in their communication is beyond the scope of this dissertation. In subsequent sections, communicative acts are grouped by the speaker's interpretation whether the acts are positive or negative answers in response to some previous communicative act.

7.2.5 Positive answers to questions for belief addition

If Sarah has interpreted a received locution as an illocution $qba_1(j, s, \psi)$, and she has updated her mental states according to the locution's post-conditions, then she may either answer negatively with a $dqba_1(s, j, \psi)$, which we will discuss in Section 7.2.6, or positively with a $gqba_1(s, j, \psi)$, which we discuss here.

Sarah is aware that John has an unbalanced mental state structure if she believes that John desires to believe ψ, and she believes that he is ignorant about ψ, i.e. $\psi \in B_s D_j B_j$ and $\psi \in B_s \widetilde{B}_j$. A similar, but stronger expression is when Sarah believes that John desires to believe ψ, and she does not believe that he believes ψ, i.e. $\psi \in B_s D_j B_j$ and $\psi \notin B_s B_j$. With the coherence relation expressed in the contraposition of equation (5.3), we have $\mathcal{M}_s \models \psi \in B_s \widetilde{B}_j \Rightarrow \mathcal{M}_s \models \psi \notin B_s B_j$. Instead of requiring that Sarah believes that John is ignorant about ψ, i.e. $\psi \in B_s \widetilde{B}_j$, Sarah has to be ignorant

[2] Whether a communicative act conveys information from a speaker to a receiver depends on the agent who judges whether the requested information is exchanged.

about whether John believes ψ, i.e. $\psi \notin B_sB_j$. In the latter situation, Sarah can balance John's mental state structure even if Sarah is not explicitly aware that John has an unbalanced mental state structure.

The set of preconditions to utter a $gqba_1(s,j,\psi)$ is that Sarah is aware that John has an unbalanced mental state structure regarding ψ, and she judges herself entitled to predicate *to believe* ψ (see eq. (7.8)). Sarah may utter a $gqba_2(s,j,\psi)$, if she is aware that John has an unbalanced mental state structure regarding ψ, and she does not desire that John is ignorant about ψ, i.e. $\psi \notin ind(D_s\widetilde{B}_j)$.

$$(\psi \in B_s), (\psi \in B_sD_jB_j), (\psi \notin B_sB_j) \;\in\; pre(gqba_1(s,j,\psi)) \tag{7.8}$$

$$(\psi \notin ind(D_s\widetilde{B}_j)), (\psi \in B_sD_jB_j), (\psi \notin B_sB_j) \;\in\; pre(gqba_2(s,j,\psi)) \tag{7.9}$$

Given the preconditions to utter a *gqba*, a receiver may derive properties of the speaker's mental states. If a receiver John interprets a locution as a $gqba_1(s,j,\psi)$, the perlocution is that he may deduce that Sarah is aware that he has an unbalanced mental state structure regarding ψ, i.e. $\psi \in B_sB_jD_sB_s$ and $\psi \in B_j\widetilde{B}_sB_j$. If John interprets that a locution has been uttered with Sarah's intent to be in response to an earlier uttered $qba_1(j,s,\psi)$, i.e. interpreted as an $gqba_1(s,j,\psi)$, then John has already updated his mental states to reflect that Sarah is aware of his unbalanced mental state structure regarding ψ. More precisely, if John interprets a received locution as a $gqba_1(s,j,\psi)$, then he may also believe that Sarah believes ψ, and he believes that she believes this, i.e. $\psi \in B_jB_s$ and $\psi \in B_jB_sB_jB_s$ (eq. (7.10)). If John interprets a received locutions as a $gqba_2(s,j,\psi)$, then he may believe that Sarah believes that he does not desire her to be ignorant about ψ.

$$(\psi \in B_jB_s), (\psi \in B_jB_sB_jB_s), (\psi \in B_jB_sD_jB_j), (\psi \in B_j\widetilde{B}_sB_j) \;\in\; post(gqba_1(s,j,\psi)) \tag{7.10}$$

$$(\psi \in B_j\widetilde{D}_s\widetilde{B}_j), (\psi \in B_jB_sD_jB_j), (\psi \in B_j\widetilde{B}_sB_j) \;\in\; post(gqba_2(s,j,\psi)) \tag{7.11}$$

Based on the listener's post-condition of a $gqba_1$, the post-condition of the $gqba_1$ is that the speaker Sarah is justified to believe that John believes that she believes ψ. After uttering a $gqba_2$, speaker Sarah is justified to believe that John believes that she does not desire John to be ignorant about ψ.

$$(\psi \in B_sB_jB_s) \;\in\; post(gqba_1(s,j,\psi)) \tag{7.12}$$

$$(\psi \in B_sB_j\widetilde{D}_s\widetilde{B}_j) \;\in\; post(gqba_2(s,j,\psi)) \tag{7.13}$$

7.2.6 Negative answers to questions for belief addition

If Sarah has an unbalanced mental state structure regarding ψ, and her decision rules cannot balance it by making her believe ψ, then, as described in Section 7.2.3, she may utter *qba*s with the intent to be answered such that she becomes explicitly justified to believe ψ. If has Sarah run out of communicative acts that, if answered, may change her mental state structure such that she becomes explicitly justified to believe ψ, then she may answer John's $qba_1(j,s,\psi)$ that she cannot provide him with justifications to believe ψ. Sarah has run out of communicative acts to become justified to be believe ψ if her mental state structure enjoys the property *finished*$(s, \psi \in B_s)$ as expressed in equation (7.79) in Section 7.7. Thus, if John utters a $qba_1(j, s, \psi)$ and Sarah has no untried communicative acts to come to believe ψ herself, then she may utter a $dqba_1(s, j, \psi)$.

Sarah is justified to utter a $dqba_1(s, j, \psi)$ if she is aware that John has an unbalanced mental state structure regarding ψ and she does not believe ψ. Additionally, Sarah has run out of communicative acts to become explicitly justified to believe ψ; this holds if $pre(dqba_1(s, j, \psi))$ is closed under condition *finished*$(s, \psi \in B_s)$. Sarah is justified to utter a $dqba_2(s, j, \psi)$ if she is aware that John has an unbalanced mental state structure regarding ψ, and she does have an indirect desire that John is ignorant about ψ.

$$(\psi \notin B_s), (\psi \in B_s D_j B_j), (\psi \notin B_s B_j) \ \in \ pre(dqba_1(s, j, \psi)) \tag{7.14}$$

$$(\psi \in ind(D_s \widetilde{B}_j)), (\psi \in B_s D_j B_j), (\psi \notin B_s B_j) \ \in \ pre(dqba_2(s, j, \psi)) \tag{7.15}$$

Listener John can derive properties of Sarah's mental states from the preconditions of a $dqba_1(s, j, \psi)$. John may believe that Sarah does not believe ψ and that he believes that she believes this, i.e. $\psi \in B_j \widetilde{B}_s$ and $\psi \in B_j B_s B_j \widetilde{B}_s$. Additionally, just as with the post-conditions of positive answers (eqs. (7.10) and (7.11)), John may become aware that Sarah is aware that John has an unbalanced mental state structure regarding ψ, i.e. $\psi \in B_s B_j D_s B_s$ and $\psi \in B_j \widetilde{B}_s B_j$. From a $dqba_2(s, j, \psi)$, listener John may derive that Sarah desires him to be ignorant about ψ.

$$(\psi \in B_j \widetilde{B}_s), (\psi \in B_j B_s B_j \widetilde{B}_s), (\psi \in B_j B_s D_j B_j), (\psi \in B_j \widetilde{B}_s B_j) \ \in \ post(dqba_1(s, j, \psi)) \tag{7.16}$$

$$(\psi \in B_j D_s \widetilde{B}_j), (\psi \in B_j B_s D_j B_j), (\psi \in B_j \widetilde{B}_s B_j) \ \in \ post(dqba_2(s, j, \psi)) \tag{7.17}$$

Speaker Sarah may derive properties of John's mental state structure. The post-condition of a $dqba_1$ is that Sarah believes that John believes that she is ignorant about ψ, i.e. $\psi \in B_s B_j \widetilde{B}_s$. The post-condition of a $dqba_2$ is that Sarah believes that John believes that she desires him to be ignorant about ψ, i.e. $\psi \in B_s B_j D_s \widetilde{B}_j$.

$$(\psi \in B_s B_j \widetilde{B}_s) \ \in \ post(dqba_1(s, j, \psi)) \tag{7.18}$$

$$(\psi \in B_s B_j D_s \widetilde{B}_j) \ \in \ post(dqba_2(s, j, \psi)) \tag{7.19}$$

7.2.7 Distributed chains of justifications

Sarah is explicitly justified to believe a proposition if, first, she regards herself entitled to know what a criterion is to predicate *to believe* the proposition, and second, this criterion has been met in her mental state structure (Section 2.3.1). We next discuss the situations in which Sarah's beliefs are entrenched in her mental state structure. Similar to the entrenchment of Sarah's belief is the entrenchment of Sarah's ignorance; without loss of generality, we will only discuss the situation in which Sarah's belief is entrenched. Sarah's belief in ϕ or her ignorance about ϕ can be a reason that she is explicitly justified to believe ψ, i.e. $\phi \in B_s / \psi \in B_s$ or $\phi \notin B_s / \psi \in B_s$. If we were to ask

Sarah why she believes ψ, she could answer that she believes ϕ or is ignorant about ϕ respectively. Intuitively, these answers provide a chain of justification.

If Sarah's belief in ϕ_1 is a criterion for her to believe ϕ_2, and her belief in ϕ_2 is a criterion for her to believe ϕ_3, then her belief in ϕ_1 provides her with an explicit justification to believe ϕ_2, and her belief in ϕ_2 provides her with an explicit justification to believe ϕ_3. The criteria create a chain of justification linking Sarah's belief in ϕ_3 with her belief in ϕ_1. Stated differently, the meaning of believing ϕ_2 and ϕ_3 describe a chain of justification from belief in ϕ_1, through her belief in ϕ_2, to her belief in ϕ_3. Sarah's belief in ϕ_1 provides her with an *indirect*, explicit justification to believe ϕ_3. If we were to ask Sarah why she believes ϕ_3, then she could answer that she believes ϕ_2. If we are somehow not satisfied with her response, Sarah could also answer that she believes ϕ_1, because her belief in ϕ_1 indirectly justifies her belief in ϕ_3. Sarah's justification to believe ϕ_1 propagates to her justification to believe ϕ_2, which propagates in her mental state structure to her justification to believe ϕ_3.

Another situation in which Sarah's belief in ψ is entrenched in her mental state structure is when her belief in ψ is justified by her belief that John believes ϕ_1, i.e. $\phi_1 \in B_sB_j/\psi \in B_s$, or her belief in ψ is justified by her belief that John is ignorant about ϕ, i.e. $\phi_1 \in B_s\widetilde{B}_j/\psi \in B_s$. If we were to ask Sarah why she believes ψ, she could answer that she believes that John believes ϕ_1, or that she believes that John is ignorant about ϕ_1, respectively. On his turn, John may justify his belief in ϕ_1 with his belief in a particular ϕ_2. In general, John's justification to believe ϕ_1 is inaccessible to Sarah. Consequently, Sarah can only answer the question why she believes ψ by referring to her belief that John believes ϕ_1. If we are not satisfied with Sarah's answer, Sarah can respond that her part of the chain of justifications stops here, and for further justification, we should ask John. This chain of justifications is distributed over the mental state structures of Sarah and John.

With the reading of the word 'correctly' under Correspondence theory of truth (Section 2.2.1), if Sarah has justified her belief in ψ with her belief that John believes ϕ, then her belief in ψ is correctly justified if John in fact justified to believe ϕ. Thus, if Sarah believes that John believes ϕ, while John does not believe ϕ, then Sarah can be said to be falsely justified to believe ψ. However, under a Wittgensteinian account of truth, if Sarah's belief that John believes ψ is not accompanied by John's belief in ϕ, she is still explicitly justified to believe ψ. This because it is the criterion that she does believe that John believes ϕ that defines when she is correct—according to her community—to believe ψ. Whether Sarah's belief that John believes ϕ is accompanied by John's belief in ϕ is a different matter. Thus, under an account of truth, according to Correspondence theory, a chain of justifications can span different mental state structures, while under a Wittgensteinian account of truth, a chain of justifications stops at the boundaries of a mental state structure. The truth is thus confined to an agent's cognitive state. This observation tallies with Wittgenstein's observation in *On Certainty* "To be sure there is justification; but justification comes to an end." [Wit72, #192] and "Knowledge is in the end based on acknowledgement." [Wit72, #378]. Wittgenstein's account of truth allows our agents correctly to predicate *to know* and *to believe* propositions based on their bounded cognitive states, while from

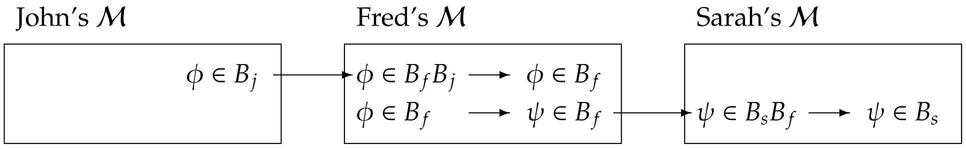
John's $\mathcal{M}$
Fred's $\mathcal{M}$
Sarah's $\mathcal{M}$
$\phi \in B_j$
$\phi \in B_f B_j \longrightarrow \phi \in B_f$
$\phi \in B_f \longrightarrow \psi \in B_f$
$\psi \in B_s B_f \longrightarrow \psi \in B_s$

Figure 7.2: A distributed chain of justification.

the agent's perspective, a chain of justifications appears to span over mental state structures of other agents.

The communicative acts qba_1, $gqba_1$ and $dqba_1$, just as any other communicative act that we will introduce in the following sections, will provide agents with explicit justifications to have beliefs about other agents' beliefs, desires, lacks of belief, and lacks of desire. The rules of the dialogue games provide Sarah with conditions in which situations she may come to believe properties of other agents. If Sarah interprets John's utterance as being a certain illocution, she is consequently explicitly justified to have beliefs about John's mental states. The beliefs of our agents will have a chain of justifications that ends in the beliefs that the agent has about the beliefs of the expert she represents, or in the belief about another agent's beliefs. If we were to ask Sarah why she believes a proposition, her answer would ultimately consist of beliefs about other agents in the multiagent system, or beliefs that are derived from the agent's knowledge base. If we were to follow the chain of justifications that makes Sarah explicitly justified to believe a certain proposition, a question would be whether such chains end in the agents' knowledge bases. Stated differently, the question would be whether the beliefs of agents are grounded in knowledge that the human experts have provided them. The answer depends on the agents' knowledge, that is, their inference rules and epistemology, this because this knowledge may construct cycles in chains of justifications. It is conceivable that Sarah's belief in ψ has a chain of justifications that ends in her belief that John believes ϕ, and John's belief in ϕ has a chain of justifications that ends in his belief that Sarah believes ψ. Because an agent's justification is, in general, inaccessible to other agents, Sarah and John may not be aware of the cycle in their chains of justifications. We will not elaborate on the implications of cyclic chains and leave unanswered whether such cycles are inimical to entrust agents with beliefs.

For example, the dialogue in example 7.2, figure 7.1, creates a distributed chain of justification as depicted in figure 7.2. Under the assumption that John's belief in ϕ originates directly from an expert he represents, and who provides John with justification to believe ϕ, Fred's belief that John believes ϕ provides Fred with an explicit justification to believe ϕ. Fred justifies his belief in ψ on his belief in ϕ. Thus, John's justification to believe ϕ propagates to Fred's justification to believe ϕ and ψ. Sarah's belief that Fred believes ψ provides Sarah with an explicit justification to believe ψ. Thus John's justification to believe ψ propagates through John's believe in ψ to Sarah's belief in ψ. The chain of justification is distributed over John's, Fred's and Sarah's cognitive states.

7.3 Questions for a Belief Retraction

This section will deal with a dialogue game that provides the rules of usage for agents to pose and answer questions whether they may decide to retract beliefs. This section is partly based on published work by Lebbink et al. [LWM05].

7.3.1 Intuitive readings

With the utterance of a question for a belief retraction, or *qbr* for short, agents ask for information which they may need to decide to be ignorant about propositions. Similar to a *qba* (Section 7.2), with a *qbr*, agents seek information such that they will be justified to perform the activity of deciding to retract a belief. With a qbr_1, an agent asks whether she is entitled, according to her community, to decide to be ignorant about a proposition. With a qbr_2, an agent asks whether the desires of the intended receiver tally with her desires to be ignorant about a proposition. An agent may utter a locution that a receiver may interpret as the illocution of granting a question for belief retraction, denoted *gqbr*. With a $gqbr_1$, the speaker provides the receiver with her testimony that she judges that a specific proposition, according to their community, need not be believed. With a $gqbr_2$, the speaker acknowledges that according to her desires the receiver may be ignorant about a proposition. An agent may also utter a locution that is interpreted as denying a question for a belief retraction, *dqba* for short. The $dqbr_1$ is the illocution with the connotation that the speaker has insufficient information whether the receiver may decide to be ignorant about the proposition. The utterance does not connote that the receiver may believe, according to their community, the proposition. (This can be achieved with a qba_1 as described in Section 7.2.) The $dqbr_2$ conveys that the speaker does not grant an agent's belief retraction because the speaker desires the agent to believe the proposition.

In short, the communicative acts have the following reading.

- $qbr_1(s, j, \psi)$ reads "May I according to the use of ψ be ignorant about ψ?"
- $qbr_2(s, j, \psi)$ reads "May I according to your desires be ignorant about ψ?"
- $gqbr_1(s, j, \psi)$ reads "According to the use of ψ, you may be ignorant about ψ."
- $gqbr_2(s, j, \psi)$ reads "According to my desires, you may be ignorant about ψ."
- $dqbr_1(s, j, \psi)$ reads "I do not have justifications for you to be ignorant about ψ."
- $dqbr_2(s, j, \psi)$ reads "According to my desires, you may not be ignorant about ψ."

Sarah is motivated to decide to be ignorant about a proposition ψ, as described in Section 6.5.1, if she has an indirect desire not to believe ψ and she believes ψ, that is, if she has an unbalanced mental state structure, i.e. $\mathcal{M}_s \models (\psi \in ind(D_s\widetilde{B_s})), (\psi \in B_s)$. If

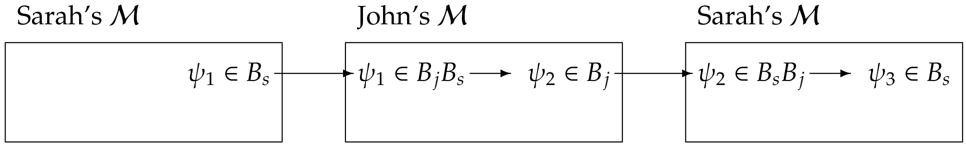

Figure 7.3: Sarah's belief in ψ_3 is indirectly justified on her belief in ψ_1.

Sarah has such an unbalanced state, and the decisions that can change her beliefs have not balanced her mental state structure, then Sarah is motivated to utter questions with the intent of providing her with information such that the decisions that will balance her mental states will become justified. If Sarah's mental state structure is closed under decision-making, and she desires to be ignorant about ψ, and she does believe ψ, then she is motivated to ask other agents whether she may become ignorant about propositions such that her decision rules will balance her mental states. This she can achieve with a qbr_1.

7.3.2 Reasons for asking approval

One might wonder why agents should ask approval from other agents to retract their beliefs even if they are, according to their community, entitled to be ignorant about propositions. A reason for co-operative agents would be that their beliefs and ignorance have to be co-ordinated in order to achieve collaborative plans. For selfish agents, dialogues in which agents ask approval for belief retraction have the following use. If Sarah believes that John believes that she believes ψ_1, i.e. $\psi_1 \in B_sB_jB_s$, and Sarah were to decide to be ignorant about ψ_1, then she may render John's beliefs, as seen from her perspective, unjustified.[3] This is because John may have used Sarah's belief in ψ_1 to ground his own belief in ψ_2. See figure 7.3 for a depiction of the following chain of justification. If Sarah has grounded her belief in ψ_3 on John's belief in ψ_2, she may render her own belief in ψ_3 unjustified if she does not consult John about her intended decision to retract her belief in ψ_1. Because Sarah is not aware of the chain of justification linking her belief in ψ_1 to her belief in ψ_3, she has a selfish reason to ask John whether he agrees with her intended decision to retract her belief in ψ_1. Coordination of this decision she can achieve with a qbr_2.

7.3.3 An example dialogue

Example 7.3. *Assume proposition ψ reads as "it is true that the patient is suffering from disease X", and ϕ reads as "it is true that the patient shows symptom Y" and Fred knows the*

[3] From Sarah's perspective, John's beliefs may be unjustified; however, from John's perspective, as described in Section 7.2.7, he may regard himself justified to have his beliefs. Stated differently, the distributed chain of justification, as seen from Sarah's perspective, may not exist, while John is not aware that the chain does not exist.

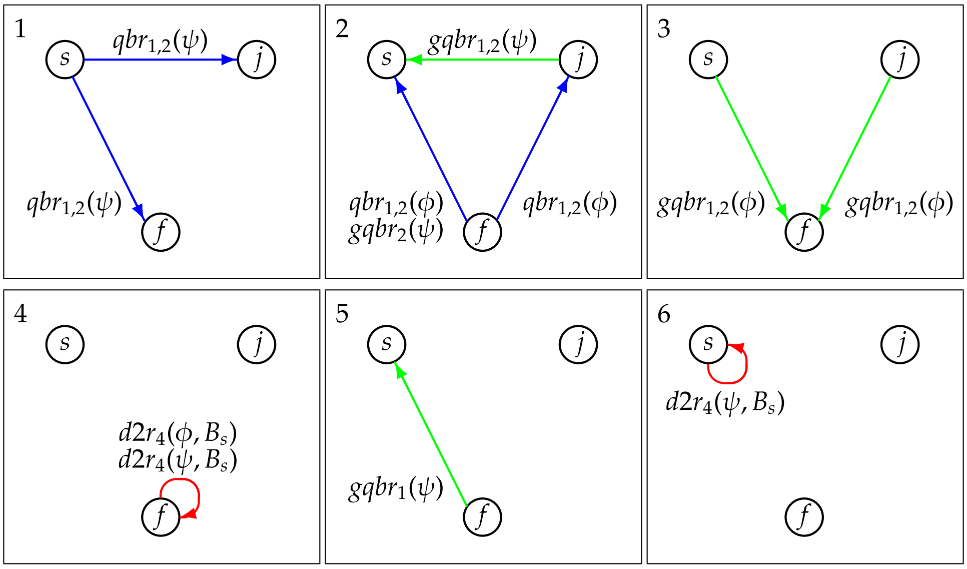

Figure 7.4: Dialogues about belief retraction with a sub dialogue (example 7.3).

inference rule from example 5.1. Additionally, assume that in previous conversation Sarah told John and Fred that she believes ψ. However, new lab results indicate that this diagnosis should be refuted. Because Sarah believes that John and Fred believe that she believes ψ, she first has to retract her previous diagnosis before she can make a new one. However, before she can retract the previous diagnosis, she asks John and Fred whether she can. See figure 7.4 for a depiction.

1. *Assume Sarah indirectly desires to be ignorant about ψ, i.e. $\psi \in ind(D_s\widetilde{B_s})$, and she believes ψ, i.e. $\psi \in B_s$, then Sarah has an unbalanced mental state structure regarding ψ. Sarah asks Fred and John whether she may retract her belief in ψ. Sarah asks Fred and John whether she is correct, according to their use of the predicate* to believe*, to retract belief in ψ, i.e. Sarah utters a $qbr_1(s, j, \psi)$ and a $qbr_1(s, f, \psi)$. Additionally, assume Sarah is believes that John and Fred believe she believe ψ, i.e. $\psi \in B_sB_jB_s$ and $\psi \in B_sB_fB_s$, the John and Fred may have used her belief in ψ to ground their own beliefs, she checks with them, i.e. Sarah utters a $qbr_2(s, j, \psi)$, and a $qbr_2(s, f, \psi)$.*

2. - *Assume John does not believe ψ, i.e. $\psi \notin B_j$, John utters $gqbr_1(j, s, \psi)$. Assume John's desires do not disagree with Sarah's proposed retraction, i.e. $\psi \notin ind(D_jB_s)$. John utters $gqbr_2(j, s, \psi)$.*
   - *Assume Fred's desires do not disagree with Sarah's proposed retraction, i.e. $\psi \notin ind(D_fB_s)$. Fred utters $gqbr_2(f, s, \psi)$.*
   - *Assume Fred believes ψ, i.e. $\psi \in B_f$, and knows inference rule $(\phi \in B_f) \rightarrowtail (\psi \in B_f) \in \mathcal{K}_f$, then the presence of his belief in ψ is cognitively entrenched in his*

belief in ϕ, i.e. $\mathcal{M}_f, \Delta \models \psi \in^{\in} B_f$. Consequently, Fred cannot agree that Sarah retracts her belief in ψ. Additionally, assume Fred believes that both John and Sarah believe that he believes ϕ, i.e. $\phi \in B_f B_j B_f$ and $\phi \in B_f B_s B_f$). Fred asks Sarah and John whether he may retract his belief in ϕ: Fred utters a $qbr_{1,2}(f, s\phi)$ and a $qbr_{1,2}(j, s\phi)$.

3. *Assume both Sarah and John do not believe ϕ, i.e. and $\phi \ni nB_s$ and $\phi \notin B_j$. Additionally, Assume both Sarah's and John's desires do not disagree with Fred's proposed retraction. Both Sarah and John respond that Fred may retract his belief in ϕ.*

4. *Fred retracts his belief in ϕ with decision $d2r_4$. As a consequence, his belief in ψ is not entrenched anymore, and he retracts ψ with decision $d2r_4$.*

5. *Fred responds to Sarah that she may retract her belief in ψ: he utters a $gqbr_1(\psi)$ to her.*

6. *Sarah may retract her belief in ψ with decision $d2r_4$.*

7.3.4 Posing questions for belief retraction

If Sarah has an unbalanced mental state structure regarding ψ and deciding to be ignorant about ψ will balance her mental state structure, then she is motivated to ask John whether she may be ignorant about ψ. John's answer could make Sarah explicitly justified to decide to retract her belief in ψ (with decision rule from equation (6.21)), and balance her mental state structure.

Sarah is justified to utter a $qbr(s, j, \psi)$ if she has an unbalanced mental state structure regarding ψ, that is, she indirectly desires to be ignorant about ψ which she believes, i.e. $\psi \in ind(D_s\widetilde{B}_s)$ and $\psi \in B_s$.

$$(\psi \in ind(D_s\widetilde{B}_s)), (\psi \in B_s) \ \in \ pre(qbr(s, j, \psi)) \tag{7.20}$$

Additionally, for a qbr_1, equal to the preconditions of a qba_1 (eq. (7.2)), Sarah should not believe that John believes ψ and she does not believe that John is ignorant about ψ, i.e. $\psi \notin B_sB_j$ and $\psi \notin B_s\widetilde{B}_j$.

$$\begin{aligned} pre(qbr(s, j, \psi)) \ &\subset \ pre(qbr_1(s, j, \psi)) \\ (\psi \notin B_sB_j), (\psi \notin B_s\widetilde{B}_j) \ &\in \ pre(qbr_1(s, j, \psi)) \end{aligned} \tag{7.21}$$

A different use of a qbr is to confer with agents whether they agree with the proposed belief retraction. With a $qbr_2(s, j, \psi)$, Sarah asks John whether he agrees with her proposed retraction to believe ψ. Sarah could ask all agents in the multiagent system for approval to become ignorant about ψ; however, she only needs to address a qbr_2 to those agents she believes could have used her belief in ψ to justify their own beliefs (cf. Section 6.4.1). If Sarah believes that John believes that she believes ψ, i.e. $\psi \in B_sB_jB_s$, and she does not believe that John does not desire her to believe ψ, i.e. $\psi \notin B_s\widetilde{D}_jB_s$, then she may ask whether she may retract her belief in ψ.

$$\begin{aligned} pre(qbr(s, j, \psi)) \ &\subset \ pre(qbr_2(s, j, \psi)) \\ (\psi \in B_sB_jB_s), (\psi \notin B_s\widetilde{D}_jB_s) \ &\in \ pre(qbr_2(s, j, \psi)) \end{aligned} \tag{7.22}$$

After receiving a locution and interpreting it to be a $qbr_1(s, j, \psi)$, the perlocution for John may be to conclude that Sarah has an unbalanced mental state structure regarding ψ, i.e. $\psi \in B_jD_s\widetilde{B}_s$ and $\psi \in B_jB_s$. Additionally, John is justified to believe that Sarah neither believes that he believes ψ nor that he is ignorant about ψ, i.e. $\psi \in B_j\widetilde{B}_sB_j$ and $\psi \in B_j\widetilde{B}_s\widetilde{B}_j$. After interpreting a locution to be a $qbr_2(s, j, \psi)$, in addition to being aware of the unbalanced mental state structure, John is justified to believe that Sarah is ignorant whether he does not desire her to believe ψ, i.e. $\psi \in B_j\widetilde{B}_s\widetilde{D}_jB_s$. John's mental state structure should change according to the following post-conditions.

$$\begin{matrix}(\psi \in B_j\widetilde{B}_sB_j), (\psi \in B_j\widetilde{B}_s\widetilde{B}_j), (\psi \in B_jD_s\widetilde{B}_s), \\ (\psi \in B_jB_s), (\psi \in B_jB_sB_jB_s)\end{matrix} \in post(qbr_1(s, j, \psi)) \tag{7.23}$$

$$\begin{matrix}(\psi \in B_j\widetilde{B}_s\widetilde{D}_jB_s), (\psi \in B_jD_s\widetilde{B}_s), \\ (\psi \in B_jB_s), (\psi \in B_jB_sB_jB_s)\end{matrix} \in post(qbr_2(s, j, \psi)) \tag{7.24}$$

After uttering a $qbr(s, j, \psi)$, Sarah is justified to believe that John believes that she desires to be ignorant about ψ, i.e. $\psi \in B_sB_jD_s\widetilde{B}_s$. Additionally, after a $qbr_2(s, j, \psi)$, Sarah is justified to believe that John believes that she believes ψ. Sarah's mental state structure has changed according to the following post-conditions.

$$(\psi \in B_sB_jB_s), (\psi \in B_sB_jD_s\widetilde{B}_s) \in post(qbr_1(s, j, \psi)) \tag{7.25}$$

$$(\psi \in B_sB_jD_s\widetilde{B}_s) \in post(qbr_2(s, j, \psi)) \tag{7.26}$$

7.3.5 Positive answers to questions for belief retraction

If Sarah has interpreted a locution as a $qbr(j, s, \psi)$, and she has updated her mental states accordingly, then she may either answer negatively with a $dqbr(s, j, \psi)$, which we will discuss in Section 7.3.6, or positively with a $gqbr(s, j, \psi)$, which we discuss now.

Sarah is justified to utter a $gqbr_1(j, s, \psi)$ if she does not believe ψ, and she is aware that John has an unbalanced mental state structure regarding ψ. Sarah may utter a $gqbr_2(j, s, \psi)$ if she believes that John desires to be ignorant about ψ, and she does not desire that John believes ψ.

$$(\psi \notin B_s), (\psi \in B_sD_j\widetilde{B}_j), (\psi \notin B_s\widetilde{B}_j) \in pre(gqbr_1(s, j, \psi)) \tag{7.27}$$

$$(\psi \notin ind(D_sB_j)), (\psi \in B_sD_j\widetilde{B}_j), (\psi \notin B_s\widetilde{B}_j) \in pre(gqbr_2(s, j, \psi)) \tag{7.28}$$

After the utterance of a $gqba_1(s, j, \psi)$, John is justified to believe that Sarah does not believe ψ, that he believes that she believes this, that Sarah believes that he desires to be ignorant about ψ, and that he believes that Sarah is ignorant whether he is ignorant about ψ. After the utterance of a $gqba_2(s, j, \psi)$, John is justified to believe that Sarah does not desire him to believe ψ.

$$(\psi \in B_j\widetilde{B}_s), (\psi \in B_jB_sB_j\widetilde{B}_s), (\psi \in B_jB_sD_j\widetilde{B}_j), (\psi \in B_j\widetilde{B}_s\widetilde{B}_j) \in post(gqbr_1(s, j, \psi)) \tag{7.29}$$

$$(\psi \in B_j\widetilde{D}_sB_j), (\psi \in B_jB_sD_j\widetilde{B}_j), (\psi \in B_j\widetilde{B}_s\widetilde{B}_j) \in post(gqbr_2(s, j, \psi)) \tag{7.30}$$

Similar post-conditions hold for the speaker. After uttering a $gqbr_1(s, j, \psi)$, Sarah is justified to believe that John believes that he is ignorant about ψ. After a $gqbr_2(s, j, \psi)$, Sarah is justified to believe that John believes that she does not desire him to believe ψ.

$$(\psi \in B_s B_j \widetilde{B}_s) \;\in\; post(gqbr_1(s, j, \psi)) \tag{7.31}$$
$$(\psi \in B_s B_j \widetilde{D}_s B_j) \;\in\; post(gqbr_2(s, j, \psi)) \tag{7.32}$$

7.3.6 Negative answers to questions for belief retraction

Similar to the negative answers to a $qba_1(j, s, \psi)$, a $qbr_1(j, s, \psi)$ can be answered negatively if Sarah has run out of methods to become justified to be ignorant about ψ. That is to say, Sarah has run out of communicative acts that may make her justified to change her mental state structure to balance it.

Sarah is justified to utter a negative response to a $qbr_1(j, s, \psi)$ if the following preconditions hold. Sarah is aware that John has an unbalanced mental state structure regarding ψ, i.e. $\psi \in B_s D_j \widetilde{B}_j$ and $\psi \notin B_s \widetilde{B}_j$, Sarah believes ψ and all her communicative acts to become justified to predicate *to be ignorant* about ψ have not made her decide to be ignorant about ψ. That is to say, all communicative acts that could make her justified to retract her belief in ψ have been tried. Sarah has run out of communicative acts to become justified to be ignorant about ψ if her mental state structure enjoys the property $\mathit{finished}(s, \psi \notin B_s)$ as expressed in equation (7.79) in Section 7.7.

$$(\psi \in B_s), (\psi \in B_s D_j \widetilde{B}_j), (\psi \notin B_s \widetilde{B}_j) \;\in\; pre(dqbr_1(s, j, \psi)) \tag{7.33}$$
$$(\psi \in ind(D_s B_j)), (\psi \in B_s D_j \widetilde{B}_j), (\psi \notin B_s \widetilde{B}_j) \;\in\; pre(dqbr_2(s, j, \psi)) \tag{7.34}$$

Receiver John may derive properties of Sarah's mental states from the preconditions of a $dqbr_1(s, j, \psi)$. From a $dqbr_1(s, j, \psi)$, John is justified to believe that Sarah is aware that he has an unbalanced mental state structure regarding ψ, i.e. $\psi \in B_j B_s D_j \widetilde{B}_j$ and $\psi \in B_j \widetilde{B}_s \widetilde{B}_j$, and John is justified to believe that Sarah believes ψ. From a $dqbr_2(s, j, \psi)$, John is justified to believe that Sarah desires him to believe ψ.

$$(\psi \in B_j B_s), (\psi \in B_j B_s B_j B_s), (\psi \in B_j B_s D_j \widetilde{B}_j), (\psi \in B_j \widetilde{B}_s \widetilde{B}_j) \;\in\; post(dqbr_1(s, j, \psi)) \tag{7.35}$$
$$(\psi \in B_j D_s B_j), (\psi \in B_j B_s D_j \widetilde{B}_j), (\psi \in B_j \widetilde{B}_s \widetilde{B}_j) \;\in\; post(dqbr_2(s, j, \psi)) \tag{7.36}$$

In a similar fashion, Sarah can derive properties of John's mental states. After uttering a $dqbr_1(s, j, \psi)$, Sarah is justified to believe that John believes that she believes ψ. After a $dqbr_2(s, j, \psi)$, Sarah is justified to believe that John believes that she desires him to believe ψ.

$$(\psi \in B_s B_j B_s) \;\in\; post(dqbr_1(s, j, \psi)) \tag{7.37}$$
$$(\psi \in B_s B_j D_s B_j) \;\in\; post(dqbr_2(s, j, \psi)) \tag{7.38}$$

7.3.7 Distributed belief revision

Next, we describe how distributed chains of justification are traversed by distributed belief revision. If Sarah has a desire to be ignorant about ψ, i.e. $\psi \in D_s \widetilde{B}_s$, and she

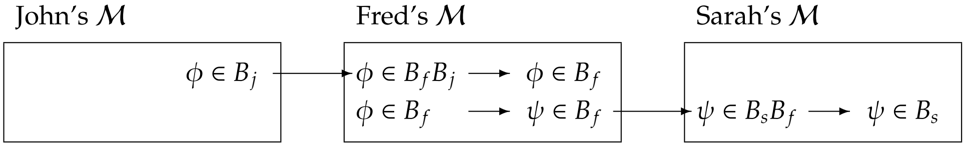

Figure 7.5: A distributed chain of justification (equal to figure 7.2).

believes ψ, i.e. $\psi \in B_s$, while her mental state structure is closed under decision-making, then her belief in ψ must be entrenched in her cognitive state, i.e. $\mathcal{M}_s \models \psi \in^{\in} B_s$. Two reasons exist why her belief in ψ can be entrenched.

1. The chain of justification (cf. Section 7.2.7) for Sarah's belief in ψ ends in her belief in ϕ that is not part of her indirect desires to be ignorant about, i.e. $\phi \notin ind(D_s\widetilde{B_s})$. Because of $\phi \notin ind(D_s\widetilde{B_s})$, the closure under decision-making does not retract Sarah's explicit justification to come to believe ψ again, that is to say, her decisions cannot retract her belief in ϕ, and thus her belief in ψ remains entrenched. Analogously, the chain of justification can also end in Sarah's ignorance about ϕ that is not part of her indirect desires to believe, i.e. $\phi \notin ind(D_s B_s)$.

2. The chain of justification for Sarah's belief in ψ ends in her belief that another agent, say John, believes ϕ. Analogously, the chain of justification can also end in her belief that John is ignorant about ϕ.

In the following example, we show that the distributed chain of justification in figure 7.5, that is created by the example dialogue from Section 7.2.2, will be traversed by the example dialogue from Section 7.3.3. Assume that the chain of justification for Sarah's belief in ψ ends in her belief that Fred believes ψ. The dialogue game presented in this section allows Sarah to request Fred to retract ψ. Due to the perlocution of the speech acts, Fred will become to believe that Sarah desires to be ignorant about ψ. With this belief, he will consider whether he can retract his belief in ψ. Either Fred's belief in ϕ is not cognitively entrenched, in which case he will retracts ψ with the appropriated decision game, and he will reports back to Sarah that he no longer believes ψ. Alternatively, his belief in ψ is entrenched, in which case he cannot retract his belief in ψ.

Assume that the chain of justification for Fred's belief in ψ ends in his belief that John believes ϕ. Fred requests John to retract ϕ. Consequently, John comes to believe that Fred desires to be ignorant about ϕ. John considers whether he can retract his belief in ϕ, which he can because ϕ is not entrenched in his cognitive state. He retracts it and reports to Fred. Fred's belief in ψ is now no longer entrenched. Fred retracts ϕ and ψ, and he reports to Sarah. Because Sarah's belief in ψ now also not longer entrenched, she retracts it and balance her mental state structure.

In general, the dialogue game presented in this section provides communicative acts that may provide the explicit justifications that may allow agents to retract or add beliefs by traversing the chain of justification backwards. The distributed chain of justification of an agent's belief is considered recursively for retraction by the different agents that provide the distributed chain.

7.4 Requesting a Belief Addition

In this section, we will present a dialogue game that provides the rules of usage for agents to request others to believe propositions. This section is partly based on published work by Lebbink et al. [LWM03c, LWM03b].

The previous two sections provided the meaning of communicative acts that enable agents to ask others for information such that they may adopt beliefs (Sections 7.2), and to ask others for information such that they may retract beliefs (Section 7.3). In this section, we provide the semantics of communicative acts in which agents request others to adopt beliefs.

7.4.1 Intuitive readings

The utterance of a request for a belief addition, or *rba* for short, will allow agents to request other agents to decide to believe propositions. With a *rba*, an agent intends to convey her desire that the intended receiver believes a certain proposition. Unlike the illocutions of a qba_1 and a qbr_1, with a *rba*, speaker Sarah does not intend to convey information whether she is justified or not to believe the proposition herself. The receiver John has to acquire justification to believe the proposition via another illocution, such as with a *qba* or a *qbr*. An agent may utter a locution that a receiver may interpret as the illocution of granting a request for belief addition, denoted *grba*. With a $grba_1$, a speaker provides a receiver with a testimony that she had judged herself justified to decide to believe the proposition she has been requested to believe. With a $grba_2$, speaker Sarah acknowledges to the intended receiver John that she agrees with his desire that she is to believe a certain proposition. With a $drba_1$, a speaker provides a receiver with a testimony that she did not judge herself justified to decide to believe the proposition. In a $drba_2$, the speaker intends to convey to the intended listener that she has an indirect desire that conflicts with the listener's desire that the speaker is to believe a proposition. John could request Sarah to believe that dinner will be served in two hours, i.e. $rba(j, s, \psi)$. Because of certain circumstances in their restaurant, Sarah could respond that she agrees to believe that, i.e. $grba_1(s, j, \psi)$. However, because their movie starts in two hours, she does not agree to desire to believe it, i.e. $drba_2(s, j, \psi)$. In short, the communicative acts have the following reading.

- $rba(s, j, \psi)$ reads "will you please believe ψ?"
- $grba_1(s, j, \psi)$ reads "I believe ψ."
- $grba_2(s, j, \psi)$ reads "I am willing to believe ψ."

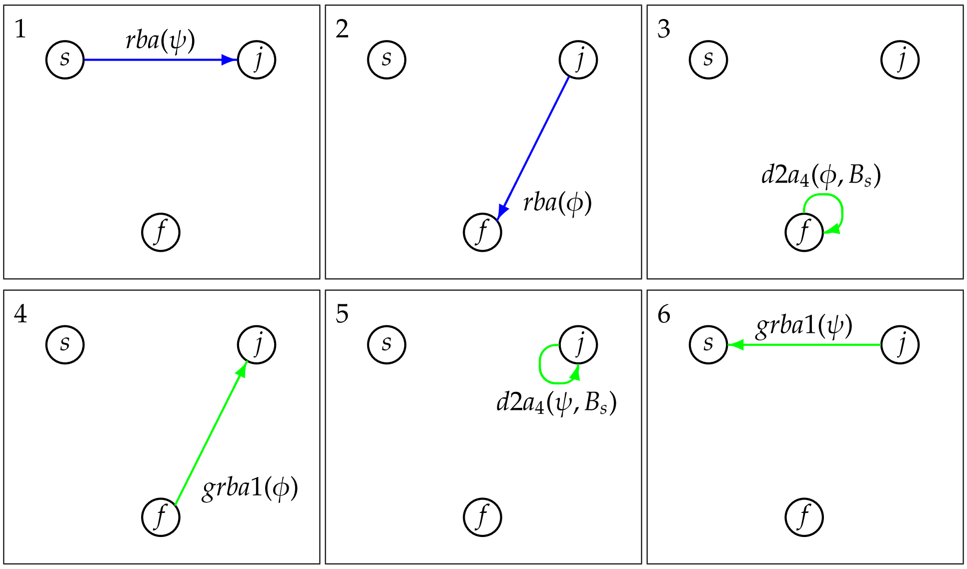

Figure 7.6: Dialogues about requests for belief addition (example 7.4).

- $drba_1(s, j, \psi)$ reads "I do not believe ψ."
- $drba_2(s, j, \psi)$ reads "I do not want to believe ψ."

Sarah is motivated to utter an *rba* to John, if she has a mental state structure in which she desires that John believes a proposition ψ, and she does not believe that John believes ψ, i.e. $\mathcal{M}_s \models (\psi \in ind(D_s B_j)), (\psi \notin B_s B_j)$.

7.4.2 An example dialogue

Example 7.4. *See figure 7.6 for a depiction. Consider the following interpretation (as in example 7.2), ψ reads as "the patient is suffering from disease X", and ϕ reads as "the patient shows symptom Y". Symptom Y is a sufficient criterion to determine disease X.*

1. *Assume Sarah desires that John believes ψ, i.e. $\psi \in D_s B_j$, and that she does not believe that John believes ψ, i.e. $\psi \notin B_s B_j$. Sarah requests John to believe ψ.*
2. *Assume that John does not believe ψ, i.e. $\psi \notin B_j$, and that a criterion for him to believe ψ is that Fred believes ϕ, for example, John knows an inference rule $(\phi \in B_j B_f) \rightarrowtail (\psi \in B_j) \in \mathcal{K}_j$. John requests Fred to believe ϕ.*
3. *Assume that Fred does not believe ϕ, i.e. $\phi \notin B_f$, and that he can comply with John's desire to believe ϕ: he decides to believe ϕ.*
4. *Fred grants John's request to believe ϕ.*

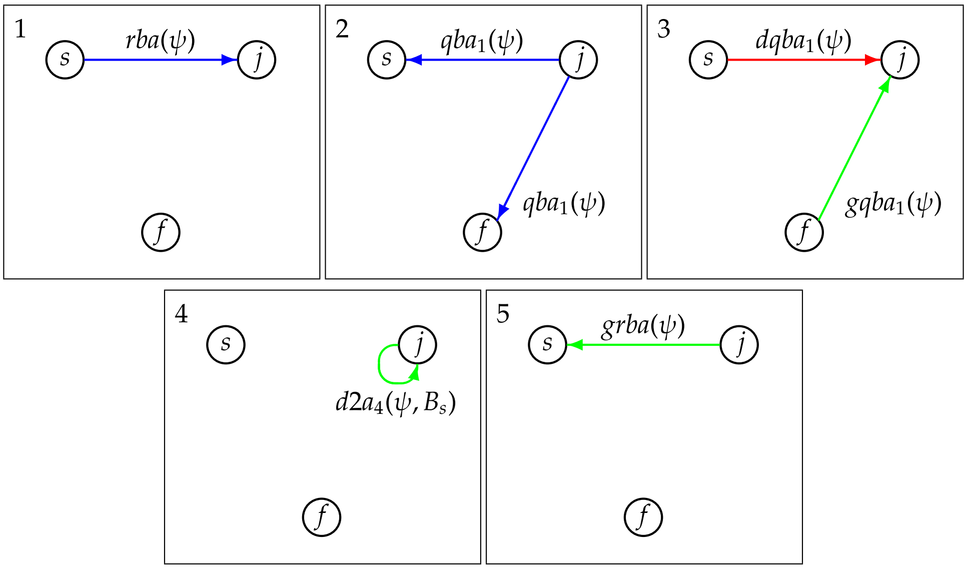

Figure 7.7: Dialogues about requests for belief addition (example 7.5).

5. *John may now decide to believe ψ.*
6. *John grants Sarah's request to believe ψ.*

Example 7.5. *See figure 7.7 for a depiction. Consider the following interpretation (as in example 7.2), ψ reads as "the patient is suffering from disease X", and ϕ reads as "the patient shows symptom Y". Symptom Y is a sufficient criterion to determine disease X.*

1. *Assume Sarah desires that John believes ψ, i.e. $\psi \in D_s B_j$, and Sarah does not believe that John believes ψ, i.e. $\psi \notin B_s$. Sarah requests John to believe ψ.*
2. *Assume that John does not believe ψ, i.e. $\psi \notin B_j$. John asks Fred whether he may believe ψ.*
3. *Assume that Fred does believe ψ, i.e. $\psi \in B_f$, he grants John's question.*
4. *John decides to believe ψ.*
5. *John grants Sarah's request to believe ψ.*

7.4.3 Requesting a belief addition

Sarah is motivated to request John to believe ψ if she indirectly desires John to believe ψ, and she does not believe that John believes ψ.

$$(\psi \in ind(D_s B_j)), (\psi \notin B_s B_j) \in pre(rba(s, j, \psi)) \tag{7.39}$$

After receiving a $rba(s, j, \psi)$, and given the communicative act's preconditions, John may deduce the following properties of Sarah's mental states. John believes that Sarah desires him to believe ψ, and that Sarah did not believe that he believes ψ. After the utterance of the request, the mental state structure of the receiver John has changed according to the following post-conditions.

$$(\psi \in B_j D_s B_j), (\psi \in B_j \widetilde{B_s} B_j) \;\in\; post(rba(s, j, \psi)) \tag{7.40}$$

A speaker may assume that the addressee derives the same post-conditions as she would have done if she had received the communicative act herself. Consequently, after having uttered an $rba(s, j, \psi)$, Sarah believes that receiver John believes that she desires him to believe ψ, i.e. $\psi \in B_s B_j D_s B_j$.

$$(\psi \in B_s B_j D_s B_j) \;\in\; post(rba(s, j, \psi)) \tag{7.41}$$

7.4.4 Granting a request for belief addition

Sarah is justified to utter the illocution $grba_1(s, j, \psi)$ in response to a $rba(j, s, \psi)$ if she believes that John has the desire that she believes ψ, and she does not believe that John believes that she believes ψ. A post-condition of a $rba(j, s, \psi)$ (eq. (7.40)) is that Sarah believes that John is ignorant whether she believes ψ, i.e. $\psi \in B_s B_j B_s$. With the coherence relation from equation (5.7), we have that Sarah does not believe that John believes that she believes ψ, i.e. $\psi \notin B_s B_j B_s$. We will use the latter as a precondition because it is less limited and will allow the application of the *grba* in more situations. In addition, if Sarah believes ψ, then she may utter a $grba_1(s, j, \psi)$. Sarah is justified to utter a $grba_2(s, j, \psi)$ if Sarah does not have an indirect desire to be ignorant about ψ.

$$(\psi \in B_s), (\psi \in B_s D_j B_s), (\psi \notin B_s B_j B_s) \;\in\; pre(grba_1(s, j, \psi)) \tag{7.42}$$

$$(\psi \notin ind(D_s \widetilde{B_s})), (\psi \in B_s D_j B_s), (\psi \notin B_s B_j B_s) \;\in\; pre(grba_2(s, j, \psi)) \tag{7.43}$$

Given these preconditions, receiver John may deduce the following properties of Sarah's mental states. From a $grba_1(s, j, \psi)$, John is justified to believe that Sarah believes ψ, and that he believes that Sarah believes that he desires her to believe ψ. Remark that if John interprets that the $grba_1(s, j, \psi)$ is in response to an $rba_1(j, s, \psi)$, then John is already justified to have $\psi \in B_j B_s D_j B_s$. From a $grba_2(s, j, \psi)$, John is justified to believe that Sarah does not desire herself to be ignorant about ψ.

$$(\psi \in B_j B_s), (\psi \in B_j B_s B_j B_s), (\psi \in B_j B_s D_j B_s) \;\in\; post(grba_1(s, j, \psi)) \tag{7.44}$$

$$(\psi \in B_j \widetilde{D_s} \widetilde{B_s}), (\psi \in B_j B_s D_j B_s) \;\in\; post(grba_2(s, j, \psi)) \tag{7.45}$$

Speaker Sarah is justified to have the following mental properties: after uttering a $grba_1(s, j, \psi)$, she may believe that John believes that she believes ψ, and after uttering a $grba_2(s, j, \psi)$, she may believe that John believes that she believes that she does not desire herself to be ignorant about ψ.

$$(\psi \in B_s B_j B_s) \;\in\; post(grba_1(s, j, \psi)) \tag{7.46}$$

$$(\psi \in B_s B_j \widetilde{D_s} \widetilde{B_s}) \;\in\; post(grba_2(s, j, \psi)) \tag{7.47}$$

7.4.5 Declining a request for belief addition

An agent is justified to utter a negative response *drba*(s, j, ψ) in similar conditions as when she is justified to utter a positive response. The difference is that in case of the negative response Sarah does not believe ψ, while in the positive response she does.

Sarah is justified to utter a *drba*$_1(s, j, \psi)$ in response to a *rba*$_1(j, s, \psi)$ if Sarah believes that John desires her to believe ψ, and she does not believe that John believes that she believes ψ, i.e. $\psi \in B_sD_jB_s$ and $\psi \notin B_sB_jB_s$. Additionally, Sarah is ignorant about ψ, i.e. $\psi \notin B_s$, and she has no communicative acts left to become justified to decide to believe ψ. Sarah has run out of communicative acts to become justified to be believe ψ if her mental state structure enjoys the property *finished*$(s, \psi \in B_s)$ as expressed in equation (7.79) in Section 7.7. Sarah is justified to utter a *drba*$_2(s, j, \psi)$ if she cannot comply with John's desire that she believes ψ because she has an indirect desire to be ignorant about ψ.

$$
\begin{aligned}
(\psi \notin B_s), (\psi \in B_sD_jB_s), (\psi \notin B_sB_jB_s) &\in \mathit{pre}(\mathit{drba}_1(s, j, \psi)) && (7.48)\\
(\psi \in \mathit{ind}(D_s\widetilde{B_s})), (\psi \in B_sD_jB_s), (\psi \notin B_sB_jB_s) &\in \mathit{pre}(\mathit{drba}_2(s, j, \psi)) && (7.49)
\end{aligned}
$$

After receiving a *drba*$_1(s, j, \psi)$, John is justified to believe that Sarah is ignorant about ψ, and that she believes that he desires her to be ignorant about ψ. After receiving a *drba*$_2(s, j, \psi)$, John is justified to believe that Sarah desires to be ignorant about ψ.

$$
\begin{aligned}
(\psi \in B_j\widetilde{B_s}), (\psi \in B_jB_sB_j\widetilde{B_s}), (\psi \in B_jB_sD_jB_s) &\in \mathit{post}(\mathit{drba}_1(s, j, \psi)) && (7.50)\\
(\psi \in B_jD_s\widetilde{B_s}), (\psi \in B_jB_sD_jB_s) &\in \mathit{post}(\mathit{drba}_2(s, j, \psi)) && (7.51)
\end{aligned}
$$

After uttering a *drba*$_1(s, j, \psi)$, Sarah is justified to believe that John believes that she is ignorant about ψ, and after uttering a *drba*$_2(s, j, \psi)$, Sarah is justified to believe that John believes that she desires to be ignorant about ψ.

$$
\begin{aligned}
(\psi \in B_sB_j\widetilde{B_s}) &\in \mathit{post}(\mathit{drba}_1(s, j, \psi)) && (7.52)\\
(\psi \in B_sB_jD_s\widetilde{B_s}) &\in \mathit{post}(\mathit{drba}_2(s, j, \psi)) && (7.53)
\end{aligned}
$$

7.5 Requesting a Belief Retraction

In this section, we will present a dialogue game that provides the rules of usage for agents to request others to retract beliefs. This section is partly based on published work by Lebbink et al. [LWM05].

7.5.1 Intuitive readings

With a request for belief retraction, or *rbr* for short, an agent provides information to another agent with the intention that the receiver retracts a belief. With the communicative act of granting a request for belief retraction, or *grbr* for short, an

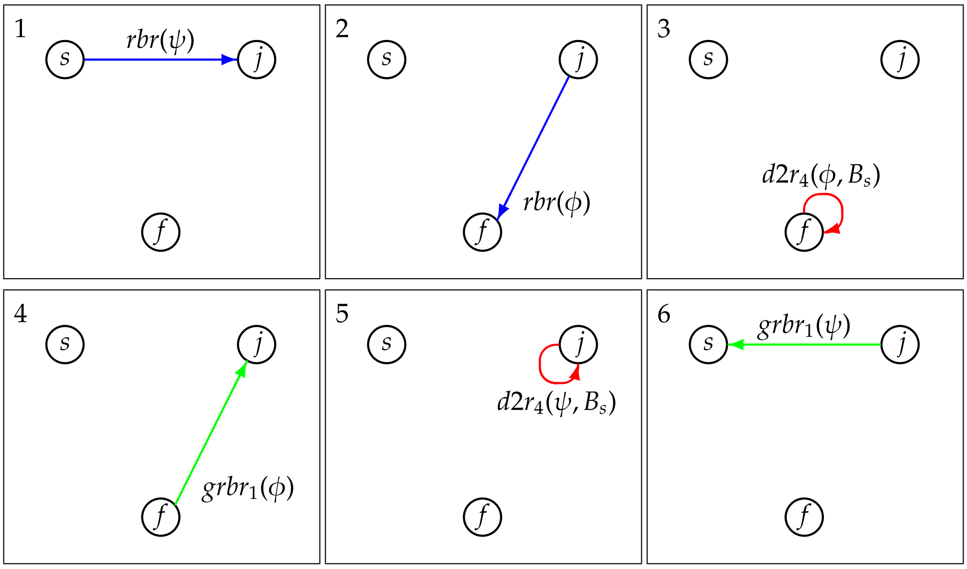

Figure 7.8: Dialogues about requests for belief retraction (example 7.6).

agent agrees to be ignorant about a proposition. With a communicative act of denying a request for belief retraction, or *drbr* for short, an agent declines to be ignorant about a proposition. In short, the communicative acts have the following reading.

- $rbr(s, j, \psi)$ reads "will you please be ignorant about ψ?"
- $grbr_1(s, j, \psi)$ reads "I do not believe ψ."
- $grbr_2(s, j, \psi)$ reads "I am willing to be ignorant about ψ."
- $drbr_1(s, j, \psi)$ reads "I believe ψ."
- $drbr_2(s, j, \psi)$ reads "I do not want to be ignorant about ψ."

Note that the reading of $grbr_1$ is equal to the reading of a $grba_1$ from Section 7.4.1. This equivalence will be discussed in Section 7.6.1 on the intuitive reading of inform statements.

Sarah has an unbalanced mental state structure if she desires that John does not believe ψ, and she does not believe that John is ignorant about ψ, i.e. $\mathcal{M}_s \models (\psi \in ind(D_s\widetilde{B_j})), (\psi \in B_sB_j)$. If Sarah is in such a state, she is motivated to request John to retract his belief in ψ, and she is motivated to provide him with information such that he becomes justified to decide to retract ψ.

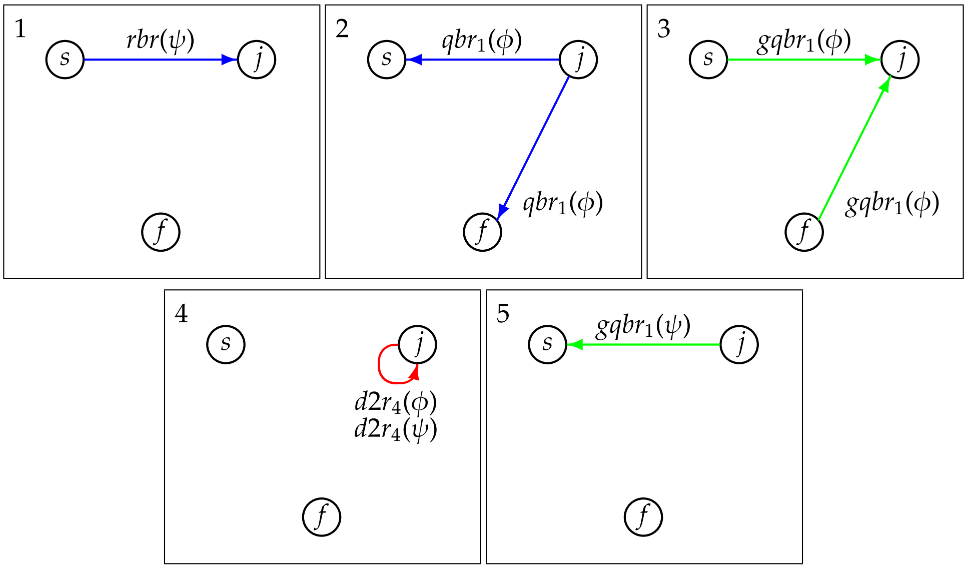

Figure 7.9: Dialogues about requests for belief retraction (example 7.7).

7.5.2 An example dialogue

Example 7.6. *See figure 7.8. Consider the following interpretation (as in example 7.2), ψ reads as "the patient is suffering from disease X", and ϕ reads as "the patient shows symptom Y". Symptom Y is a sufficient criterion to determine disease X.*

1. *Assume that Sarah desires that John does not believe ψ, i.e. $\psi \in D_s\widetilde{B_j}$, and she believes that John does believe ψ, i.e. $\psi \in B_sB_j$. Sarah request John to retract ψ.*

2. *Assume that John believes ψ, i.e. $\psi \in B_j$, and that ψ is entrenched in his mental state structure due to his belief that Fred believes ϕ, i.e. $\phi \in B_jB_f/\psi \in B_j$. Thus, John indirectly desires that Fred is ignorant about ϕ; this would make his belief in ψ un-entrenched. John requests Fred to retract ϕ.*

3. *Assume Fred believes ϕ, i.e. $\phi \in B_f$, but his belief in ϕ is not entrenched in his mental state structure. With decision rule $d2r_4$ he retracts his belief in ϕ.*

4. *Fred does not believe ϕ (anymore), and grants John's request to retract ϕ.*

5. *Assume that John's belief in ψ is now not entrenched anymore, and he can thus retract his belief in ψ.*

6. *John does not believe ψ (anymore), and grants Sarah's request to retract ψ.*

Example 7.7. *See figure 7.9. Consider the following interpretation (as in example 7.2), ψ reads as "the patient is suffering from disease X", and ϕ reads as "the patient shows symptom Y". Symptom Y is a sufficient criterion to determine disease X.*

1. *Assume that Sarah desires that John does not believe ψ, i.e. $\psi \in D_s\widetilde{B}_j$, and she believes that John does believe ψ, i.e. $\psi \in B_sB_j$. Sarah request John to retract ψ.*
2. - *Assume that John believes ψ, i.e. $\psi \in B_j$.*
   - *Assume that ψ is entrenched in John's mental state structure due to his belief in ϕ, i.e. $\phi \in B_j / \psi \in B_j$.*
   - *Assume that John indirectly desires to be ignorant about ϕ and ψ. John may thus asks both Sarah and Fred whether he may be ignorant about ϕ.*
3. *Sarah and Fred both respond affirmative that John may be ignorant about ϕ.*
4. *John first retracts his belief in ϕ, and, as a result, his belief in ψ is not entrenched anymore. John then retracts his belief in ψ.*
5. *John does not believe ψ (anymore), and grants Sarah's request to retract ψ.*

7.5.3 Requesting a belief retraction

Sarah is motivated to utter an $rbr(s, j, \psi)$ if the following two criteria hold. Sarah desires that John is ignorant about proposition ψ, and she does not believe that John is ignorant about ψ, i.e. $\psi \notin B_s\widetilde{B}_j$. The latter condition could be replaced by a more informative condition that Sarah believes that John believes ψ, i.e. $\psi \in B_sB_j$. However, as captured in the coherence relation from equation (5.3), the latter condition tends to restrict the application of the *rbr* unnecessarily.

$$(\psi \in ind(D_s\widetilde{B}_j)), (\psi \notin B_s\widetilde{B}_j) \;\in\; pre(rbr(s, j, \psi)) \tag{7.54}$$

John judges that Sarah uses the same preconditions for a *rbr* as he does. Based on this, John is justified to have the following mental states after the $rbr(s, j, \psi)$ has been uttered. John believes that Sarah desires that he is ignorant about ψ, i.e. $\psi \in B_jD_s\widetilde{B}_j$. Secondly, he believes that Sarah is ignorant about whether he is ignorant about ψ, i.e. $\psi \in B_j\widetilde{B}_s\widetilde{B}_j$. If the precondition of the $rbr(s, j, \psi)$ had been $\psi \in B_sB_j$ instead of $\psi \notin B_s\widetilde{B}_j$, John would have been justified to have $\psi \in B_jB_sB_j$ instead of $\psi \in B_j\widetilde{B}_s\widetilde{B}_j$. Because Sarah used a less restricting condition, John can only come to believe a less informative statement. This observation is consistent with the coherence relation from equation (5.4), i.e. $B_sB_jB_s \subseteq B_s\widetilde{B}_s\widetilde{B}_j$.

$$(\psi \in B_jD_s\widetilde{B}_j), (\psi \in B_j\widetilde{B}_s\widetilde{B}_j) \;\in\; post(rbr(s, j, \psi)) \tag{7.55}$$

Sarah may also judge that John uses the same post-conditions for a *rbr* as she does. She is thus justified to believe that he believes that she desires him to be ignorant about ψ.

$$(\psi \in B_sB_jD_s\widetilde{B}_j) \;\in\; post(rbr(s, j, \psi)) \tag{7.56}$$

7.5.4 Granting a request for belief retraction

Sarah may utter a $grbr_1(s, j, \psi)$ if she believes that John desires her to be ignorant about ψ, and she is ignorant about ψ. Sarah may utter a $grbr_2(s, j, \psi)$ if she does not indirectly desire to believe ψ.

$$(\psi \notin B_s), (\psi \in B_s D_j \widetilde{B}_s), (\psi \in B_s \widetilde{B}_j \widetilde{B}_s) \;\in\; pre(grbr_1(s, j, \psi)) \tag{7.57}$$
$$(\psi \notin ind(D_s B_s)), (\psi \in B_s D_j \widetilde{B}_s), (\psi \in B_s \widetilde{B}_j \widetilde{B}_s) \;\in\; pre(grbr_2(s, j, \psi)) \tag{7.58}$$

After receiving a $grbr_1(s, j, \psi)$, John is explicitly justified to believe that Sarah is ignorant about ψ, and that Sarah believes that he desires her to be ignorant about ψ. After receiving a $grbr_1(s, j, \psi)$, John is explicitly justified to believe that Sarah does not desire him to be ignorant about ψ.

$$(\psi \in B_j \widetilde{B}_s), (\psi \in B_j B_s B_j \widetilde{B}_s), (\psi \in B_j B_s D_j \widetilde{B}_s) \;\in\; post(grbr_1(s, j, \psi)) \tag{7.59}$$
$$(\psi \in B_j \widetilde{D}_s B_j), (\psi \in B_j B_s D_j \widetilde{B}_s) \;\in\; post(grbr_2(s, j, \psi)) \tag{7.60}$$

After uttering a $grbr_1(s, j, \psi)$, Sarah is explicitly justified to believe that John believes that she is ignorant about ψ, and after uttering a $grbr_2(s, j, \psi)$, Sarah is explicitly justified to believe that John believes that she does not desire him to be ignorant about ψ.

$$(\psi \in B_s B_j \widetilde{B}_s) \;\in\; post(grbr_1(s, j, \psi)) \tag{7.61}$$
$$(\psi \in B_s B_j \widetilde{D}_s B_j) \;\in\; post(grbr_2(s, j, \psi)) \tag{7.62}$$

7.5.5 Declining a request for belief retraction

Sarah is justified to utter a $drbr(s, j, \psi)$ in response to a $rbr(j, s, \psi)$ if the following criteria hold. Sarah believes that John desires her to be ignorant about ψ i.e. $\psi \in B_s D_j \widetilde{B}_s$, and she does not believe that he is ignorant about whether she is ignorant about ψ, i.e. $\psi \in B_s \widetilde{B}_j \widetilde{B}_s$. A $drbr_1(s, j, \psi)$ is allowed, if, in addition to these two conditions, Sarah believes ψ, i.e. $\psi \in B_s$, and Sarah has no untried communicative acts that may make her justified to become ignorant about ψ. Sarah has run out of communicative acts to become justified to be believe ψ if her mental state structure enjoys the property *finished*$(s, \psi \notin B_s)$ as expressed in equation (7.79) in Section 7.7. A $drbr_2(s, j, \psi)$ is allowed if Sarah also indirectly desires to believe ψ, i.e. $\psi \in ind(D_s B_s)$.

$$(\psi \in B_s), (\psi \in B_s D_j \widetilde{B}_s), (\psi \in B_s \widetilde{B}_j \widetilde{B}_s) \;\in\; pre(drbr_1(s, j, \psi)) \tag{7.63}$$
$$(\psi \in ind(D_s B_s)), (\psi \in B_s D_j \widetilde{B}_s), (\psi \in B_s \widetilde{B}_j \widetilde{B}_s) \;\in\; pre(drbr_2(s, j, \psi)) \tag{7.64}$$

From the preconditions of a $drbr_1(s, j, \psi)$, receiver John may derive properties of Sarah's mental states. John is justified to believe that Sarah believes ψ, and that he believes that she believes this, and that Sarah believes that he desires her to be ignorant about ψ. The post-conditions of a $drbr_2(s, j, \psi)$ are that John is justified to believe that Sarah desires to believe ψ.

$$\begin{aligned}(\psi \in B_jB_s), (\psi \in B_jB_sB_jB_s), (\psi \in B_jB_sD_j\widetilde{B_s}) &\in post(drbr_1(s, j, \psi)) && (7.65)\\ (\psi \in B_jD_sB_s), (\psi \in B_jB_sD_j\widetilde{B_s}) &\in post(drbr_2(s, j, \psi)) && (7.66)\end{aligned}$$

After uttering a $drbr_1(s, j, \psi)$, Sarah is justified to believe that John believes that she believes ψ. After uttering a $drbr_2(s, j, \psi)$, Sarah is justified to believe that John believes that she desires to believe ψ.

$$\begin{aligned}(\psi \in B_sB_jB_s) &\in post(drbr_1(s, j, \psi)) && (7.67)\\ (\psi \in B_sB_jD_sB_s) &\in post(drbr_2(s, j, \psi)) && (7.68)\end{aligned}$$

7.6 Inform Statement about Changed Belief and Desire

7.6.1 Intuitive reading

In an inform statement, abbreviated *is*, an agent communicates information about her changed mental state, so that the receiver obtains a correct belief about her mental state. If Sarah has adopted a belief, then she may inform John that she has changed her mental state with the intent that John believes that she has a new belief. The communicative act has the following reading, with $ms \in MSN_s$:

- $is(s, j, \psi, ms)$ reads "I have $\psi \in ms$."

Specifically, if ms equals B_s, then $is(s, j, \psi, B_s)$, has a reading equal to the $grba_1(s, j, \psi)$ from Section 7.4.1, and the $drbr_1(s, j, \psi)$ from Section 7.5.1: "I believe ψ." Analogously, if ms equals $\widetilde{B}_s$, then $is(s, j, \psi, \widetilde{B}_s)$ has a reading equal to the $drbr_1(s, j, \psi)$ and the $grba_1(s, j, \psi)$: "I do not believe ψ."

We could define one locution with the reading "I believe ψ" that the agents have to interpret to be either a $grba_1$, $drbr_1$ or is (cf. Section 7.2.4 on interpreting communicative acts). Interpretation should then be based on other previously uttered communicative acts. We chose to make syntactically different locutions that need not be interpreted. This makes it possible for agents to judge from the syntactical properties of incoming communication to tell which illocution the speaker has used. This approach is only feasible if the number of speech acts is reasonably small. Because, if our agents would need thousands of different speech acts, the computational costs would outweigh the complexity of interpreting whether a single locution is one of many different meanings. If communicative acts were interpreted by agents, the agent's communication language could be considerable smaller because fewer communicative acts would be needed to convey information. Because of pragmatic reasons, such as complexity of the framework, we chose our agents not to interpret communicative acts.

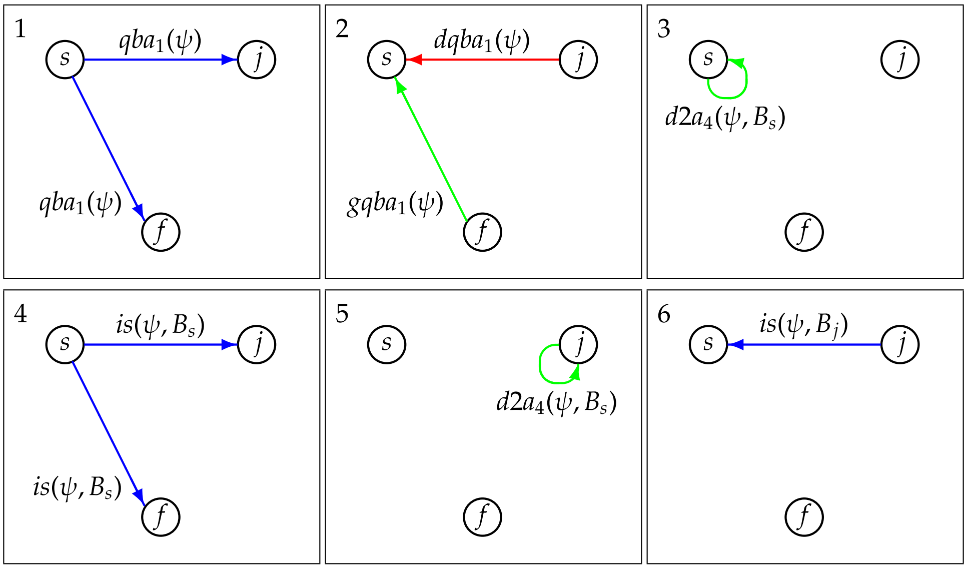

Figure 7.10: Dialogue with inform statements about belief change (example 7.8).

7.6.2 An example dialogue

Example 7.8 (inform about changed belief)**.** *Assume that Sarah desires to believe ϕ which she does not believe,*

- $\mathcal{M}_s \models (\psi \notin B_s), (\phi \notin B_s), (\phi \in D_s B_s)$ *and* $(\psi \in B_s) \rightarrowtail (\phi \in B_s) \in \mathcal{K}_s$*,*
- $\mathcal{M}_j \models (\psi \notin B_j), (\phi \notin B_j)$
- $\mathcal{M}_f \models (\psi \in B_f), (\phi \notin B_f)$

See figure 7.10 for a depiction.

1. *Sarah asks Fred and John whether she may believe ψ.*
2. *As a result, among others, Sarah believes that John and Fred believe that she does not believe ψ. Assume that John does not believe ψ: he answers Sarah's question negatively, and assume that Fred does believe ψ: he answers Sarah's question positively.*
3. *Assume that Sarah, based on her belief that Fred believes ψ, decides to believe ψ.*
4. *From situation 1, Sarah believes that John and Fred believe that she does not believe ψ; however, she now does believe ψ. Sarah informs John and Fred that she believes ψ.*
5. *After receiving the $is(s, j, \psi, B_s)$, John is justified to believe that Sarah believes ψ. Assume that John, based on his belief that Sarah believes ψ, decides to believe ψ.*
6. *From situation 2, John believes that Sarah believes that he does not believe ψ; however, he now does believe ψ. John informs Sarah that he believes ψ.*

7.6.3 Inform statement

Sarah is justified to utter an $is(s, j, \psi, B_s)$ if she believes that John believes that she is ignorant about ψ, while she believes ψ. Analogously, Sarah is justified to utter an $is(s, j, \psi, \widetilde{B}_s)$ if she believes that John believes that she believes ψ, while she is ignorant about ψ.

$$(\psi \in B_s B_j \widetilde{B}_s), (\psi \in B_s) \;\in\; pre(is(s, j, \psi, B_s)) \tag{7.69}$$
$$(\psi \in B_s B_j B_s), (\psi \notin B_s) \;\in\; pre(is(s, j, \psi, \widetilde{B}_s)) \tag{7.70}$$

The post-conditions are straightforward, after receiving an $is(s, j, \psi, B_s)$, the receiver John is explicitly justified to believe that Sarah believes ψ, and that he believes that she believes this. After receiving an $is(s, j, \psi, \widetilde{B}_s)$, John is explicitly justified to believe that Sarah is ignorant about ψ, and that he believes that she believes this.

$$(\psi \in B_j B_s), (\psi \in B_j B_s B_j B_s) \;\in\; post(is(s, j, \psi, B_s)) \tag{7.71}$$
$$(\psi \in B_j \widetilde{B}_s), (\psi \in B_j B_s B_j \widetilde{B}_s) \;\in\; post(is(s, j, \psi, \widetilde{B}_s)) \tag{7.72}$$

The post-conditions for the speaker of the inform statement are also straightforward. After uttering an $is(s, j, \psi, B_s)$, speaker Sarah is explicitly justified to believe that John believes that she believes ψ. After uttering an $is(s, j, \psi, \widetilde{B}_s)$, Sarah is explicitly justified to believe that John believes that she is ignorant about ψ.

$$(\psi \in B_s B_j B_s) \;\in\; post(is(s, j, \psi, B_s)) \tag{7.73}$$
$$(\psi \in B_s B_j \widetilde{B}_s) \;\in\; post(is(s, j, \psi, \widetilde{B}_s)) \tag{7.74}$$

7.6.4 Correct representations

Next, we describe several properties of the dialogue games. If John believes that Sarah believes a proposition, then Sarah believes that John believes that she believes the proposition, is a property that remains valid under communication. That is to say, if John believes that Sarah believes ψ, i.e. $\psi \in B_j B_s$, and Sarah believes that John believes that she believes ψ, i.e. $\psi \in B_s B_j B_s$, thus, $B_j B_s \subseteq B_s B_j B_s$, then, after the closure of their mental state structures under communication, still holds $B_j B_s \subseteq B_s B_j B_s$.

Proposition 7.1. *Suppose $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_s B_j B_s \in \mathcal{M}_s$ and $B_j B_s \in \mathcal{M}_j$ with $B_j B_s \subseteq B_s B_j B_s$, then after any communication yielding $\mathcal{M}'_s$ and $\mathcal{M}'_j$ with $B_s B_j B'_s \in \mathcal{M}'_s$ and $B_j B'_s \in \mathcal{M}'_j$ holds $B_j B'_s \subseteq B_s B_j B'_s$. Sketch of proof in Appendix A.4.*

A property of the dialogue game about inform statements is that if Sarah's mental state structure is closed under inform statements then if Sarah believes that John believes that she believes ψ, then Sarah does also believe ψ, i.e. $B_s B_j B_s \subseteq B_s$. Stated differently, if Sarah is not allowed to utter inform statements to John, then Sarah will be aware that John has represent her beliefs correctly if what she believes that John believes about her beliefs is actually believed by her. Sarah is said to be aware that

John has represented her beliefs incorrectly if and only if $B_sB_jB_s \nsubseteq B_s$, if this situation holds, then because an inform statement is applicable, Sarah will inform John about his beliefs.

Proposition 7.2. *If Sarah is not allowed to utter an inform statement to John that she is ignorant about ψ, i.e. $\mathcal{M}_s \not\models pre(is(s, j, \psi, \widetilde{B_s}))$, then Sarah's beliefs about John's beliefs about her beliefs are correct, i.e. $\mathcal{M}_s \models \psi \in B_sB_jB_s \Rightarrow \mathcal{M}_s \models \psi \in B_s$. Proof in Appendix A.4.*

Another important property is that agents' beliefs about other agents' beliefs are correct. John is said to have correctly represented Sarah's beliefs if what he believes about Sarah's beliefs is actually believed by Sarah, i.e. $B_jB_s \subseteq B_s$. However, because John has no access to Sarah's true beliefs, he cannot be aware that he has correctly represented her beliefs. Still the dialogue games provide this property if holds $B_jB_s \subseteq B_sB_jB_s$.

Proposition 7.3. *If Sarah is not allowed to utter an inform statement to John that she is ignorant about ψ, i.e. $\mathcal{M}_s \not\models pre(is(s, j, \psi, \widetilde{B_s}))$, and $B_jB_s \subseteq B_sB_jB_s$, then John's beliefs about Sarah's beliefs are correct, i.e. $\mathcal{M}_j \models \psi \in B_jB_s \Rightarrow \mathcal{M}_s \models \psi \in B_s$. Proof in Appendix A.4.*

The previous three propositions about agents' beliefs have their duals about agents' ignorance. A property of the dialogue games is that if Sarah's and John's cognitive states are closed under communication and John believes that Sarah is ignorant about ψ, then Sarah believes that John believes that she is ignorant about ψ, i.e. $B_j\widetilde{B_s} \subseteq B_sB_j\widetilde{B_s}$.

Proposition 7.4. *Suppose $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_sB_j\widetilde{B_s} \in \mathcal{M}_s$ and $B_j\widetilde{B_s} \in \mathcal{M}_j$ with $B_j\widetilde{B_s} \subseteq B_sB_j\widetilde{B_s}$, then after any communication yielding $\mathcal{M}'_s$ and $\mathcal{M}'_j$ with $B_sB_j\widetilde{B'_s} \in \mathcal{M}'_s$ and $B_j\widetilde{B'_s} \in \mathcal{M}'_j$ holds $B_j\widetilde{B'_s} \subseteq B_sB_j\widetilde{B'_s}$. Sketch of proof in Appendix A.4.*

If Sarah is not allowed to utter inform statements to John, then Sarah will be aware that John has represent her ignorance correctly if what she believes that John believes about her beliefs is actually not believed by her.

Proposition 7.5. *If Sarah is not allowed to utter an inform statement to John that she believes ψ, i.e. $\mathcal{M}_s \not\models pre(is(s, j, \psi, B_s))$, then Sarah's beliefs about John's about her ignorance are correct, i.e. $\mathcal{M}_s \models \psi \in B_sB_j\widetilde{B_s} \Rightarrow \mathcal{M}_s \models \psi \notin B_s$. Proof in Appendix A.4.*

Another important property is that John's beliefs about Sarah's ignorance are correct. John is said to have correctly represented Sarah's ignorance if what he believes about Sarah's beliefs is in fact not believed by Sarah. This property only holds if $B_j\widetilde{B_s} \subseteq B_sB_j\widetilde{B_s}$.

Proposition 7.6. *If Sarah is not allowed to utter an inform statement to John that she believes ψ, i.e. $\mathcal{M}_s \not\models pre(is(s, j, \psi, B_s))$, and $B_j\widetilde{B_s} \subseteq B_sB_j\widetilde{B_s}$, then John's beliefs about Sarah's ignorance are correct, i.e. $\mathcal{M}_j \models \psi \in B_j\widetilde{B_s} \Rightarrow \mathcal{M}_s \models \psi \notin B_s$. Proof in Appendix A.4.*

7.7 Run Out of Communicative Acts

Next, we provide the property of Sarah's mental states in which she has completely run out of communicative acts to become explicitly justified to decide to believe or to decide to be ignorant about a proposition.

Informally, Sarah has run out of questions for a belief addition (with a subscript 1) regarding proposition ψ if she may not utter a qba_1 and all her uttered qba_1s have been answered. Sarah may not utter a $qba_1(s, a, \psi)$ with $a \in \mathcal{A} \setminus s$, i.e. $\forall a \in \mathcal{A} \setminus s\ \mathcal{M}_s \not\models pre(qba_1(s, a, \psi))$. Sarah is said to be aware that if she has uttered a $qba_1(s, a, \psi)$, then it has been answered if $\mathcal{M}_s \models post(qba_1(s, a, \psi))$, and a has answered either positively, i.e. $\mathcal{M}_s \models post(gqba_1(a, s, \psi))$, or negatively, i.e. $\mathcal{M}_s \models post(dqba_1(a, s, \psi))$. Equation (7.75) expresses that Sarah has run out of qba_1s regarding ψ because she cannot utter a qba_1 to any of the agents in the multiagent system, and, as seen from Sarah's perspective, all her questions have been answered.

$$\forall a \in \mathcal{A} \setminus s \Big(\mathcal{M}_s \not\models pre(qba_1(s, a, \psi)) \wedge \Big(\mathcal{M}_s \models post(qba_1(s, a, \psi)) \Rightarrow \Big(\mathcal{M}_s \models post(gqba_1(a, s, \psi)) \vee \mathcal{M}_s \models post(dqba_1(a, s, \psi))\Big)\Big)\Big) \tag{7.75}$$

Similar to the situation in which Sarah has run out of qba_1s, Sarah has run out of qbr_1s in the following situation.

$$\forall a \in \mathcal{A} \setminus s \Big(\mathcal{M}_s \not\models pre(qbr_1(s, a, \psi)) \wedge \Big(\mathcal{M}_s \models post(qbr_1(s, a, \psi)) \Rightarrow \Big(\mathcal{M}_s \models post(gqbr_1(a, s, \psi)) \vee \mathcal{M}_s \models post(dqbr_1(a, s, \psi))\Big)\Big)\Big) \tag{7.76}$$

Sarah has run out of *rba*s in the following situation.

$$\forall a \in \mathcal{A} \setminus s \Big(\mathcal{M}_s \not\models pre(rba(s, a, \psi)) \wedge \Big(\mathcal{M}_s \models post(rba(s, a, \psi)) \Rightarrow \Big(\mathcal{M}_s \models post(grba_1(a, s, \psi)) \vee \mathcal{M}_s \models post(drba_1(a, s, \psi))\Big)\Big)\Big) \tag{7.77}$$

Sarah has run out of *rbr*s in the following situation.

$$\forall a \in \mathcal{A} \setminus s \Big(\mathcal{M}_s \not\models pre(rbr(s, a, \psi)) \wedge \Big(\mathcal{M}_s \models post(rbr(s, a, \psi)) \Rightarrow \Big(\mathcal{M}_s \models post(grbr_1(a, s, \psi)) \vee \mathcal{M}_s \models post(drbr_1(a, s, \psi))\Big)\Big)\Big) \tag{7.78}$$

Sarah cannot become explicitly justified to believe ψ through communication if she has run out of communicative acts that may change her mental states such that she may become explicitly justified to decide to believe ψ.

1. Sarah has run out of qba_1s to become justified to believe ϕ, if for all partial cognitive preconditions $\psi \in B_s$ for Sarah to believe ϕ, i.e. $\psi \in B_s/\phi \in B_s$, holds that she has run out of qba_1s regarding ψ, as expressed in equation (7.75), for all agents in the multiagent system.

2. Sarah has run out of qbr_1s to become justified to believe ϕ, if for all partial cognitive preconditions $\psi \notin B_s$ for Sarah to believe ϕ, i.e. $\psi \notin B_s/\phi \in B_s$, holds that she has run out of qbr_1s regarding ψ, as expressed in equation (7.76), for all agents in the multiagent system.

3. Sarah has run out of *rba*s to become justified to believe ϕ, if for all partial cognitive preconditions $\psi \in B_sB_a$ for Sarah to believe ϕ, i.e. $\psi \in B_sB_a/\phi \in B_s$, holds that she has run out of rba_1s regarding ψ, as expressed in equation (7.77), for all agents in the multiagent system.

4. Sarah has run out of *rbr*s to become justified to believe ϕ, if for all partial cognitive preconditions $\psi \in B_s\widetilde{B_a}$ for Sarah to believe ϕ, i.e. $\psi \in B_s\widetilde{B_a}/\phi \in B_s$, holds that she has run out of rbr_1s regarding ψ, as expressed in equation (7.78), for all agents in the multiagent system.

This description that Sarah has run out of communicative acts to become justified to believe ψ, is formalised with condition $\mathit{finished}(s, \psi \in B_s)$. In analogous fashion, $\mathit{finished}(s, \psi \notin B_s)$ formalises that Sarah has run out of communicative acts to become justified to be ignorant about ψ. In the following chapter, we will also use $\mathit{finished}(s, \psi \in B_sB_j)$ and $\mathit{finished}(s, \psi \in B_s\widetilde{B_j})$ to describe that Sarah has run out of communicative acts to convince John to believe ψ or convince him to be ignorant about ψ respectively.

$$
\begin{aligned}
&\mathit{finished}(s, \pi) \Leftrightarrow \\
&\qquad \mathcal{M}_s \models (\phi \in B_s/\pi) \Rightarrow (7.75) \wedge \\
&\qquad \mathcal{M}_s \models (\phi \notin B_s/\pi) \Rightarrow (7.76) \wedge \\
&\qquad \forall a \in \mathcal{A} \setminus s \left(\mathcal{M}_s \models (\phi \in B_sB_a/\pi) \Rightarrow (7.77)\right) \wedge \\
&\qquad \forall a \in \mathcal{A} \setminus s \left(\mathcal{M}_s \models (\phi \in B_s\widetilde{B_a}/\pi) \Rightarrow (7.78)\right)
\end{aligned}
\tag{7.79}
$$

7.8 Concluding Remarks

In order to communicate information, as described by speech act theory, the utterances that our agents use to communicate must have certain properties. The locutions that agents send to each other need to have, according to the agents, a use that is common knowledge. If a speaker regards the use of an utterance common knowledge, i.e. the rules of usage are shared, then the intended receiver of the utterance should be capable of recognising that the speaker has directed the utterance with the aim of inducing certain effects in the receiver. A dialogue game is a set of dialogue rules that describe when an agent may utter a communicative act conform to the shared use of the act, and the dialogue rules describe the changes to the agent's mental states after she has received or uttered the communicative act.

We defined the communicative act in which agents ask whether they may decide to believe propositions, i.e. speech acts *qba*, *gqba* and *dqba*. The agent utters such communicative acts with the intent to achieve a state in which she is explicitly justified to add a belief. A receiving agent interprets these communicative acts, and updates her mental states accordingly, after which she may respond. An agent may respond with a positive or with a negative answer. The effects of these communications are that a chain of justifications is created which is distributed over the mental state structures of the dialogue participants. That is to say, according to the agent's mental state structure, a proposition may be viewed to be justified based on the mental state structures of other agents.

Three other dialogue games provide the semantics of communicative act in which agents ask whether they may decide to be ignorant about propositions, i.e. speech acts *qbr*, *gqbr* and *dqbr*. This game has the reverse effect of the game with questions for a belief addition: the agent intends to achieve a state in which she is explicitly justified to retract a belief. A distributed chain of justification is traversed backwards recursively when agents ask others whether they may decide to be ignorant about propositions.

The dialogue game with requests for a belief addition, i.e. speech acts *rba*, *grba* and *drba*, provides the semantics of communicative acts in which agents request others whether they will decide to believe propositions. These acts are uttered with the intent that the receiver changes her cognitive such that she becomes explicitly justified to believe a certain proposition. The dialogue game with requests for a belief retraction, i.e. speech acts *rbr*, *grbr* and *drbr*, provides the semantics of communicative acts in which agents request other whether they will decide to be ignorant about propositions. This game has the reverse effect of the game with requests for a belief addition: the agent intends to achieve a state in which she believes that the intended receiver is explicitly justified to be ignorant about a particular proposition. In an additional dialogue game, the agent will inform other agents about her changed mental states.

Our agents may utter communicative acts according to the dialogue rules that the agents are entitled to know. In the following chapter, we assume that our agents utter communicative acts with a semantics as defined in this chapter. That is to say, the agents' set of dialogue rules consists of the dialogue rules with the preconditions and post-conditions as defined in this chapter.

Chapter 8

Agreeing to Disagree

> *The scientific theory I like best is that the rings of Saturn are composed entirely of lost airline luggage.*
>
> Mark Russell [Ros94, p. 439]

Some propositions and theories can be very appealing and seem to follow straightforwardly from established truths and opinions. We may easily integrate these propositions and theories with our thoughts, while others may not. For example, we could believe that the rings of Saturn are entirely composed of lost airline luggage; others could believe this to be absolute nonsense. Whatever we believe, humans and agents may very likely have disagreements on what to believe. Most of the time our disagreements can be resolved; however, occasionally disagreements may seem irresolvable. Would such an irresolvable disagreement be a bad sign? Would an irresolvable disagreement on the lost luggage influence our judgement whether a patient suffers from a depression? We may assume it would most likely not. It would be useful just to set aside irresolvable disagreements and focus attention on solvable problems. This setting aside is what we seem to be doing when we are tempted to quarrel over the question what tasty food is and what not. "There's no accounting for taste" is a saying that addresses the tacit agreement that disputes about taste are useless; this because we seem to acknowledge that differences in taste exist and cannot be resolved. We seem to agree that we may disagree about taste. Our agents should be made capable, in the same line of thought, to agree that irresolvable disagreements are irresolvable, that is, our agents should be made capable of agreeing to disagree.

First, in Section 8.1 we will define when an agent has a disagreement with another agent about the truth value of a proposition, and when an agent is justified to believe that she has a disagreement. Additionally, we describe how agents should change their beliefs to resolve disagreements. We turn in Section 8.2 to provide inference rules that allow the agents to resolve their disagreements. In Section 8.3 we describe when an agent, if resolving their disagreement proves impossible, is justified to believe that she has an irresolvable disagreement. We provide several inference rules that allow our agents to communicate their irresolvable disagreements with the intent to resolve them. This communication may result in an agreement to disagree. In Section 8.4 we discuss that, according to Aumann's 'no disagreement' theorem, agents cannot agree to disagree. We discuss how this theorem adds to the interpretation of our agent's agreement to disagree. We provide concluding remarks in Section 8.5. This chapter is partly based on published work by Lebbink et al. [LWM03b, LWM03c, LWM04a, LWM04b].

8.1 Disagreements

Agents can disagree over what to believe because they can have different beliefs. Another truism is that these disagreements are either resolvable or irresolvable. This section addresses what it takes for propositions to be contradictory and for agents to disagree, when agents can be said to be aware of disagreements, and how agents may resolve their disagreements.

8.1.1 Contradictory propositions

Two propositions are said to differ if their formulae or their assigned truth values differ. For example, propositions $p : \mathsf{t}$ and $q : \mathsf{t}$ are different because $p \neq q$, and propositions $p : \mathsf{t}$ and $p : \mathsf{u}$ are different because $\mathsf{t} \neq \mathsf{u}$. Two propositions are said to contradict if and only if they have equal formulae and the information that is represented by their truth values contradicts. Two pieces of information are said to contradict if the corresponding truth values are not comparable in order $\leq_k$, i.e. the truth values θ_1 and θ_2 contradict if and only if $\theta_1 \not\gtrless_k \theta_2$. In a four-valued (multi-valued) logic only truth value t contradicts with f and *vice versa*. In a nine-valued logic, several combinations of truth values contradict. For example, truth value $0 \times {}^1\!/_2$ contradicts with truth value 1×0 and ${}^1\!/_2 \times 0$.

Two propositions do not contradict if and only if they have equal formulae and the information represented by their truth value is non-contradictory. Propositions $p : \theta_1$ and $p : \theta_2$ do not contradict if and only if $\theta_1 \gtrless_k \theta_2$. For example, the propositions $p : \mathsf{u}$ and $p : \mathsf{i}$ do not contradict with all other propositions $p : \theta$ with a truth value θ from any bilattice.

8.1.2 Disagreeing and agreeing agents

If Sarah and John believe that the formula of a proposition. is assigned *at most* a truth value that is contradictory, then they disagree about the truth value of the proposition. Sarah and John disagree about a proposition with formula p if, for example, Sarah believes that p is true and John believes that p is false, and Sarah and

John do not believe more about p. For example, if Sarah believes that p has *at least* truth value t, i.e. $p : \mathsf{t} \in B_s$, then she can still believe that p is inconsistent, i.e. $p : \mathsf{i} \in B_s$. If Sarah believes that p is inconsistent, then she also believes that p has at least truth values t and f. If Sarah also believes that p has a least truth value f, then she does *not* disagree with John over p. Thus, the truth values of formula p that Sarah believes *at most* should be contradictory with the truth value of p that John believes *at most*. Sarah believes that formula p is assigned *at most* truth value θ if $p : \theta$ is part of the set of k-maximal elements of her belief state (def. 4.8), i.e. $p : \theta \in max_k(B_s)$. Thus, as given in equation (8.1), Sarah and John disagree about a proposition with formula p if Sarah believes that p has *at most* truth value θ_1, and John believes that p has *at most* truth value θ_2, and θ_1 and θ_2 are contradictory.

$$(\mathcal{M}_s \models p : \theta_1 \in max_k(B_s)) \wedge (\mathcal{M}_j \models p : \theta_2 \in max_k(B_j)) \wedge (\theta_1 \not\geq_k \theta_2) \tag{8.1}$$

Sarah and John are said to agree about a proposition if they believe the proposition's formula to have at most a truth value that does not contradict.

Example 8.1. *Next we provide the situation in which Sarah and John disagree over the truth value of p. Assume that Sarah and John use a bilattice with* $11^2 = 121$ *distinct truth values, and thus,* $\mathsf{t} = {}^{10}/_{10} \times {}^{0}/_{10}$, $\mathsf{f} = {}^{0}/_{10} \times {}^{10}/_{10}$, $\mathsf{u} = {}^{0}/_{10} \times {}^{0}/_{10}$, *and* $\mathsf{i} = {}^{10}/_{10} \times {}^{10}/_{10}$.

1. *Assume that Sarah believes that a certain proposition or event p is in* 70% *true and in* 30% *false, and that John believes that p is in* 30% *true and in* 70% *false. We represent Sarah's and John's beliefs as* $p : {}^{7}/_{10} \times {}^{3}/_{10} \in max_k(B_s)$, *and* $p : {}^{3}/_{10} \times {}^{7}/_{10} \in max_k(B_j)$. *Thus, suppose:* $\mathcal{M}_s$ *with* $B_s \in \mathcal{M}_s$ *with* $p : \theta_1 \in max_k(B_s)$, *and* $\mathcal{M}_j$ *with* $B_j \in \mathcal{M}_j$ *with* $p : \theta_2 \in max_k(B_j)$.
2. *If* $\theta_1 = {}^{7}/_{10} \times {}^{3}/_{10}$ *and* $\theta_2 = {}^{3}/_{10} \times {}^{7}/_{10}$, *then, because* $\theta_1 \not\geq_k \theta_2$, *Sarah and John disagree about the truth value of p.*

Equation (8.1) describes the criterion when *we* can observe that Sarah and John disagree over the truth value of a formula. In the following section (Section 8.1.3), we will describe the agents' decisions that may resolve their disagreements. However, because we did not describe the criterion when agents can themselves observe that they disagree over the truth value of a formula, we will in Section 8.1.4 provide the criterion when agents are said to be aware of their disagreements. Based on such a disagreement-awareness, an agent may act with the aim of resolving their disagreement (Section 8.1.5).

8.1.3 Resolving disagreements

A disagreement between Sarah and John can be resolved in different ways: either Sarah or John adds a proposition, or one retracts a proposition, such that the disagreement is resolved. Combinations in which both Sarah and John add a proposition, or one agent retracts and the other adds a proposition, or one agent both adds and retracts propositions to resolve the disagreement, will not be considered. Assume as

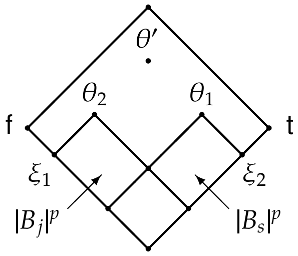

Figure 8.1: Resolve a disagreement by adding beliefs ($\theta' = \theta_1 \oplus_k \theta_2$).

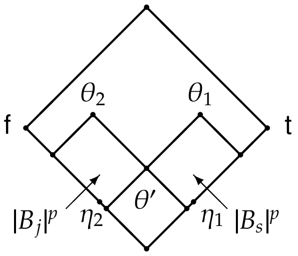

Figure 8.2: Resolve a disagreement by retracting beliefs ($\theta' = \theta_1 \otimes_k \theta_2$).

in equation (8.1) that Sarah disagrees with John about the truth value of formula p. We discuss four different ways to resolve their disagreement.

The disagreement is resolved if either Sarah or John decides to believe at least $p:(\theta_1 \oplus_k \theta_2)$. For all $\theta' \in \mathcal{B}$ we have $\theta \leq_k (\theta' \oplus_k \theta)$, thus $\theta_1 \gneqq_k (\theta_1 \oplus_k \theta_2)$. John's decision to believe $p:(\theta_1 \oplus_k \theta_2)$ will resolve their disagreement. Analogously, Sarah's decision to believe $p:(\theta_1 \oplus_k \theta_2)$ will also resolve their disagreement. See figure 8.1 for a depiction of the truth values.

1. To make Sarah believe at least $p:(\theta_1 \oplus_k \theta_2)$, it suffices if she decides to believe a certain $p:\xi_1$ which, combined with her current belief $p:\theta_1$, will make her believe $p:(\theta_1 \oplus_k \theta_2)$. If Sarah decides to believe $p:\xi_1$ with $(\theta_1 \oplus_k \xi_1) = (\theta_1 \oplus_k \theta_2)$ then, due to the closure of her belief state under rule R3 from Section 4.3.1, she will come to believe $p:(\theta_1 \oplus_k \theta_2)$. However, different values for ξ_1 will resolve the disagreement. Sarah will only be interested in the smallest truth value with respect to order $\leq_k$ that resolves the disagreement. Truth value ξ_1 is the abbreviation for the truth value with the least amount of information, i.e. the k-meet of all such truth values ξ_1, that resolve the disagreement.

$$\xi_1 \equiv \bigotimes_k \Big(\{\theta' \in \mathcal{B} \mid (\theta_1 \oplus_k \theta') = (\theta_1 \oplus_k \theta_2)\}\Big) \tag{8.2}$$

Proposition 8.1. *A disagreement between Sarah and John (eq. (8.1)) will be resolved if Sarah decides to believe $p:\xi_1$ (eq. (8.2)). Proof in Appendix A.5.*

2. An analogous argument holds for John. If John is made to believe a least $p:(\theta_1 \oplus_k \theta_2)$, he has to decide to believe at least a certain $p:\xi_2$ with $(\theta_2 \oplus_k \xi_2) = (\theta_1 \oplus_k \theta_2)$. Deciding to believe $p:\xi_2$, combined with his current belief $p:\theta_2$ and due to the closure of B_j under rule R3, will make him believe $p:(\theta_1 \oplus_k \theta_2)$. John is also only interested in the truth value with the least possible amount of information that resolves the disagreement. Truth value ξ_2 is the abbreviation

for the truth value with the least amount of information, i.e. the k-meet of all such truth values ξ_2, that resolve the disagreement.

$$\xi_2 \;\equiv\; \bigotimes_k \Big(\{\theta' \in \mathcal{B} \mid (\theta_2 \oplus_k \theta') = (\theta_1 \oplus_k \theta_2)\}\Big) \tag{8.3}$$

Proposition 8.2. *A disagreement between Sarah and John (eq. (8.1)) will be resolved if John decides to believe $p : \xi_2$ (eq. (8.3)). Proof in Appendix A.5.*

Example 8.2. *Next, we provide example truth values that Sarah may use to resolve her disagreement with John. Continued from example 8.1. We have: $\mathcal{M}_s$ with $B_s \in \mathcal{M}_s$ with $p : \theta_1 \in max_k(B_s)$ with $\theta_1 = {}^7\!/_{10} \times {}^3\!/_{10}$, and $\mathcal{M}_j$ with $B_j \in \mathcal{M}_j$ with $p : \theta_2 \in max_k(B_j)$ with $\theta_2 = {}^3\!/_{10} \times {}^7\!/_{10}$.*

1. *We have $\xi_1 = {}^0\!/_{10} \times {}^7\!/_{10}$, e.g. $\mathcal{M}'_s = add_{ms}(p : \xi_1, B_s)(\mathcal{M}_s)$ with $B'_s \in \mathcal{M}'_s$ with $p : \theta'_1 \in max_k(B'_s)$ with $\theta'_1 = {}^7\!/_{10}\times{}^3\!/_{10}\oplus_k{}^0\!/_{10}\times{}^7\!/_{10} = {}^7\!/_{10}\times{}^7\!/_{10}$. Because $\theta_2 \leq_k \theta'_1$, the disagreement is resolved.*

2. *We have $\xi_2 = {}^7\!/_{10} \times {}^0\!/_{10}$, e.g. $\mathcal{M}'_s = add_{ms}(p : \xi_2, B_j)(\mathcal{M}_j)$ with $B'_j \in \mathcal{M}'_j$ with $p : \theta'_2 \in max_k(B'_s)$ with $\theta'_2 = {}^3\!/_{10}\times{}^7\!/_{10}\oplus_k{}^7\!/_{10}\times{}^0\!/_{10} = {}^7\!/_{10}\times{}^7\!/_{10}$. Because $\theta_1 \leq_k \theta'_2$, the disagreement is resolved.*

The propositions $p : \xi_1$ and $p : \xi_2$ are the propositions with the least amount of information that resolve the disagreement. Both agents could decide to believe propositions that are more informative; however, to resolve their issues, we assume that agents change their beliefs as little as possible.

Retracting beliefs can have the same effect as the opposite decision of adding beliefs: a disagreement can be resolved if either Sarah or John retracts certain beliefs such that in their remaining belief state they come to believe at most $p : (\theta_1 \otimes_k \theta_2)$. Proposition $p : (\theta_1 \otimes_k \theta_2)$ is the most Sarah may believe to agree with John, and $p : (\theta_1 \otimes_k \theta_2)$ is also the most John may believe to agree with Sarah. See figure 8.2 for a depiction.

3. If Sarah is to believe at most $p : (\theta_1 \otimes_k \theta_2)$, then she has to become ignorant about a certain $p : \eta_1$. Truth value η_1 has less information than θ_1, i.e. $\eta_1 \leq_k \theta_1$; otherwise Sarah would not retract anything. Additionally, η_1 does not have less information than $(\theta_1 \otimes_k \theta_2)$, i.e. $\eta_1 \nleq_k (\theta_1 \otimes_k \theta_2)$; otherwise Sarah would retract more than necessary. Truth value η_1 is the abbreviation for the truth value with the least amount of information, i.e. the k-meet of all such truth values, that apply for both criteria.

$$\eta_1 \;\equiv\; \bigotimes_k \Big(\{\theta' \in \mathcal{B} \mid (\theta' \leq_k \theta_1) \wedge (\theta' \nleq_k (\theta_1 \otimes_k \theta_2))\}\Big) \tag{8.4}$$

Proposition 8.3. *A disagreement between Sarah and John (eq. (8.1)) will be resolved if Sarah decides to be ignorant about $p : \eta_1$ (eq. (8.4)). Proof in Appendix A.5.*

4. An analogous argument holds for John. If John is to believe at most $p : (\theta_1 \otimes_k \theta_2)$, he has to become ignorant about a certain $p : \eta_2$. Truth value η_2 has less information than θ_2, i.e. $\eta_2 \leq_k \theta_2$; otherwise John would not retract anything. Additionally, η_2 does not have less information than $\theta_1 \otimes_k \theta_2$, i.e. $\eta_2 \nleq_k (\theta_1 \otimes_k \theta_2)$; otherwise John would retract more than necessary. Truth value η_2 is the abbreviation for the truth value with the least amount of information, i.e. the k-meet of all such truth values, that apply for both criteria. See figure 8.2 for a depiction.

$$\eta_2 \equiv \bigotimes_k \Big(\{\theta' \in \mathcal{B} \mid (\theta' \leq_k \theta_2) \wedge (\theta' \nleq_k (\theta_1 \otimes_k \theta_2))\} \Big) \tag{8.5}$$

Proposition 8.4. *A disagreement between Sarah and John (eq. (8.1)) will be resolved if John decides to be ignorant about $p : \eta_2$ (eq. (8.5)). Proof in Appendix A.5.*

Example 8.3. *Next, we provide example truth values that Sarah may use to resolve her disagreement with John. Continued from example 8.1. We have: $\mathcal{M}_s$ with $B_s \in \mathcal{M}_s$ with $p : \theta_1 \in max_k(B_s)$ with $\theta_1 = {}^7\!/_{10} \times {}^3\!/_{10}$, and $\mathcal{M}_j$ with $B_j \in \mathcal{M}_j$ with $p : \theta_2 \in max_k(B_j)$ with $\theta_2 = {}^3\!/_{10} \times {}^7\!/_{10}$.*

1. *We have $\eta_1 = {}^4\!/_{10} \times {}^0\!/_{10}$, e.g. $\mathcal{M}'_s = retr_{ms}(p : \eta_1, B_s)(\mathcal{M}_s)$ with $B'_s \in \mathcal{M}'_s$ with $\theta'_1 \in max_k(B'_s)$ with $\theta'_1 = {}^3\!/_{10} \times {}^3\!/_{10}$. Because $\theta'_1 \leq_k \theta_2$, the disagreement is resolved.*

2. *We have $\eta_2 = {}^0\!/_{10} \times {}^4\!/_{10}$, e.g. $\mathcal{M}'_j = retr_{ms}(p : \eta_2, B_j)(\mathcal{M}_j)$ with $B'_j \in \mathcal{M}'_j$ with $\theta'_2 \in max_k(B'_j)$ with $\theta'_2 = {}^3\!/_{10} \times {}^3\!/_{10}$. Because $\theta'_2 \leq_k \theta_1$, the disagreement is resolved.*

8.1.4 Disagreement-awareness

Disagreeing agents can become aware of their disagreement, as we will describe now. We assume that our agents' mental states are closed under communication of inform statements (Section 7.6). Given this assumption and proposition 7.2, it follows that Sarah is aware that John has represented her beliefs correctly, and with proposition 7.5, it follows that Sarah is aware that John has represented her ignorance correctly. Additionally, given this assumption and proposition 7.3, it follows that Sarah's beliefs about John's beliefs are correct, and with proposition 7.6, it follows that Sarah's beliefs about John's ignorance are also correct. If we would not assume that agents are aware that their beliefs are correctly represented, we had to describe situations in which Sarah would be aware that John is incorrectly aware of a disagreement with her, or situations in which Sarah would be aware that John is incorrectly not aware of disagreement with her. Given the assumption, three different levels of disagreement-awareness are identified.

1. The lowest level of disagreement-awareness is that Sarah disagrees with John about the truth value of a formula; however, she is not aware they disagree. This unawareness is solely because Sarah does not have beliefs about John's beliefs from which she can become aware of their disagreement.

2. The first level of a disagreement-awareness is that Sarah is aware that she disagrees with John. Figure 8.3 depicts the following mental state structures, and equation (8.6) provides a sentence in $\mathcal{L}_{MS}$ that defines when Sarah has a first-order disagreement disagreement-awareness with John.

   - Sarah believes at most that p has truth value θ_1, i.e. $p:\theta_1 \in max_k(B_s)$.
   - Ideally, if Sarah believes that John believes that formula p has *at most* truth value θ_2 and θ_1 disagrees with θ_2, then she disagrees with John about p, i.e. $\theta_1 \not\gtrless_k \theta_2$. However, Sarah does *not* believe what John believes at most; she only believes what John believes *at least*. The best approximation that Sarah can have of what John believes at most is by the combined information from her beliefs about his beliefs and her beliefs about his ignorance. The hypothetical truth value of the formula p that John believes at most, say θ', as seen from Sarah's perspective, has equal or more information than the k-maximal elements of her beliefs about John's beliefs about p, i.e. $p:\theta_3 \in max_k(B_sB_j)$ and $\theta_3 \leq_k \theta'$.
   - Sarah's approximation of what John believes at most, proposition $p:\theta'$, is not part of Sarah's belief about John's ignorance. According to Sarah, her belief about John's beliefs provides a lower bound to what he believes at most, and her beliefs about his ignorance provides an upper bound to what he believes at most. Sarah may believe at most that John believes at least $p:\theta_3$, i.e. $p:\theta_3 \in max_k(B_sB_j)$. If a proposition $p:\tau_2$ that would resolve the disagreement, if believed by John or Sarah, i.e. $\tau_2 \leq_k (\theta_1 \oplus_k \theta_3)$, is not believed by John, i.e. $p:\tau_2 \in B_s\widetilde{B}_j$, and θ_1 disagrees with θ_3, i.e. $\theta_1 \not\gtrless_k \theta_3$, then Sarah is aware of a disagreement with John about the truth value of formula p.

   Sarah is aware of the first-order disagreement if $\mathcal{M}_s \models$ equation (8.6) with $\theta_1 \not\gtrless_k \theta_3$ and $\tau_2 \leq_k (\theta_1 \oplus_k \theta_3)$.

$$(p:\theta_1 \in max_k(B_s)) \wedge (p:\theta_3 \in max_k(B_sB_j)) \wedge (p:\tau_2 \in B_s\widetilde{B}_j) \qquad (8.6)$$

Proposition 8.5. *If Sarah has a first-order disagreement-awareness with John (eq. (8.6)), then she has a disagreement with John (eq. (8.1)). Proof in Appendix A.5.*

Example 8.4. *Next, we provide the situation in which Sarah has a first-order disagreement-awareness with John about the truth value of p. Continued from example 8.1. Suppose:* $\mathcal{M}_s$ *with* $B_s, B_sB_j, B_s\widetilde{B}_j \in \mathcal{M}_s$ *with* $p:\theta_1 \in max_k(B_s)$, $p:\theta_3 \in max_k(B_sB_j)$ *and* $p:\tau_2 \in B_s\widetilde{B}_j$.

1. *If* $\theta_1 = {}^7/_{10} \times {}^3/_{10}$ *and* $\theta_3 = {}^3/_{10} \times {}^3/_{10}$, *then because* $\theta_3 \leq_k \theta_1$ *Sarah does not have a first-order disagreement-awareness with John about the truth value of p.*
2. *If* $\theta_1 = {}^7/_{10} \times {}^3/_{10}$ *and* $\theta_3 = {}^2/_{10} \times {}^6/_{10}$, *then because* $\theta_3 \not\gtrless_k \theta_1$ *Sarah believes that John has a belief that disagrees with her belief. Additionally, if* $\tau_2 = {}^4/_{10} \times {}^6/_{10}$, *then she*

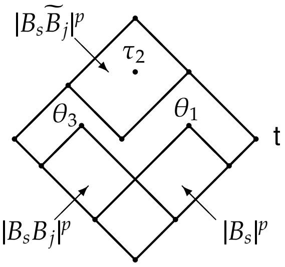

Figure 8.3: disagreement-awareness about formula p.

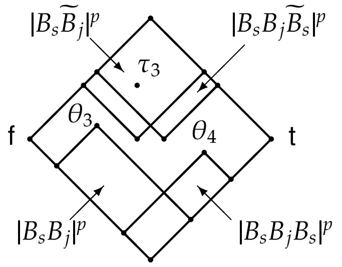

Figure 8.4: Awareness of disagreement-awareness about formula p.

believes that John cannot have a belief in which they agree, i.e. $\tau_2 \leq_k (\theta_1 \oplus_k \theta_3)$, i.e. $^4/_{10} \times {}^6/_{10} \leq_k {}^7/_{10} \times {}^6/_{10}$. Thus, Sarah has a first-order disagreement-awareness with John about the truth value of p.

3. The second level of a disagreement-awareness is described by the situation in which Sarah is aware that John is aware that they disagree. Figure 8.4 depicts the following mental state structures, and equation (8.7) provides a sentence in $\mathcal{L}_{MS}$ that defines when Sarah has a second-order disagreement-awareness with John.

   - Ideally, if Sarah believes that John believes that p has *at most* truth value θ_3, and that John believes that she believes that p has *at most* a truth value θ_4 that disagrees with θ_3, then she is aware that John is aware that they disagree. Just as in to the previous paragraph, we have $p\colon \theta_3 \in max_k(B_sB_j)$.
   - Sarah believes at most that John believes at least that she believes at least $p : \theta_4$, i.e. $p : \theta_4 \in max_k(B_sB_jB_s)$. Note that θ_4 need not be equal to θ_1 that Sarah actually assigns to p (eq. (8.1)); Sarah may deliberately keep John ignorant about what she actually believes, i.e. $\theta_4 \leq_k \theta_1$. Thus, as seen from Sarah's perspective, John is aware that they disagree if $\theta_3 \nleq_k \theta_4$. The proposition that would, from John's perspective as seen from Sarah's perspective, resolve the disagreement is $p : (\theta_3 \oplus_k \theta_4)$.
   - As seen from Sarah's perspective, John does not believe the proposition $p : (\theta_3 \oplus_k \theta_4)$ that would resolve the disagreement, i.e. $p : \tau_2 \in B_s\widetilde{B_j}$ with $\tau_2 \leq_k (\theta_3 \oplus_k \theta_4)$.
   - Additionally, Sarah believes that John believes that she is ignorant about the proposition that would resolve the second-order disagreement-awareness, i.e. $p : \tau_3 \in B_sB_j\widetilde{B_s}$ with $\tau_3 \leq_k (\theta_3 \oplus_k \theta_4)$.

Sarah is aware of the second-order disagreement if $\mathcal{M}_s \models$ equation (8.7) with $\theta_3 \nleq_k \theta_4$, $\tau_2 \leq_k (\theta_3 \oplus_k \theta_4)$ and $\tau_3 \leq_k (\theta_3 \oplus_k \theta_4)$.

$$\begin{aligned}&(p : \theta_3 \in max_k(B_sB_j)) \wedge (p : \theta_4 \in max_k(B_sB_jB_s)) \wedge \\ &(p : \tau_2 \in B_s\widetilde{B_j}) \wedge (p : \tau_3 \in B_sB_j\widetilde{B_s})\end{aligned} \tag{8.7}$$

Proposition 8.6. *If Sarah has a second-order disagreement-awareness with John (eq. (8.7)), then Sarah has a first-order disagreement-awareness with John (eq. (8.6)). Proof in Appendix A.5.*

Example 8.5. *Next, we provide the situation in which Sarah has a second-order disagreement-awareness with John about the truth value of p. Continued from example 8.4. Suppose:* $\mathcal{M}_s$ *with* $B_s, B_sB_j, B_s\widetilde{B}_j, B_sB_jB_s$, $B_sB_j\widetilde{B}_s \in \mathcal{M}_s$ *with* $p : \theta_1 \in max_k(B_s)$, $p : \theta_3 \in max_k(B_sB_j)$, $p : \theta_4 \in max_k(B_sB_jB_s)$, $p : \tau_2 \in B_s\widetilde{B}_j$ *and* $p : \tau_3 \in B_sB_j\widetilde{B}_s$.

1. *Assume* $\theta_1 = {}^7\!/_{10} \times {}^3\!/_{10}$, $\theta_3 = {}^2\!/_{10} \times {}^6\!/_{10}$ *and* $\tau_2 = {}^4\!/_{10} \times {}^6\!/_{10}$, *then Sarah has a first-order disagreement-awareness with John about the truth value of p, see example 8.4(2).*

2. *If* $\theta_4 = {}^2\!/_{10} \times {}^1\!/_{10}$, *then because* $\theta_4 \leq_k \theta_3$ *Sarah is not aware that John is aware of their disagreement, i.e. Sarah does not have a second-order disagreement-awareness about the truth value of p.*

3. *If* $\theta_4 = {}^5\!/_{10} \times {}^1\!/_{10}$, *i.e.* $\theta_3 \not\geq_k \theta_4$, *and if* $\tau_3 = {}^5\!/_{10} \times {}^4\!/_{10}$, *i.e.* $\tau_3 \leq_k (\theta_3 \oplus_k \theta_4)$, *i.e.* ${}^5\!/_{10} \times {}^4\!/_{10} \leq_k {}^5\!/_{10} \times {}^6\!/_{10}$, *then Sarah has a second-order disagreement-awareness with John about the truth value of p.*

Additional situations in which either Sarah's or John's beliefs are incorrect, or situations in which further levels of awareness, such as third-order disagreement-awareness, would emerge, are not needed for our agents, and are not discussed.

8.1.5 Resolving disagreement-awareness

Next, we address propositions that resolve a disagreement-awareness. Assume that Sarah is aware that John is aware that they disagree about the truth value of formula p (eq. (8.7)). As described in Section 8.1.3, Sarah resolves the disagreement if she decides to believe $p : \xi_1$ (eq. (8.2)). Because Sarah does not have access to John's true beliefs, and thus does not believe what John believes *at most*, she does not have access to $p : \xi_1$; consequently, she cannot resolve the disagreement itself. The best Sarah can do is to resolve the disagreement that she can become aware of. As far as Sarah can tell, she has resolved the disagreement with John if she has resolved her first-order disagreement-awareness. Because John may not have informed Sarah about his true beliefs, Sarah can be aware that she has resolved the disagreement, while John is aware that he still has a first-order disagreement with Sarah.

In the following discussion, four situations are distinguished, and three special propositions are defined that, as seen from Sarah's perspective, resolve her disagreement-awareness. Table 8.1 summarises the four situations and figure 8.5 depicts the truth values of the special propositions.

1. Sarah resolves her first-order disagreement-awareness if she comes to believe at least $p : (\theta_1 \oplus_k \theta_3)$. Because Sarah does not have access to John's true beliefs, i.e. $p : \theta_2 \in max_k(B_j)$, and only has beliefs about his beliefs, i.e. $p : \theta_3 \in max_k(B_sB_j)$, deciding to believe $p : (\theta_1 \oplus_k \theta_3)$ resolves Sarah's first-order awareness of the

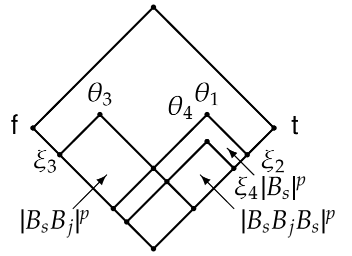

Figure 8.5: Resolving disagreement-awareness about p by adding beliefs.

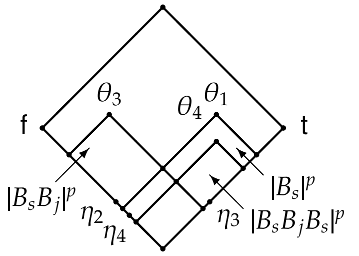

Figure 8.6: Resolving disagreement-awareness about p by retracting beliefs.

disagreement. To make Sarah believe at least $p : (\theta_1 \oplus_k \theta_3)$, it suffices if she decides to believe a certain $p : \xi_3$ which, combined with her current belief $p : \theta_1$, would make her believe $p : (\theta_1 \oplus_k \theta_3)$. Truth value ξ_3 is the abbreviation for the truth value with the least amount of information that resolves her first-order disagreement-awareness.

$$\xi_3 \equiv \bigotimes_k \Big(\{\theta' \in \mathcal{B} \mid (\theta_1 \oplus_k \theta') = (\theta_1 \oplus_k \theta_3)\}\Big) \tag{8.8}$$

Proposition 8.7. *Sarah's first-order disagreement-awareness with John (eq. (8.6)) will be resolved if Sarah decides to believe $p : \xi_3$ (eq. (8.8)). Proof in Appendix A.5.*

2. Sarah resolves her second-order disagreement-awareness if she comes to believe a certain $p : \theta'$ and communicates to John that she believes $p : \xi_3$, e.g. with the communicative act $is(s, j, p : \xi_3, B_s)$, and thus adds to her beliefs that John has about her beliefs, that she believes $p : \xi_3$, i.e. $add_{ms}(p : \xi_3, B_sB_jB_s)(\mathcal{M}_s)$.

Proposition 8.8. *Sarah's second-order disagreement-awareness with John (eq. (8.6)) will be resolved if Sarah decides to believe $p : \xi_3$ (eq. (8.8)). Proof in Appendix A.5.*

3. From Sarah's perspective, John resolves Sarah's first-order disagreement-awareness if John decides to believe $p : \xi_2$ (eq. (8.3)). If John were to come to believe $p : \xi_2$ and thus, from Sarah's perspective, come to believe $p : (\theta_1 \oplus_k \theta_3)$ (cf. eq. (8.6)), then John would resolve Sarah's first-order disagreement-awareness.

Proposition 8.9. *Sarah's first-order disagreement-awareness with John (eq. (8.6)) will be resolved if John decides to believe $p : \xi_2$ (eq. (8.3)). Proof in Appendix A.5.*

4. From Sarah's perspective, John resolves Sarah's second-order disagreement-awareness if John comes to believe a certain $p : (\theta_3 \oplus_k \theta_4)$. Sarah may want to keep John ignorant about her true beliefs; from her perspective, John only believes that she believes at least $p : \theta_4$, i.e. $p : \theta_4 \in max_k(B_sB_jB_s)$. As described in

Section 8.1.4, Sarah can become aware that John is aware of their disagreement, and thus, from Sarah's perspective, if John comes to believe $p:(\theta_3 \oplus_k \theta_4)$, then the second-order disagreement-awareness would be resolved (figure 8.4). Note that if John comes to believe $p:(\theta_3 \oplus_k \theta_4)$ then his first-order disagreement-awareness need *not* be resolved.

From Sarah's perspective, John need not decide to believe $p:(\theta_3 \oplus_k \theta_4)$, he only has to decide to believe a certain $p:\xi_4$ which, combined with Sarah's beliefs about his current beliefs $p:\theta_3 \in max_k(B_sB_j)$, would resolve his first-order disagreement-awareness.

$$\xi_4 \equiv \bigotimes_k \Big(\{\theta' \in \mathcal{B} \mid (\theta_3 \oplus_k \theta') = (\theta_4 \oplus_k \theta_3)\}\Big) \tag{8.9}$$

Proposition 8.10. *Sarah's second-order disagreement-awareness with John (eq. (8.7)) will be resolved if John decides to believe $p:\xi_4$ (eq. (8.9)). Proof in Appendix A.5.*

The difference between $p:\xi_2$ that John needs to decide to believe to resolve the first-order disagreement-awareness, and $p:\xi_4$ that John needs to decide to believe to resolve the second-order disagreement-awareness, reflects the difference between what Sarah believes at most, i.e. $p:\theta_1 \in max_k(B_s)$ and what, according to Sarah, John believes about Sarah's beliefs, i.e. $p:\theta_4 \in max_k(B_sB_jB_s)$. That is: $\xi_4 \leq_k \xi_2$ reflects $\theta_4 \leq_k \theta_1$.

Example 8.6. *Next, we provide example truth values that Sarah may use to resolve her disagreement-awareness with John. Continued from example 8.5. Suppose: $\mathcal{M}_s$ with B_s, B_sB_j, $B_s\widetilde{B}_j$, $B_sB_jB_s$, $B_sB_j\widetilde{B}_s \in \mathcal{M}_s$ with $p:\theta_1 \in max_k(B_s)$, $p:\theta_3 \in max_k(B_sB_j)$, $p:\theta_4 \in max_k(B_sB_jB_s)$, $p:\tau_2 \in B_s\widetilde{B}_j$ and $p:\tau_3 \in B_sB_j\widetilde{B}_s$.*

1. *Assume: $\theta_1 = {}^7\!/_{10} \times {}^3\!/_{10}$, $\theta_3 = {}^2\!/_{10} \times {}^6\!/_{10}$, $\theta_4 = {}^5\!/_{10} \times {}^1\!/_{10}$, $\tau_2 = {}^4\!/_{10} \times {}^6\!/_{10}$, $\tau_3 = {}^5\!/_{10} \times {}^4\!/_{10}$, i.e. Sarah has a second-order disagreement-awareness with John about the truth value of p, see example 8.5(3).*

2. *We have $\xi_3 = {}^0\!/_{10} \times {}^6\!/_{10}$, e.g. $\mathcal{M}'_s = add_{ms}(p:\xi_3, B_sB_jB_s)(\mathcal{M}_s)$ with $B_sB_jB'_s \in \mathcal{M}'_s$ with $p:\theta'_4 \in max_k(B_sB_jB'_s)$ with $\theta'_4 = {}^5\!/_{10} \times {}^1\!/_{10} \oplus_k {}^0\!/_{10} \times {}^6\!/_{10} = {}^5\!/_{10} \times {}^6\!/_{10}$. Because $\theta_3 \leq_k \theta'_4$, the second-order disagreement-awareness is resolved.*

3. *We have $\xi_4 = {}^5\!/_{10} \times {}^0\!/_{10}$, e.g. $\mathcal{M}'_s = add_{ms}(p:\xi_4, B_sB_j)(\mathcal{M}_s)$ with $B_sB'_j \in \mathcal{M}'_s$ with $p:\theta'_3 \in max_k(B_sB'_j)$ with $\theta'_3 = {}^2\!/_{10} \times {}^6\!/_{10} \oplus_k {}^5\!/_{10} \times {}^0\!/_{10} = {}^5\!/_{10} \times {}^6\!/_{10}$. Because $\theta_4 \leq_k \theta'_3$, the second-order disagreement-awareness is resolved.*

The opposite of deciding to add beliefs to resolve a disagreement-awareness, is that Sarah or John decide to retract certain beliefs such that they come to agree. Four situations are distinguished, and three special propositions defined that, as seen from Sarah's perspective, resolve her disagreement-awareness. Figure 8.6 depicts the truth values, and table 8.2 summarises the four situations.

Decision:	*Resolves:*
$d2a(s, p : \xi_3, B_s)$	Sarah's 1-order disagreement-awareness
$d2a(s, p : \xi_3, B_sB_jB_s)$	Sarah's 2-order disagreement-awareness
$d2a(s, p : \xi_2, B_sB_j)$	Sarah's 1-order disagreement-awareness
$d2a(s, p : \xi_4, B_sB_j)$	Sarah's 2-order disagreement-awareness

Table 8.1: Overview of the decisions to add beliefs that resolve Sarah's awareness of the disagreement.

1. Sarah resolves her first-order disagreement-awareness if she decides to be ignorant about propositions such that she comes to believe at most $p : (\theta_1 \otimes_k \theta_3)$. Because Sarah does not have access to John's true beliefs, i.e. $p : \theta_2 \in max_k(B_j)$, and only has beliefs about his beliefs, i.e. $p : \theta_3 \in max_k(B_sB_j)$, deciding to believe at most $p : (\theta_1 \otimes_k \theta_3)$ resolves Sarah's first-order awareness of the disagreement. Proposition $p : (\theta_1 \otimes_k \theta_3)$ is the most Sarah may believe if she is to agree with John, i.e. $(\theta_1 \otimes_k \theta_3) \gtrless_k \theta_3$.

   To come to believe at most $p : (\theta_1 \otimes_k \theta_3)$, Sarah must decide to be ignorant about a certain $p : \eta_3$. Truth value η_3 should have less information than θ_1, i.e. $\eta_3 \leq_k \theta_1$, and η_3 should not have less information than $(\theta_1 \otimes_k \theta_3)$, i.e. $\eta_3 \nleq_k (\theta_1 \otimes_k \theta_3)$, and η_3 is the smallest in order $\leq_k$.

$$\eta_3 \equiv \bigotimes_k \Big(\{\theta' \in \mathcal{B} \mid (\theta' \leq_k \theta_1) \wedge (\theta' \nleq_k (\theta_1 \otimes_k \theta_3))\}\Big) \tag{8.10}$$

Proposition 8.11. *Sarah's first-order disagreement-awareness with John (eq. (8.6)) will be resolved if Sarah decides to be ignorant about $p : \eta_3$ (eq. (8.10)). Proof in Appendix A.5.*

2. Sarah resolves her second-order disagreement-awareness if she retracts a belief such that she comes to believe at most $p : (\theta_3 \otimes_k \theta_4)$. Sarah believes at most that John believes that she believes that formula p has at least truth value θ_4, i.e. $\theta_4 \in max_k(B_sB_jB_s)$, while in fact Sarah believes at most $p : \theta_1$, i.e. $p : \theta_1 \in max_k(B_s)$, with $\theta_4 \leq_k \theta_1$.

   Because of $\theta_4 \leq_k \theta_1$ (proposition 7.2) we have $(\theta_3 \otimes_k \theta_4) \leq_k (\theta_1 \otimes_k \theta_3)$, and if Sarah is to come to believe at most $p : (\theta_3 \otimes_k \theta_4)$, then she may have to decide to retract more beliefs than necessary. If Sarah believes at most $p : (\theta_3 \otimes_k \theta_4)$, then she is unnecessarily ignorant about information, this because believing at most $(\theta_1 \otimes_k \theta_3)$ would also have resolved the disagreement. Consequently, if Sarah is to change her beliefs as little as possible, she is to resolve her first-order disagreement-awareness.[1]

Proposition 8.12. *Sarah's second-order disagreement-awareness with John (eq. (8.7)) will be resolved if Sarah decides to be ignorant about $p : \eta_3$ (eq. (8.10)). Proof in Appendix A.5.*

[1] If Sarah resolves the first-order disagreement, then she has to admit that $\theta_4 \leq_k \theta_1$ and $\theta_4 \neq \theta_1$, that is, she implicitly has to admit that she knowingly kept John unaware of her true beliefs.

Decision:	*Resolves:*
$d2r(s, p : \eta_3, B_s)$	Sarah's 1-order disagreement-awareness
$d2r(s, p : \eta_3, B_sB_jB_s)$	Sarah's 2-order disagreement-awareness
$d2r(s, p : \eta_2, B_sB_j)$	Sarah's 1-order disagreement-awareness
$d2r(s, p : \eta_4, B_sB_j)$	Sarah's 2-order disagreement-awareness

Table 8.2: Overview of the decisions to retract beliefs that resolve Sarah's awareness of the disagreement.

3. From Sarah's perspective, John resolves Sarah's first-order disagreement-awareness if he decides to be ignorant about a certain $p : \eta_2$ (eq. (8.5)), and thus would come to believe at most $p : (\theta_1 \otimes_k \theta_3)$.

Proposition 8.13. *Sarah's first-order disagreement-awareness with John (eq. (8.6)) will be resolved if John decides to be ignorant about $p : \eta_2$ (eq. (8.5)). Proof in Appendix A.5.*

4. From Sarah's perspective, John resolves Sarah's second-order disagreement-awareness if he decides to be ignorant about propositions such that according to Sarah, John comes to believe at most $p : (\theta_3 \otimes_k \theta_4)$. Sarah believes that John believes that she believes $p{:}\theta_4$, i.e. $p{:}\theta_4 \in max_k(B_sB_jB_s)$. Thus, John would resolve the disagreement from his perspective, as seen from Sarah's perspective, if he becomes to believe at most $p : (\theta_3 \otimes_k \theta_4)$. Thus, if John comes to believe at most $p : (\theta_3 \otimes_k \theta_4)$, Sarah's second-order disagreement-awareness is resolved.

   To come to believe at most $p : (\theta_3 \otimes_k \theta_4)$, John must decide to be ignorant about a certain $p : \eta_4$. Truth value η_4 should have less information than θ_3, i.e. $\eta_4 \leq_k \theta_3$, and η_4 should not have less information than $(\theta_3 \otimes_k \theta_4)$, i.e. $\eta_4 \not\leq_k (\theta_3 \otimes_k \theta_4)$, and η_4 is the smallest in order $\leq_k$.

$$\eta_4 \;\equiv\; \bigotimes_k \Big(\{\theta' \in \mathcal{B} \mid (\theta' \leq_k \theta_3) \wedge (\theta' \not\leq_k (\theta_3 \otimes_k \theta_4))\} \Big) \tag{8.11}$$

Proposition 8.14. *Sarah's second-order disagreement-awareness with John (eq. (8.7)) will be resolved if John decides to be ignorant about $p : \eta_4$ (eq. (8.11)). Proof in Appendix A.5.*

Example 8.7. *Next, we provide example truth values that Sarah may use to resolve her disagreement-awareness with John. Continued from example 8.5. Suppose: $\mathcal{M}_s$ with B_s, B_sB_j, $B_s\widetilde{B}_j$, $B_sB_jB_s$, $B_sB_j\widetilde{B}_s \in \mathcal{M}_s$ with $p{:}\theta_1 \in max_k(B_s)$, $p{:}\theta_3 \in max_k(B_sB_j)$, $p{:}\theta_4 \in max_k(B_sB_jB_s)$, $p : \tau_2 \in B_s\widetilde{B}_j$ and $p : \tau_3 \in B_sB_j\widetilde{B}_s$.*

1. *Assume: $\theta_1 = {}^{7}/_{10} \times {}^{3}/_{10}$, $\theta_3 = {}^{2}/_{10} \times {}^{6}/_{10}$, $\theta_4 = {}^{5}/_{10} \times {}^{1}/_{10}$, $\tau_2 = {}^{4}/_{10} \times {}^{6}/_{10}$, $\tau_3 = {}^{5}/_{10} \times {}^{4}/_{10}$, i.e. Sarah has a second-order disagreement-awareness with John about the truth value of p, see example 8.5(3).*

2. *We have $\eta_3 = {}^3\!/_{10} \times {}^0\!/_{10}$, e.g. $\mathcal{M}_s' = retr_{ms}(p : \eta_3, B_s)(\mathcal{M}_s)$ with $B_s' \in \mathcal{M}_s'$ with $p : \theta_1' \in max_k(B_s')$ with $\theta_1' = {}^2\!/_{10} \times {}^3\!/_{10}$. Because $\theta_1' \leq_k \theta_3$, the second-order disagreement-awareness is resolved.*

3. *We have $\eta_4 = {}^0\!/_{10} \times {}^2\!/_{10}$, e.g. $\mathcal{M}_s' = retr_{ms}(p : \eta_4, B_s B_j')(\mathcal{M}_s)$ with $B_s B_j' \in \mathcal{M}_s'$ with $p : \theta_3' \in max_k(B_s B_j')$ with $\theta_3' = {}^1\!/_{10} \times {}^2\!/_{10}$. Because $\theta_1' \leq_k \theta_1$, the second-order disagreement-awareness is resolved.*

8.2 Acting to resolve Disagreements

If agents have disagreements, and given that their disagreements are unsatisfactory states to maintain, then agents may take such states of disagreement to motivate actions with the intent to resolve and come to agree on issues. This section addresses four inference rules that may lead to the possible actions that resolve a disagreement-awareness. By adding certain propositions to Sarah's desires, Sarah will come to have an unbalanced mental state structure, and this unbalanced state will motivate her to decide to believe propositions with the intent to balance her mental state structure. In this balanced state, Sarah will have resolved her disagreement-awareness. If the decision rules cannot balance her mental state structure, then Sarah will use dialogue rules with the intent to convey and in return receive information such that the disagreement can be resolved. If Sarah changes her desires, the decision and dialogue games are 'instructed' to resolve the disagreement: the interlocking of all games can be used to resolve disagreements. Assume for the remainder of this chapter that Sarah is aware that John is aware that they disagree about the truth value of formula p, i.e. $\mathcal{M}_s \models$ equation (8.7).

8.2.1 Questions to resolve disagreements

Sarah resolves her disagreement-awareness with John, as described in Section 8.1.5, if she decides to believe $p : \xi_3$ (eq. (8.8)). If Sarah's decision to believe $p : \xi_3$ does not conflict with her desires, as described Section 6.5.3, then she is motivated to desire to believe $p : \xi_3$. We equip Sarah with an inference rule that will make her add the desire to believe $p : \xi_3$, if, first, she is aware of a second-order disagreement-awareness, $\mathcal{M}_s \models$ equation (8.7), and deciding to believe $p : \xi_3$ resolves the disagreement-awareness; second, believing $p : \xi_3$ is coherent with her desires, i.e. $\mathcal{M}_s \models$ *coherent*$(p : \xi_3 \in B_s, \Delta_s)$.

$$\Big(((8.7) \wedge \mathit{coherent}(p : \xi_3 \in B_s, \Delta_s)) \rightarrowtail (p : \xi_3 \in D_s B_s)\Big) \in \mathcal{K}_s \tag{8.12}$$

If Sarah desires to believe $p : \xi_3$, then all her beliefs that will make her explicitly justified to believe $p : \xi_3$ will be part of her indirect desires to believe and her indirect desires to be ignorant. These indirect desires will make her justified, if she is allowed to do so, to decide to believe $p : \xi_3$. If Sarah lacks the grounds to decide to believe

$p:\xi_3$, i.e. her decision rules cannot balance her mental state structure, then she is motivated to communicate with other agents with the intent to acquire information such that she will become explicitly justified to decide to believe $p:\xi_3$ afterwards. Sarah will ask other agents with the communicative act *qba*s and *qbr*s whether she may decide to add or retract beliefs that will make her justified to decide to believe $p:\xi_3$.

Sarah also resolves her disagreement-awareness with John if she decides to be ignorant about $p:\eta_3$ (eq. (8.10)), and thus decides to believe at most $p:(\theta_1 \otimes_k \theta_3)$. If Sarah's decision to be ignorant about $p:\eta_3$ does not conflict with her desires, then she is motivated to desire to be ignorant about $p:\eta_3$. We equip Sarah with an inference rule that will make her adopt the desire to be ignorant about $p:\eta_3$, if, first, she is aware of a second-order disagreement-awareness, i.e. $\mathcal{M}_s \models$ equation (8.7), and deciding to be ignorant about $p:\eta_3$ resolves the disagreement-awareness; second, being ignorant about $p:\eta_3$ is coherent with her desires, i.e. $\mathcal{M}_s \models \mathit{coherent}(p:\eta_3 \notin B_s, \Delta_s)$.

$$\Big(((8.7) \wedge \mathit{coherent}(p:\eta_3 \notin B_s, \Delta_s)) \rightarrowtail (p:\eta_3 \in D_s\widetilde{B_s})\Big) \in \mathcal{K}_s \tag{8.13}$$

8.2.2 Requests to resolve disagreements

As seen from Sarah's perspective, John resolves their disagreement-awareness if he decides to believe $p:\xi_4$ (eq. (8.9)). If John's decision to believe $p:\xi_4$ is coherent with Sarah's desires about John's beliefs, i.e. $\mathcal{M}_s \models \mathit{coherent}(p:\xi_4 \in B_sB_j, \Delta_s)$, then Sarah is motivated to desire that John believes $p:\xi_4$. We equip Sarah with an inference rule that will make her adopt the desire that John believes $p:\xi_4$.

$$\Big(((8.7) \wedge \mathit{coherent}(p:\xi_4 \in B_sB_j, \Delta_s)) \rightarrowtail (p:\xi_4 \in D_sB_j)\Big) \in \mathcal{K}_s \tag{8.14}$$

As seen from Sarah's perspective, John resolves their disagreement-awareness if he decides to be ignorant about $p:\eta_4$ (eq. (8.11)). If John's decision to be ignorant about $p:\eta_4$ is coherent with Sarah's desires about John's ignorance, i.e. $\mathcal{M}_s \models \mathit{coherent}(p:\eta_4 \in B_s\widetilde{B}_j, \Delta_s)$, then Sarah is motivated to desire that John is ignorant about $p:\eta_4$. We equip Sarah with the inference rule that will make her adopt the desire that John is ignorant about $p:\eta_4$.

$$\Big(((8.7) \wedge \mathit{coherent}(p:\eta_4 \in B_s\widetilde{B}_j, \Delta_s)) \rightarrowtail (p:\eta_4 \in D_s\widetilde{B}_j)\Big) \in \mathcal{K}_s \tag{8.15}$$

Example 8.8. *Next, we provide a situation in which Sarah resolves her disagreement-awareness with John. Continued from example 8.6. Suppose:* $\mathcal{M}_s$ *with* $B_s, B_sB_j, B_s\widetilde{B}_j, B_sB_jB_s, B_sB_j\widetilde{B}_s \in \mathcal{M}_s$ *with* $p:\theta_1 \in \mathit{max}_k(B_s)$, $p:\theta_3 \in \mathit{max}_k(B_sB_j)$, $p:\theta_4 \in \mathit{max}_k(B_sB_jB_s)$, $p:\tau_2 \in B_s\widetilde{B}_j$ *and* $p:\tau_3 \in B_sB_j\widetilde{B}_s$.

1. *Assume:* $\theta_1 = {}^7\!/_{10} \times {}^3\!/_{10}$, $\theta_3 = {}^2\!/_{10} \times {}^6\!/_{10}$, $\theta_4 = {}^5\!/_{10} \times {}^1\!/_{10}$, $\tau_2 = {}^4\!/_{10} \times {}^6\!/_{10}$, $\tau_3 = {}^5\!/_{10} \times {}^4\!/_{10}$, *i.e. Sarah has a second-order disagreement-awareness with John about the truth value of p, see example 8.5(3). Thus,* $\xi_3 = {}^0\!/_{10} \times {}^6\!/_{10}$, *see example 8.6(2).*

2. *With the inference rule from equation (8.12), Sarah will adopt the desire to believe* $p : \xi_3$, *i.e.* $p : {}^{0}\!/_{10} \times {}^{6}\!/_{10} \in D_s B_s$.

3. *Assume that Sarah knows the following inference rule.*

$$\big((q : {}^{7}\!/_{10} \times {}^{0}\!/_{10} \in B_s) \rightarrowtail (p : {}^{7}\!/_{10} \times {}^{0}\!/_{10} \in B_s)\big) \in \mathcal{K}_s \tag{8.16}$$

   From $p : {}^{0}\!/_{10} \times {}^{6}\!/_{10} \in D_s B_s$ *and the inference rule from equation (8.16) follows* $q : {}^{7}\!/_{10} \times {}^{0}\!/_{10} \in ind(D_s B_s)$. *If Sarah decides to add* $q : {}^{7}\!/_{10} \times {}^{0}\!/_{10}$ *to her belief state, then, if she closes her belief state under decision-making, due to inference rule from equation (8.16), she will balance her mental state structure.*

4. *Assume that Sarah believes,* $q : {}^{10}\!/_{10} \times {}^{0}\!/_{10}$. *If her mental state structure is closed under decision-making, due to inference rule from equation (8.16), she will add* $p : {}^{7}\!/_{10} \times {}^{0}\!/_{10}$ *to her belief state. If she adds this, then due to rule R3 she will believe* $p : {}^{7}\!/_{10} \times {}^{7}\!/_{10}$, *and agree with John: she has resolved her disagreement.*

8.3 To Agree to Disagree

We now describe the situation in which agents may become aware that their disagreements are irresolvable. An agent may regard her disagreement with another agent irresolvable if she has exhausted all possible actions that could have resolved the disagreement. Stated differently, if all possible communicative acts that, according to the agent, can resolve the disagreement have been tried and have been answered, yet the disagreement remains, then the disagreement is irresolvable.

8.3.1 Irresolvable disagreements

In the previous section, we described the decision rules that allow Sarah to change her desires such that she will become motivated to communicate with the intent to resolve her disagreement. Next, we describe the properties of Sarah's mental state structure in which the disagreement with John has not been resolved.

1. If Sarah decides to believe at least $p : \xi_3$ (eq. (8.8)), then her disagreement-awareness would be resolved. However, her communicative acts that could make her explicitly justified to decide to believe $p : \xi_3$ have failed, i.e. $\mathcal{M}_s$ enjoys the property *finished*$(s, p : \xi_3 \notin B_s)$.

2. From Sarah's perspective, if John decides to believe at least $p : \xi_4$ (eq. (8.9)), then Sarah's disagreement-awareness would be resolved. However, Sarah's communicative acts that could make John explicitly justified to decide to adopt $p : \xi_4$, as seen from her perspective, have failed, i.e. $\mathcal{M}_s$ enjoys the property *finished*$(s, p : \xi_4 \in B_s B_j)$.

3. If Sarah decides to be ignorant about $p : \eta_3$ (eq. (8.10)), then her disagreement-awareness would be resolved. However, her communicative acts that could make her explicitly justified to decide to retract $p : \eta_3$ have failed, i.e. $\mathcal{M}_s$ enjoys the property $\mathit{finished}(s, p : \eta_3 \notin B_s)$.

4. From Sarah's perspective, if John decides to be ignorant about $p : \eta_4$ (eq. (8.11)), then Sarah's disagreement-awareness would be resolved. However, Sarah's communicative acts that could make John explicitly justified to decide to retract $p : \eta_4$, as seen from her perspective, have failed, i.e. $\mathcal{M}_s$ enjoys the property $\mathit{finished}(s, p : \eta_4 \in B_s\widetilde{B_j})$.

Sarah has run out of communicative acts if these four properties hold in her mental state structure, i.e. $\mathcal{M}_s$ enjoys the property as expressed by equation (8.17).

$$\begin{aligned} &\mathit{finished}(s, p : \xi_3 \in B_s) \wedge \mathit{finished}(s, p : \xi_4 \in B_s\underset{\sim}{B_j}) \wedge \\ &\mathit{finished}(s, p : \eta_3 \notin B_s) \wedge \mathit{finished}(s, p : \eta_4 \in B_s\widetilde{B_j}) \end{aligned} \tag{8.17}$$

We assume that the property as expressed by equation (8.17) can be tested on the agent's mental state structure. The agent does not need to have an epistemic attitude, such as beliefs, about whether the property is true or false. However, we assume that the property can be used in the antecedent to inference rules to express that the agent has run out of communicative acts to reach a certain mental state structure.

8.3.2 Believing irresolvable disagreements

If Sarah disagrees with John, and she has tried everything to resolve their disagreement, then, as seen from her perspective, she is justified to decide to believe that their disagreement is irresolvable. A special purpose formula will be used to express that Sarah has, from her perspective, an irresolvable disagreement with John about the truth value of formula p. Formula $\mathsf{irdis}(s, p : \theta_4, j, p : \theta_3)$ states, as expressed in equation (8.7), that Sarah disagrees with John about the truth value of p, i.e. among other properties that $p : \theta_4 \in B_sB_jB_s$ and $p : \theta_3 \in B_sB_j$. For $s, j \in \mathscr{A}$ with $s \neq j$, and $p : \theta_4, p : \theta_3 \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$, then $\mathsf{irdis}(s, p : \theta_4, j, p : \theta_3) \in \mathcal{F}$.[2]

We assume that Sarah knows the following two inference rules. The antecedent for Sarah to decide to believe that $\mathsf{irdis}(s, p : \theta_4, j, p : \theta_3)$ is true, i.e. that their disagreement is irresolvable, is that, from Sarah's perspective, she has a disagreement, (eq. (8.7)), and the disagreement is irresolvable (eq. (8.17)).

$$\Big(((8.7) \wedge (8.17)) \rightarrowtail (\mathsf{irdis}(s, p : \theta_4, j, p : \theta_3) : \mathsf{t} \in B_s)\Big) \in \mathcal{K}_s \tag{8.18}$$

[2] Assume that irresolvable disagreements cannot be made about an irresolvable disagreement, i.e. nesting the irdis formula within an irdis is not possible. We will not discuss whether Sarah can have an irresolvable disagreement with John whether they have an irresolvable disagreement on the truth value of p, i.e. $\mathsf{irdis}(s, \mathsf{irdis}(s, p : \theta_4, j, p : \theta_3) : \mathsf{t}, j, \mathsf{irdis}(s, p : \theta_4, j, p : \theta_3) : \mathsf{f}) \notin \mathcal{F}$.

If Sarah ceases to regard the disagreement as irresolvable, or ceases to think that she has a disagreement, then she will become ignorant about the disagreement.

$$\Big((\neg(8.7) \wedge (8.17)) \rightarrowtail (\mathsf{irdis}(s, p:\theta_4, j, p:\theta_3):\mathsf{t} \notin B_s)\Big) \in \mathcal{K}_s \tag{8.19}$$

8.3.3 Communicating irresolvable disagreements

If Sarah believes she has an irresolvable disagreement with John, then she desires that John believes that she has this irresolvable disagreement. If John were to believe that Sarah has an irresolvable disagreement with him, he could help to resolve it. If Sarah knows the following inference rule, then she will act to make John believe that she has an irresolvable disagreement with him. The antecedent for desiring that John believes that $\mathsf{irdis}(s, p{:}\theta_4, j, p{:}\theta_3)$ is true, is that Sarah believes she has this irresolvable disagreement.

$$\Big((\mathsf{irdis}(s, p:\theta_4, j, p:\theta_3):\mathsf{t} \in B_s) \rightarrowtail (\mathsf{irdis}(s, p:\theta_4, j, p:\theta_3):\mathsf{t} \in D_s B_j)\Big) \in \mathcal{K}_s \tag{8.20}$$

If Sarah ceases to believe that $\mathsf{irdis}(s, p:\theta_4, j, p:\theta_3)$ is true, then she does not desire that John believes this.

$$\Big((\mathsf{irdis}(s, p:\theta_4, j, p:\theta_3):\mathsf{t} \in \widetilde{B_s}) \rightarrowtail (\mathsf{irdis}(s, p:\theta_4, j, p:\theta_3):\mathsf{t} \notin D_s B_j)\Big) \in \mathcal{K}_s \tag{8.21}$$

The dialogue game of requesting belief additions (Section 7.4) will do the rest and will make John believe that Sarah has an irresolvable disagreement: Sarah will request John to believe $\mathsf{irdis}(s, p:\theta_4, j, p:\theta_3):\mathsf{t}$, and John will respond either with a negative or an affirmative answer.

8.3.4 Irresolvable disagreement-awareness

Next, we define inference rules that allow Sarah to decide to believe certain consequences from her belief that John believes that he has an irresolvable disagreement with her.

Because Sarah is assumed to have a disagreement, as given in equation (8.7), Sarah's mental state structure has, among other properties, $p:\theta_3 \in max_k(B_s B_j)$ and $p:\theta_4 \in max_k(B_s B_j B_s)$. In addition to Sarah's disagreement, assume that John believes he has the following irresolvable disagreement with Sarah: $\mathsf{irdis}(j, p{:}\theta_5, s, p{:}\theta_6){:}\mathsf{t} \in B_j$. John's mental state structure has, among others properties, $p:\theta_5 \in max_k(B_j B_s)$ and $p:\theta_6 \in max_k(B_j B_s B_j)$. For convenience, take $\kappa \equiv \mathsf{irdis}(j, p:\theta_5, s, p:\theta_6)$. Assume that John requested Sarah to believe $\kappa:\mathsf{t}$, i.e. $rba(j, s, \kappa:\mathsf{t})$ (cf. Section 7.4). Sarah thus believes that John has an irresolvable disagreement with her about p.

$$\mathcal{M}_s \models (\kappa:\mathsf{t} \in B_s B_j),\ (p:\theta_3 \in max_k(B_s B_j)),\ (p:\theta_4 \in max_k(B_s B_j B_s)) \tag{8.22}$$

In this mental state structure, Sarah may decide to believe properties about John's mental state structure.

1. The situation $\theta_6 \not\leq_k \theta_3$ occurs when, according to Sarah, she is incorrectly informed about John's beliefs. Sarah believed that John believed $p\!:\!\theta_3$; however, from Sarah's belief about John's irresolvable disagreement $\kappa : \mathsf{t} \in B_sB_j$, she can conclude that John believes that she allegedly believes that he believes $p : \theta_6$, i.e. $p : \theta_6 \in max_k(B_jB_sB_j)$. If $\theta_6 \not\leq_k \theta_3$, then Sarah may add $p : \theta_6$ to her beliefs about John's beliefs.

$$\Big(((\kappa : \mathsf{t} \in B_sB_j) \wedge (p : \theta_3 \in max_k(B_sB_j))) \rightarrowtail (p : \theta_6 \in B_sB_j)\Big) \in \mathcal{K}_s \tag{8.23}$$

2. The situation $\theta_4 \not\leq_k \theta_5$ occurs when, according to Sarah, John is incorrectly informed about her beliefs. From Sarah's belief about John's irresolvable disagreement $\mathsf{irdis}(j, p : \theta_5, s, p : \theta_6) : \mathsf{t} \in B_sB_j$, she can conclude that John believes that Sarah believes $p\!:\!\theta_5$, i.e. $p\!:\!\theta_5 \in max_k(B_jB_s)$. However, Sarah herself believes that John believes that she believes $p : \theta_4$. If $\theta_4 \not\leq_k \theta_5$, then Sarah may add her belief about John's belief about hers, i.e. she decides to add $p : \theta_5$ to $B_sB_jB_s$.

$$\Big(((\kappa : \mathsf{t} \in B_sB_j) \wedge (p : \theta_4 \in max_k(B_sB_jB_s))) \rightarrowtail (p : \theta_5 \in B_sB_jB_s)\Big) \in \mathcal{K}_s \tag{8.24}$$

   From the dialogue game to inform about changed beliefs from Section 7.6, the communicative act *is* will allow Sarah to inform John that he is incorrectly informed about her belief. Such communication may resolve, as a last resort, the disagreement on the truth value of p.

3. Sarah may decide to believe that John has an irresolvable disagreement with her about the truth value of p if she believes she has an irresolvable disagreement with him about p. Sarah believes she has an irresolvable disagreement with John, i.e. $\kappa_1 \equiv \mathsf{irdis}(s, p : \theta_4, j, p : \theta_3)$ and $\kappa_1 : \mathsf{t} \in B_s$, and if she believes that John believes he has an irresolvable disagreement with her about the same formula, i.e. $\kappa_2 \equiv \mathsf{irdis}(j, p : \theta_6, s, p : \theta_5)$ and $\kappa_2 : \mathsf{t} \in B_sB_j$, and, according to Sarah, John is not incorrectly informed about her beliefs, i.e. $\theta_6 \leq_k \theta_3$ and $\theta_4 \leq_k \theta_5$, then she may believe that John has an irresolvable disagreement with her.

$$\Big(((\kappa_1 : \mathsf{t} \in B_s) \wedge (\kappa_2 : \mathsf{t} \in B_sB_j)) \rightarrowtail (\kappa_2 : \mathsf{t} \in B_s)\Big) \in \mathcal{K}_s \tag{8.25}$$

8.3.5 Agreeing or disagreeing to disagree

From Sarah's belief that John has an irresolvable disagreement with her, the previous section described three inference rules that allow her to change her beliefs about John's beliefs, such that she may resolve the disagreement, or she decides to believe that he indeed has an irresolvable disagreement with her. Next, we will allow Sarah to confront John with what she believes about his irresolvable disagreement.

The inference rules that have been described in Section 8.2 changed the agents' desires such that the dialogue games might possibly resolve the disagreement. If Sarah's mental state structure is closed under decision-making, and thus closed

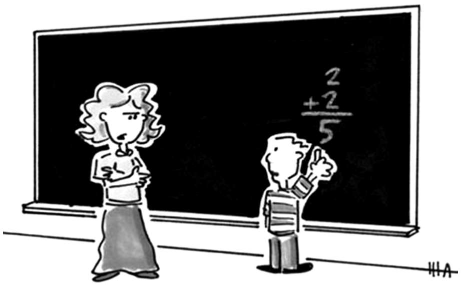

Figure 8.7: Sarah rejects John's request that she is to believe that he has an irresolvable disagreement with her.

under the decision rules and inference rules from the previous section (Section 8.3.4), then Sarah will be allowed to communicate (cf. the deliberation cycle, Section 5.3.4). Two options exist:

1. If Sarah does not believe that John has an irresolvable disagreement with her, i.e. inference rule from equation (8.25) has not been used to change her beliefs, then Sarah denies John's request to believe that he has an irresolvable disagreement. See also figure 8.7. Observe that in such a situation, Sarah has changed her beliefs about John, and is allowed to communicate an inform statement. This inform statement will update Sarah's and John's beliefs, and either resolve the disagreement, or John will be allowed to request that Sarah believes a new and updated irresolvable disagreement.

2. If Sarah believes that John has an irresolvable disagreement with her, then Sarah can grant John's request to believe that he has an irresolvable disagreement with her. Sarah agrees to disagree. After granting John's request, Sarah believes that John believes that she also has an irresolvable disagreement with him, as expressed in the following equation. $\kappa_1 \equiv \textsf{irdis}(s, p : \theta_4, j, p : \theta_3)$ and $\kappa_2 \equiv \textsf{irdis}(j, p : \theta_6, s, p : \theta_5))$ with $\theta_6 \leq_k \theta_3$ and $\theta_4 \leq_k \theta_5$.

$$\mathcal{M}_s \models (\kappa_1 : \mathsf{t} \in B_s),\ (\kappa_2 : \mathsf{t} \in B_s B_j),\ (\kappa_2 : \mathsf{t} \in B_s B_j B_s) \tag{8.26}$$

If Sarah is allowed to respond affirmatively to Sarah's request to believe that she has an irresolvable disagreement with him, Sarah is also allowed to request John to believe that she has an irresolvable disagreement him. If Sarah had first requested to believe that they have an irresolvable disagreement, then Sarah's mental state structure would turn out slightly different.

$$\mathcal{M}_s \models (\kappa_1 : \mathsf{t} \in B_s),\ (\kappa_1 : \mathsf{t} \in B_s B_j),\ (\kappa_1 : \mathsf{t} \in B_s B_j B_s) \tag{8.27}$$

Example 8.9. *Next we provide a situation in which John becomes aware that Sarah has agreed to disagree with him about the truth value of p. Continued from example 8.5. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j, B_sB_jB_s \in \mathcal{M}_s$ with $p:\theta_1 \in max_k(B_s)$, $p:\theta_3 \in max_k(B_sB_j)$ and $p:\theta_4 \in max_k(B_sB_jB_s)$. Suppose: $\mathcal{M}_j$ with $B_j, B_jB_s, B_jB_sB_j \in \mathcal{M}_j$ with $p:\theta_2 \in max_k(B_j)$, $p:\theta_5 \in max_k(B_jB_s)$ and $p:\theta_6 \in max_k(B_jB_sB_j)$.*

1. *Assume: $\theta_1 = {}^7/_{10} \times {}^3/_{10}$, $\theta_3 = {}^2/_{10} \times {}^6/_{10}$, $\theta_4 = {}^5/_{10} \times {}^1/_{10}$, see example 8.5(3).*
2. *Assume: $\theta_2 = {}^3/_{10} \times {}^7/_{10}$, $\theta_5 = {}^6/_{10} \times {}^2/_{10}$, $\theta_6 = {}^1/_{10} \times {}^5/_{10}$.*
3. *Assume that John is aware that he disagrees with Sarah as expressed by equation (8.7). Additionally, assume that John has run out of communicative acts to convince Sarah to agree with him, that is, his mental state structure enjoys property (8.17).*
4. *From equation (8.7), property (8.17) and the inference rule from equation (8.18) follows: $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_j$.*
5. *From $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_j$ and closure under the inference rule from equation (8.20) follows $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in D_jB_s$.*
6. *From $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in D_jB_s$, and $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \notin B_jB_s$ and the preconditions to utter a request for a belief addition (eq. (7.39)), John requests Sarah to believe $\mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t}$.*
7. *After communication we have, among other properties, $\mathcal{M}_s \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_sB_j$ and $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_jB_sB_j$.*
8. *We have $\theta_6 \leq_k \theta_3$, i.e. ${}^1/_{10} \times {}^5/_{10} \leq_k {}^2/_{10} \times {}^6/_{10}$, thus, according to Sarah, John is not completely informed about her beliefs, but his beliefs are not incorrect.*
9. *We have $\theta_4 \leq_k \theta_5$, i.e. ${}^5/_{10} \times {}^1/_{10} \leq_k {}^6/_{10} \times {}^2/_{10}$, thus, according to Sarah, John is not completely informed about what she believes about his beliefs, but his beliefs are not incorrect.*
10. *From $\theta_6 \leq_k \theta_3$, $\theta_4 \leq_k \theta_5$, $\mathcal{M}_s \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_sB_j$, and the inference rule from equation (8.25) follows $\mathcal{M}_s \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_s$.*
11. *From $\mathcal{M}_s \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_s$ and $\mathcal{M}_s \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_sD_jB_s$ and the preconditions to grant a request for a belief addition (eq. (7.42)), Sarah grants John's request to believe $\mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t}$.*
12. *After communication we have, among other properties, $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_jB_s$ and $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_jB_sB_j$.*
13. *From $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_j$, $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_jB_s$ and $\mathcal{M}_j \models \mathsf{irdis}(j, p: {}^1/_{10} \times {}^5/_{10}, s, p: {}^6/_{10} \times {}^2/_{10}) : \mathsf{t} \in B_jB_sB_j$ follows that John is aware that Sarah has agreed to disagree with him about the truth value of p.*

8.4 Interpretation of an Agreement to Disagree

Next, we provide an interpretation (Section 8.4.2) of our agents' agreements to disagree in the context of Aumann's proof that such an agreement cannot exist (Section 8.4.2).

8.4.1 The 'no disagreement' theorem

In *Agreeing to Disagree* [Aum76], Aumann provided a formal definition of common knowledge and used it to prove that two agents cannot "agree to disagree". If agents start from a common prior and update, using Bayes' rule, the probability of an event on the basis of private information, then it cannot be common knowledge between them that on agent assigns a certain probability to the event, and another agent assigns a different probability to the event. In other words, if their posteriors[3] of the event are common knowledge, then the posteriors must coincide. The fact that Sarah has a different opinion from John is a piece of information that should induce Sarah and John to revise their opinions. Such a process of revision will continue until consensus is reached. Aumann's surprising result is that even if agents justify their beliefs with private information, common knowledge of their opinions and a common prior probability distribution implies that their opinions cannot be different.

Translated to our agents, Aumann's result states the following. The meaning of speech acts is common prior, which is common knowledge, and the agents updates her beliefs, i.e. her posteriors, based on utterances of speech acts. If Sarah and John start from a common prior, such as their belief that a patient suffers from a certain disease, then after receiving communication from Fred that the patient has a severe fever, Sarah and John update their belief about the patient's disease. If Sarah and John have only used knowledge that they regard common knowledge, then their new beliefs would have to be the same. Agents cannot disagree.

Since agents often hold different beliefs and they are aware they do so, the 'no disagreement' result raises the question of explaining the origin of differences in belief. Aumann's result leaves us with two options: either we admit that common knowledge of agents' beliefs fails, or deny the 'common prior' assumption (see Halpern [Hal98]). First, we address the common prior assumption, then we will turn to the idea of incorrect common knowledge.

The common prior assumption, also known as the 'Harsanyi Doctrine' [Har68], says that differences in beliefs among agents can be completely explained by differences in information. Essentially, the idea is that agents start out with identical prior beliefs (the common prior) and then condition their beliefs on the information that they later receive. If their new beliefs differ, it must be because they have received different information. Effectively, this states that if agents all have the same knowledge, then they ought to have the same subjective probability assignments. Translated to our agents, if agents start out with identical beliefs and knowledge, that is, with

[3] The posterior probability, also known as the posterior, is a revised probability obtained by updating a prior probability after receiving new information, cf. Grinstead [GS97, p. 155].

identical cognitive states (the common prior), and then update their cognitive state after receiving a speech act, because their cognitive states are equal, they interpret the act equally, and thus their cognitive states remain equal. If their cognitive states do differ, it must be because they received different speech acts.

We will not deny the common prior assumption. If we do not deny the common prior assumption, but still wish to explain the differences of agents' beliefs, we have to question whether agents' common knowledge can be incorrect. Next, we explore this idea and conclude what this means for our agents' agreement to disagree.

8.4.2 Alleged common knowledge

As said in Section 2.3, facts in reality do not determine the truth of propositions for our agents. It are the criteria that hold in the agents' cognitive states that determine the truth of propositions. These criteria have been (tacitly) agreed upon by the agent's community to determine when an agent may regard propositions to constitute a truth. Furthermore, the community has agreed on the criteria that have to hold in an agent's cognitive state to express when she may regard herself entitled to predicate *to believe* propositions.

An agent's view of the meaning of speech acts, as well as of propositions and decisions, is that she regards herself entitled to predicate *to know* that her use of the speech acts is common knowledge. That is, the agent knows a criterion rule that provides her with a criterion to use some speech act, and, according to the agent, all agents in her community know the criterion to use the speech act. Notwithstanding an agent's justification to know the meaning of some speech act, because she grounds this knowledge on her cognitive state, other agents may still use the speech act differently. If the truth-conditions of a proposition had been independent of the agent's cognitive state, such as proposed by Correspondence theory, then an agent could not be justified to know a truth while another agent is also justified to know a contradictory truth. A similar argument holds for speech acts. If the criteria for correct usage had been independent of the agent's cognitive state, then an agent could not attest that a proposition is a truth while another agent is correct to attest that the proposition is not a truth. If a fact in reality provides the truth-condition of a proposition held by two agents, the two agents cannot have a different account of the truth. If the truth-conditions of a proposition or a speech act are properties of an agent's cognitive state, then, because cognitive states are not shared, agents can predicate to believe propositions that contradict, while nevertheless they can be justified to believe them.

Thus, with a Wittgensteinian account of meaning, agents can be justified to know that the use of a speech act is common knowledge, while in an absolute sense, as defined for example by Aumann, the agents do not have common knowledge. If agents are perfectly justified to know that the use of a speech act is common knowledge, then this common knowledge can be the source of the differences in belief that agents often seem to have. Aumann used his definition of common knowledge to prove the celebrated result that says that agents cannot "agree to disagree" about their beliefs. We provided in intuitive analysis that agents that use Wittgenstein's use account of meaning, can agree to disagree about their beliefs.

Speech act theory, as described in Section 7.1.1, explains why a speaker can communicate information with a speech act. The receiver of the speech act is aware in which situations the speaker may utter the act. Additionally, the receiver assumes the speaker intends to be communicate, and thus searches for an interpretation of the speech acts in which the speaker *does* communicate. Under this assumption and with this awareness, the receiver can retrieve the speaker's intentions, and based on these intentions can derive information. However, under a Wittgensteinian interpretation of meaning, a speaker and receiver need not have a shared use of the speech act. Speaker Sarah utters a speech act with the intention to convey certain information in accordance with what *she* regards as the shared use of the act in an absolute sense. However, receiver John may interpret this act in accordance to what *he* regards as the shared use of the act, and if his usage of the act is not equal to Sarah's, he may 'retrieve' different information. If both agents do not become aware that the information that Sarah intended John to retrieve is not equal to the information that John has retrieved, they may continue to be justified to believe that their communication has been effective. From our perspective, their communication has not been effective, but if this infectiveness remains unobserved by the agents, then, from their perspective, the difference between the information that was intended to be communicated and that has been retrieved, may just as well not exist. Our agents need not doubt the meaning of propositions, decisions, and speech acts, as long as they do not observe differences in the way they use them.

An agreement to disagree between two agents is an indication that the two agents have a method to justify beliefs that is not shared. Common knowledge is used in the following three methods.

- First, an agent regards herself entitled to know that the preconditions of speech acts are shared, and that this use is common knowledge. It is because of the common knowledge that the agent can communicate information with speech acts. This common knowledge may not be common after all.

- Second, an agent regards herself entitled to know that the preconditions of decisions are shared, and that this shared use is common knowledge. Because of the shared use, agents can draw the same conclusions, and can agree about beliefs.

- Third, the agent's epistemology does not describe shared use after all. That is to say, the use of the contents of the epistemology is not common knowledge in the agent's community.

In short, if an agent has agreed to disagree, then she can doubt the meaning of speech acts, decisions, or propositions respectively.

Another source for the disagreement between our agents is that they can have inference rules in their knowledge base that allow them to have irresolvable disagreements. If our agents were allowed and capable to change and communicate their knowledge, then the agreement to disagree could be extended to encompass the situation that agents also failed to change their knowledge. Such an extended definition of an agreement to disagree, together with Aumann's no disagreement theorem, can be taken to mean that our agents failed to have a piece of common knowledge that they are justified to regard as common knowledge.

8.5 Concluding Remarks

In this chapter, we provided inference rules that together with the decision rules from Chapter 6 and the dialogue rules from Chapter 7 enable our agents to exchange information, and in the worst-case scenario, to agree to disagree. First, we defined four inference rules that allow Sarah to communicate propositions that, from her perspective on the disagreement, may resolve the disagreement (Section 8.2). We then provided two rules that describe when Sarah is explicitly justified to believe that she has an irresolvable disagreement with John, and when she is explicitly justified to retract that she has such a disagreement (Section 8.3.2). Because Sarah's beliefs about an irresolvable disagreement are based on her mental state structure only, John need not regard their disagreements as irresolvable. We have defined two inference rules that derive new desires such that the dialogue games from Chapter 7 will communicate Sarah's belief about her disagreement with John to him (Section 8.3.3). Based on Sarah's belief that John believes that from his perspective they have an irresolvable disagreement, Sarah can update her beliefs about John's beliefs; what Sarah can learn from the communicated irresolvable disagreements is described in two inference rules. Additionally, as defined in Section 8.3.4, if Sarah also has an irresolvable disagreement with John, then she is justified to decide to believe that John has an irresolvable disagreement with her. The dialogue games from Chapter 7 will make Sarah communicate that she agrees with John's request to believe they have an irresolvable disagreement, or communicate that she disagrees. That is to say, Sarah can disagree to disagree, or agree to disagree.

If Sarah and John agree to disagree about the truth value of a proposition, then they cannot resolve their disagreement themselves. We observed that either:

1. The agents know inference rules that justify them to believe propositions over which they disagree; or

2. The agents regard the use of certain speech acts, decisions, or the meaning of epistemic propositions common knowledge, while in fact this use is not shared between Sarah and John. That is to say, that what they *regarded* as common knowledge, is in fact not common knowledge.

An agreement to disagree is an indication that the shared use of a previously used speech act or decision is not shared after all. Instead of providing the criteria when certain information is common knowledge, with an agreement to disagree, we can provide a criterion when certain information is not common knowledge. In the absence of a method to obtain access to the criteria that another agent use when making a statement, we can never define common knowledge in the absolute sense as Aumann did in his definition of common knowledge. A history of use in which speech acts and decisions have not led to an agreement do disagree can be taken as part of the criterion that speech acts and decisions are common knowledge. That is to say, the absence of an agreement to disagree is part of the criterion for common knowledge.

An agreement to disagree has not only a theoretical use. As said, the company Emotional Brain is developing a software program that supports physicians to diagnose patients with psychopathological syndromes. The software is implemented by means of agents who will report their assessments, as well as their inability to diagnose to the physician. The agents will report their agreements to disagree to their responsible human experts with the question to change their knowledge such that disagreements become resolvable.[4] If agents consult their responsible experts for every disagreement they encounter, then, possibly, the experts are consulted more often than necessary. In general, the physician or expert should only be consulted if agents cannot find information to conclude on their assessments or cannot resolve disagreements themselves. Because our agents have operational definitions when their disagreements are irresolvable, they can agree to disagree. Because time of experts is expensive, agents should only consult them as a last resort. Experts may be consulted if agents agree to disagree. We could argue that it is computationally expensive for agents to reach agreements to disagree, we assume, however, that these computational costs do not outweigh the costs of an expert's time.

[4] How agents report their assessments, their lack of assessments, or their agreements to disagree to physicians and experts, are implementation specific issues, and are beyond the scope of this thesis.

Chapter 9

Conclusions and Future Research

A modus vivendi is a manner of living; a way of life, or a temporary agreement between contending parties pending a final settlement.[1]

Our objective was to design a multiagent system in which agents each can represent a medical expert or a domain of relevant knowledge. Additionally, we set to provide our agents with decision and communication games that will make them act in accordance with the conventions of the experts that the agents represent. We will conclude that the beliefs that our agents may come to reach are justifiable in the expert's domain. Additionally, we present two future extensions to our architecture. One extension that may be interesting for the company Emotional Brain is that agents may question other agents about their grounds to believe propositions. In another extension, the agents can communicate about the meaning of propositions, decisions and possibly even speech acts.

In Section 9.1, we conclude that because our agents act in accordance with the relevant conventions, they can represent an expert's opinion or a domain of knowledge. In Section 9.2, we describe a possible dialogue game that provides the meaning of communicative acts that would allow agents to enquire into the grounds that agents have to justify their beliefs. In Section 9.3, we describe a dialogue game that provides the meaning of communicative acts that would allow agents to communicate about the meaning of propositions.

[1] Dictionary definition of modus vivendi on Answers.com. The American Heritage© Dictionary of the English Language, Fourth Edition Copyright © 2004 by Houghton Mifflin Company. Published by Houghton Mifflin Company.

9.1 Conclusions

9.1.1 Our design of a multiagent system

We have designed a multiagent system in which agents have a set of dialogue rules that allow them to communicate. Agents communicate with the aim of changing their own and other agents' mental state structures such that they become justified to make decisions. Additionally, agents have a set of decision rules that allow them to change their mental state structures with the aim of balancing their unbalanced mental state structures. That is to say, our agents apply dialogue and decision rules, as described by the deliberation cycle, to become explicitly justified to believe propositions they desire to believe, or to become explicitly justified to be ignorant about propositions they desire to be ignorant about.

An agent's knowledge base can be used to customise the agent such that she will represent one domain of interest or one medical expert. By providing an agent with specific desires, the multiagent system is 'instructed' to derive specific beliefs from the knowledge that is distributed over the agents' knowledge bases. In this design, the agents' dialogue and decision games cannot be changed easily without the need to reengineer other dialogue and decision games. However, both games can be programmed by providing the knowledge bases of agents with generic inference rules that allow the agent to make generic decisions. As described in Chapter 8, by providing our agents with a small set of inference rules, they are programmed to agree to disagree. It is a topic for future research to investigate whether sets of inference rules can be defined that allow agents to communicate the grounds of justification (Section 9.2) or to communicate the sets of inference rules that allow them to communicate about the meaning of propositions (Section 9.3).

9.1.2 Justified beliefs only

The software program that is being developed by Emotional Brain should assist physicians in their decision-making processes by proposing decisions and providing accompanying justifications. A physician needs to be aware of the grounds that the program has used to construct and justify the proposed decision. That is, the program will need the agents' justifications for having certain beliefs that contributed to the decision. The physician needs these justifications to become confident that, for example, the decision to prescribe a certain medication is indeed justified. If the physician is to agree with the program's proposed decision, she has to be confident that she can convince other physicians that her decision, which is based on the program's decision, is justified. If the program provides no insight into the reasons why the physician's community would agree on the proposed decision, the physician can only accept the decision at face value. Lacking justifying grounds makes it impossible, by definition, to convince a physician that the proposed decision is justified.

Next, we will conclude that the beliefs that our agents can have are justified in the experts' community. That is to say, if Sarah comes to believe a certain proposition, then she comes to have the belief in accordance with the practices of her expert's community. Because our agents' decision-making and communication behaviour is governed by decision and dialogue rules that (are assumed to) reflect shared use by the medical experts, the effects of decision-making and communication are in accordance with the experts' practices. Furthermore, we assume that if an agent has performed steps that are justifiable separately, then the combined result is also justifiable. Thus, if an agent was justified to update her beliefs after communication, and then was justified to update her beliefs after decision-making, then the composition of both update actions also yields a justified belief state. In our design, the following three situations for establishing new beliefs have been described.

1. Sarah can establish her beliefs based on her personal knowledge that is part of her knowledge base $\mathcal{K}_s$. A medical expert has compiled this knowledge base and fully approves and supports its contents. The decision rule $d2a_1$, that allows Sarah to draw conclusions if antecedents have been met in her mental state structure, reflects shared use in the expert's community. Together with this rule, Sarah's knowledge base provides her with domain specific knowledge that allows her to become explicitly justified to believe, for example, that the patient suffers from a certain disease. If Sarah were to be blamed for incorrectly believing that a patient has a certain disease, she can defend herself by stating that she has acted in accordance with the practices of her expert's community and according to the knowledge of her expert.

2. Sarah can establish beliefs based on the meaning of being in a state of believing a proposition that is described by her epistemology $\mathcal{E}_s$. Medical experts that represent their community have compiled this epistemology, and consequently, we may assume that all experts of the community will support its contents. Similar to the decision rule that allows Sarah to derive epistemic properties with inference rules from her knowledge base, the decision rule $d2a_2$ allows her to derive epistemic properties with inference rules from her epistemology. Because the meaning of believing, as described in the epistemology, is shared in the community, Sarah may become explicitly justified to believe a proposition if its criterion holds in her mental state structure. Additionally, if Sarah is aware that the criterion holds in John's mental state structure, then she may come explicitly justified to believe that John is justified to believe the proposition. If Sarah were to be blamed for incorrectly believing that John has a certain belief, then she can defend herself by stating that she has acted in accordance with the meaning of a decision, and the meaning of believing the proposition, i.e. in accordance with the practices of her expert's community.

3. Sarah can establish beliefs based on her beliefs about other agents' beliefs. Sarah is explicitly justified to have beliefs about other agents' properties of mental states as defined by the perlocutions of speech acts. Because the meaning of

speech acts is shared, the perlocutions of speech acts are also shared. This shared meaning not only allows agents to communicate, but also provides them with explicit justification to believe properties about other agents' mental states in accordance with the experts' conventions. Thus, agents may conform to other agents' beliefs based on their beliefs about other agents' beliefs. The decision processes in which an agent conforms to another agent's beliefs are constrained to conform to those beliefs that the agent desires to believe. Stated differently, an agent's desires guide her decision processes, see the definition of the decision $d2a_4$. If Sarah were to be blamed for incorrectly believing that a patient has a certain disease, then she can defend herself by stating that she has acted in accordance with her decision rules. Additionally, she can state that her beliefs about other agents' beliefs are also established in accordance with the shared use of speech acts, i.e. she has acted in accordance with the practices of her expert's community.

Practically, by making the medical experts agree, firstly, that decision and dialogue rules are conform their uses in their community, secondly, that Sarah's epistemology reflects the standards set by her community, and thirdly, that the domain knowledge of Sarah is the responsibility of certain human experts from the community, then the community should agree that the beliefs that Sarah can come to have are justified in the community. If Sarah's behaviour is governed by conventions and is based on justified beliefs, then Sarah and experts could not have acted differently.

9.1.3 Software providing decision support

Sarah's ability to propose justified decisions is independent of her ability to communicate why her proposed decisions are indeed justified. Sarah can test whether her proposed decision is justified by testing whether the separate steps, which make her propose the decision, are justifiable. Sarah will not be able to explain why these steps are justified; she only regards these steps justified because a (human) medical expert regards the steps justifiable.

To convince a physician that the proposed decisions are justified, the program has to provide the physician with information such that the physician will become convinced the decisions are indeed justified. It still has to be investigated how the different steps that contributed to the decision can be communicated to the physician. We assume that if the physician is informed about the grounds that justify the steps, then she will acknowledge that the decision is used conform the criteria set by her community. If all steps are justified, then it is unreasonable not to call the decision justified. Given the grounds that justify the steps to construct the program's proposed decision, the physician can, in principle, come to make the *same* decision. The program will not claim to *make* a decision, but will provide a proposed decision, and will thus be a decision *support* system.

Experts can be perfectly justified to believe that their contribution to Sarah's knowledge base constitutes undisputed knowledge. However, it is very well conceivable that other experts do not share this belief. An expert may believe that a certain piece of information constitutes knowledge, while another expert believes that a contradicting piece of information constitutes knowledge. These two experts may justify the knowledge bases of two agents, and these knowledge bases may provide the agents with justification to have contradictory beliefs. If human experts become aware that they have conflicting views on when they may call something a truth, or for example, when they may administer some drug, they have to resolve their disagreement, and change their knowledge. Consequently, agents can encounter irresolvable disagreements on their beliefs, and if they do, they should inform their responsible experts, because, unlike the experts, the agents are not allowed to change their knowledge that resulted in the disagreement. In this respect, an agreement to disagree between two agents is a temporary agreement pending a final settlement that will be provided by the experts who are responsible for the agents' knowledge. The responsible experts should resolve their disagreement and change the agent's knowledge bases such that the agent's irresolvable disagreements will become resolvable.

What does it mean for Emotional Brain if two agents have agreed to disagree? If John and Fred have agreed to disagree about the truth value of a formula, then, as described in Section 3.1, Sarah could be justified to believe a paraconsistent proposition. Sarah is justified in doing so, because the meaning of believing the proposition, as described by either her decision rules, epistemology or knowledge base, is supplied by the medical expert she represents. An agent who has a paraconsistent belief is in a state that needs investigation by the medical experts who supplied knowledge that led to the belief's justification. Because the agent is justified to have the belief, as described in Section 3.3, as long as the situation is not resolved, the agent should cope with the paraconsistency.

Because agents' beliefs are justified by different games that represent shared agreements between experts, the beliefs will have meaning for the experts. If Sarah's belief were grounded in the application of some set of rules that are not grounded in the practices of the medical experts, a physician may doubt the validity of Sarah's beliefs, and may think that her beliefs do not express authority. Sarah's behaviour is governed exclusively by her decision and dialogue games, her knowledge base and epistemology. Ultimately, this behaviour changes her beliefs and lack of beliefs. Because these beliefs are obtained by behaviour that reflects the shared agreement of medical experts, the beliefs will have a meaning for these medical experts. These beliefs with a meaning for humans are the 'products' of our multiagent system.

9.2 Dialogues about Grounds of Justification

A useful extension to the five dialogue games from Chapter 7 would be the game that allows agents to ask other agents why they believe or are ignorant about propositions. Such communicative acts would provide agents with questions to ask for the grounds that other agents have used to justify their beliefs and lack of beliefs.

As said in Section 7.1.3, persuasion dialogues are dialogues in which a speaker seeks to convince a listener to accept to believe a particular proposition (McBurney and Parsons, [MP01a]; Dignum et al. [DDKV00] and Prakken [Pra00]). Dialogue games in Law have similar objectives, for example, disputes in which participants have conflicting arguments (see Prakken and Sartor [PS98, PS96]). Agent can use their grounds for justification as arguments that may persuade other agents to adopt beliefs. In addition to multiagent systems, in the domain of Artificial Intelligence and Law, computational and logical models of argumentation and of reasoning with conflicting information have been proposed (see Prakken and Vreeswijk [PV02], Bench-capon et al. [JBCP03]). An argumentation system essentially includes a logical language, a definition of the argument concept, a definition of the attack relation between arguments and finally a definition of acceptability (see Vreeswijk [Vre97]). In an argumentation models, a proof of a formula takes the form of a dialogue tree, in which each branch of the tree is a dialogue and the root of the tree is an argument for the formula. Every move in a dialogue or argumentation system consists of an argument.

Next, we sketch two communicative acts that may be classified as belonging to information seeking dialogues.

- A communicative act $why(s, j, \psi \in B_s)$ could be interpreted by receiver John that speaker Sarah desires to believe on what grounds he has justified his belief in ψ. Sarah may utter it with the intention that the act will be responded by John with the justifications that John has for believing ψ.

- A communicative act $ground(s, j, \psi \in B_s, \phi \in B_sB_j)$ could be interpreted by receiver John that speaker Sarah has justified her belief in ψ on her belief that John believes ϕ. Analogously, communicative act $ground(s, j, \psi \in B_s, \phi \in B_s)$ could be interpreted by John that Sarah has justified her belief in ψ on her belief in ϕ.

Another approach that would allow our agents to communicate the grounds of beliefs is by introducing special purpose propositions and inference rules as part of the agent's knowledge base. These special purpose propositions can be derived with inference rules, and can be communicated with the existing dialogue games. This approach of equipping our agents with such inference rules has the advantage over introducing new dialogue games that the agent architecture does not have to be changed. Whether adding dialogue rules or adding inference rules is feasible, is part of future research.

This functionality is imperative for the end users of the system, i.e. the physician who enters patient data, from which the multiagent system will provide the physician with support to diagnose the patient. The physician will take it for granted to have access to explanations why the system proposes, for example, that a patient suffers from a depression. Without this functionality, the software program would not likely gain acceptance by its users, particularly if potential harmful consequences for people are involved.

9.3 Dialogues about Meaning

Another extension to the dialogue game is one in which agents enquire about the meaning of propositions.

As discussed in Section 2.4, our agents are said to grasp the meaning of propositions if they use the propositions according to the criteria that have been agreed upon by their community. In such situations, our agents are said to know that a criterion is part of a criterion rule that defines the correct use of the proposition. For our agents, we have implemented these criterion rules with inference rules, and the antecedent of the inference rule represents the criterion that the agent's community has tacitly agreed upon. Although agents may be justified to know that the use of a proposition is common knowledge in a community, i.e. to know the meaning of a proposition, other agents from that community may still be justified to know a different meaning. As long as an agent does not observe differences in the way she and others use a proposition, she need not doubt her knowledge of the proposition's meaning.

If Sarah has agreed to disagree with John about the truth value of a proposition, then she has a motive to doubt her knowledge of the proposition's meaning. If Sarah doubts whether she knows the meaning of a particular proposition, she may ask John whether she is correct to predicate *to know* that her use of the proposition is shared. She may also ask John whether he is correct to predicate *to know* that his use of the proposition is shared. In such a dialogue, the agents communicate about the content of their epistemology $\mathcal{E}$. For our agents, who represent medical experts, contemplating the meaning of propositions is beyond their jurisdiction. Only experts are entitled to change the agents' knowledge and epistemology.

If agents agree to disagree about a proposition, then they cannot resolve their disagreement with their current dialogue games. The agents agree about their disagreement, and they still desire to resolve their disagreement; it is only because they cannot resolve it that they have made an agreement. As said before, an agreement to disagree is a temporary agreement between contending parties pending a final settlement. Future research would include the dialogue about changing the agent's knowledge that resulted in the disagreement such that the disagreement may become resolvable.

Bibliography

[AA98] Ofer Arieli and Arnon Avron. The value of the four values. *Artificial Intelligence*, 102(1):97–141, 1998.

[Acz96] Amir D. Aczel. *Fermat's last theorem : Unlocking the secret of an ancient mathematical problem.* Four Walls Eight Windows, New York, 1996.

[AMP00] Leila Amgoud, Nicolas Maudet, and Simon Parsons. Modelling Dialogues using argumentation. In E.H. Durfee, editor, *Proceedings of the Fourth International Conference on Multi-Agent Systems (ICMAS 2000)*, pages 31–38, Boston, MA, USA, 2000. IEEE Press.

[And96] Floy E. Andrews. The principle of excluded middle then and now: Aristotle and principia mathematica. Available on `http://www.mun.ca/animus/1996vol1/andrews.htm`, 1996.

[APM00] Leila Amgoud, Simon Parsons, and Nicolas Maudet. Arguments, dialogue, and negotiation. In W. Horn, editor, *Proceedings of the Fourteenth European Conference on Artificial Intelligence (ECAI 2000)*, pages 338–342, Berlin, Germany, 2000. IOS Press.

[Aud98] Robert Audi. *Epistemology: a contemporary introduction to the theory of knowledge*. Routledge, 1998.

[Aum76] Robert Aumann. Agreeing to disagree. *Annals of Statistics*, 4:1236–9, 1976.

[Aus62] John L. Austin. *How to do Things with Words.* Harvard University Press, Cambridge Mass., 1962.

[AvH03] Grigoris Antoniou and Frank van Harmelen. Web Ontology Language: OWL. In S. Staab and R. Studer, editors, *Handbook on Ontologies in Information Systems*. Springer-Verlag, 2003.

[Bar88] Jon Barwise. Three views of common knowledge. In M.Y. Yardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 365–79, San Francisco, 1988. Morgan Kaufman.

[Bar00] Jonathan Baron. *Thinking and Deciding*. Cambridge University Press, Cambridge, UK, third edition, 2000.

[Baz98] V. A. Bazhanov. Toward the Reconstruction of the Early History of Paraconsistent Logic: The Prerequisites of N.A.Vasiliev's Imaginary Logic. *Logique et Analyse*, 161–162–163:17–20, 1998.

[Bel77] Nuel D. Belnap Jr. A useful four-valued logic. In J. Michael Dunn and G. Epstein, editors, *Modern Uses of Multiple-Valued Logic*, pages 8–37, Dordrecht, 1977. D. Reidel.

[Beu01] Robbert-Jan Beun. On the Generation of Coherent Dialogue: A Computational Approach. *Pragmatics & Cognition*, 9(1):37–68, 2001.

[BHS04] Leopoldo Bertossi, A. Hunter, and T. Schaub. Introduction to Inconsistency Tolerance. In *Inconsistency Tolerance*, volume 3300 of *LNCS*, pages 1–14. Springer Verlag, Berlin, Germany, 2004.

[BLHL91] Tim Berners-Lee, James Hendler, and Ora Lasilla. The semantic web: new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, May 1991.

[Bro86] Rodney A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23, March 1986.

[Cla78] Keith L. Clark. Negation as failure. In Gallaire and Minker [GM78], pages 293–322.

[Col00] Marco Colombetti. A commitment-based approach to agent speech acts and conversations. In *Proceedings of the Workshop on Agent Languages and Communication Policies*, pages 21–29, Barcelona, 2000. 4th International Conference on Autonomous Agents.

[CP79] Philip R. Cohen and C. Raymond Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212, 1979.

[Dav86] David Davidson. A coherence theory of truth and knowledge. In Ernest LePore, editor, *Truth and Interpretation, Perspectives on the Philosophy of Donald Davidson*, pages 307–19. Basil Blackwell, Oxford, 1986.

[DBD$^+$06] J. van Diggelen, R.J. Beun, F. Dignum, R.M. van Eijk, and J.-J.Ch. Meyer. ANEMONE: An effective minimal ontology negotiation environment. *Proceedings of the Fifth International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2006.

[DDKV00] Frank Dignum, B. Dunin-Keplicz, and R. Verbrugge. Agent theory for team formation by dialogue. In Cristiano Castelfranchi and Yves Lespérance, editors, *Pre-Proceedings of the Seventh International Workshop on Agent Theories, Architectures and Languages (ATAL 2000)*, pages 141–156, Boston, USA, 2000.

[DDKV01] Frank Dignum, B. Dunin-Keplicz, and R. Verbrugge. Creating collective intentions through dialogue. *Logic Journal of the IGPL*, 9(2):305–319, 2001.

[Den81] Daniel C. Dennett. True believers: The intentional strategy and why it works. In A.F. Heath, editor, *Scientific Explanations*. Oxford University Press, 1981.

[Dig00] Frank Dignum. Agent communication and cooperative information agents. In Klusch and Kerschberg [KK00], pages 191–207.

[dLW97] Mark d'Inverno, Michael Luck, and Michael J. Wooldridge. Cooperation structures. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 600–5, 1997.

[Doy79] Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12:231–272, 1979.

[DP02] B. A. Davey and H. A. Priestley. *Introduction to Lattice and Order*, volume Second Edition. Cambridge University Press, 2002.

[DR76] R. O. Duda and R. Reboh. AI and decision making: the PROSPECTOR experience. *Artificial Intelligence Applications for Business*, pages 111–147, 1976.

[Dum78] Michael Dummett. *Truth and other enigmas*. Duckworth, London, 1978.

[DW98] Keith Derose and Ted A. Warfield. *Skepticism: A Contemporary Reader*. Oxford University Press, New York, 1998.

[Ell03] Sara Ellenbogen. *Wittgenstein's Account of Truth*. SUNY series in philosophy. State University of New York Press, 2003.

[Esh88] L. Eshelman. MOLE: a knowledge-acquisition tool for cover-and-differentiate systems. In Marcus [Mar88], pages 37–80.

[FFMM94] T. Finin, R. Fritzson, D. McKay, and R. McEntire. KQML as an Agent Communication Language. In N. Adam, B. Bhargava, and Y. Yesha, editors, *Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM'94)*, pages 456–463, Gaithersburg, MD, USA, 1994. ACM Press.

[FIP02] FIPA. Communicative act library specification. standard sc00037j. Technical report, Foundation for Intelligent Physical Agents, 2002.

[Fit90] Melvin Fitting. Bilattices in logic programming. In George Epstein, editor, *The Twentieth International Symposium on Multiple-Valued Logic*, pages 238–246. IEEE, 1990.

[Fit91] Melvin Fitting. Bilattices and the semantics of logic programming. *Journal of Logic Programming*, 11:91–116, 1991.

[Gär88] Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. A Bradford book. The MIT Press, 1988.

[Gin88] Matthew L. Ginsberg. Multivalued Logics: A Uniform Approach to Reasoning in Artificial Intelligence. *Computational Intelligence*, 4:265–316, 1988.

[GK96] Barbara J. Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[GM78] Herve Gallaire and Jack Minker, editors. *Logic and Data Bases*. Plenum Press, New York, 1978.

[Gri57] Paul Grice. Meaning. *The Philosophical Review*, 64:377–88, 1957.

[Gri68] Paul Grice. Utterer's meaning, sentence-meaning and word-meaning. *Foundations of Language*, 4:225–42, 1968.

[Gri75] Paul Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, volume 3 of *Syntax and Semantics*, pages 41–58. Academic Press, New York, USA, 1975.

[Gri89] Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1989.

[Gru93a] Thomas R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[Gru93b] Thomas R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43:907–928, 1993.

[GS90] G.B. Guchanan and E.H. Shortliffe. *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Cambridge: The MIT Press, 1990.

[GS97] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*. American Mathematical Society, 1997. 2nd Rev edition.

[Hab84] Jürgen Habermas. *The Theory of Communicative Action: Volume I: Reasons and the Rationalization of Society*. Heinemann, London, 1984. (Published in German 1981).

[Hal98] Joseph Y. Halpern. Characterizing the Common Prior Assumption. In Itzhak Gilboa, editor, *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Seventh Conference (TARK 1998)*, pages 133–146. Morgan Kaufmann, July 22 1998.

[Har68] John C. Harsanyi. Games with incomplete information played by "bayesian" players, parts i, ii, iii. *Management Science*, 14:159–82, 320–34, 486–502, 1968.

[Hin62] Jaako Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, New York, 1962.

[Hin01] Koen Hindriks. *Agent Programming Languages: Programming with Mental Models*. PhD thesis, Universiteit Utrecht, Utrecht, The Netherlands, 2001.

[HMP01] David Hitchcock, Peter McBurney, and Simon Parsons. A framework for deliberation dialogues. In H. V. Hansen, C. W. Tindale, J. A. Blair, and R. H. Johnson, editors, *Proceedings of the Fourth Biennial Conference on the Ontario Society of the Study of Argumentation (OSSA 2001)*, Windsor, Ontario, Canada, 2001.

[Hor98] Paul Horwich. *Meaning*. Clarendon Press, Oxford , UK, 1998.

[HS97a] Michael N. Huhns and Munindar P. Singh. Agents and multiagent systems: Themes, approaches, and challenges. [HS97b], pages 1–23.

[HS97b] Michael N. Huhns and Munindar P. Singh, editors. *Readings in Agents*, San Francisco, CA, 1997. Morgan Kaufmann.

[Hul00] Joris Hulstijn. *Dialogue Models for Inquiry and Transaction*. PhD thesis, Universiteit Twente, Enschede, The Netherlands, 2000.

[Jam95] William James. *Pragmatism*. Dover Publications, New York, 1995. Originally published: London; New York: Longman Green, 1907.

[Jam97] William James. *The Meaning of Truth*. Great Books in Philosophy. Prometheus Books, New York, 1997. Originally published: New York: Longmans, Geen, and Co., 1911.

[JBCP03] H. Hohmann J.M. Bench-Capon, J.B. Freeman and H. Prakken. Computational models, argumentation theories and legal practice. In C. Reed and T.J. Norman, editors, *Argumentation Machines. New Frontiers in Argument and Computation*, Kluwer Argumentation Library, pages 85–120. Kluwer Academic Publishers, Boston/Dordrecht/London, 2003.

[Jen01] Nicholas R. Jennings. An agent-based approach for building complex software systems. *Communications of the ACM*, 44(4):35–41, 2001.

[JSW98] Nicholas R. Jennings, Katia P. Sycara, and Michael J. Wooldridge. A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, 1:7–38, 1998.

[Kah88] G. Kahn. MORE: from observing knowledge engineers to automating knowledge acquisition. In Marcus [Mar88], pages 7–35.

[KK00] Matthias Klusch and Larry Kerschberg, editors. *Cooperative Information Agents IV*, volume 1860 of *LNAI*. Springer, 2000.

[Kle50] Stephen C. Kleene. *Introduction to Methmathematics*. D. Van Nostrand, Princeton, NJ, 1950.

[Kli04] Gyula Klima. The Medieval Problem of Universals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Winter 2004.

[KM91] Hirofumi Katsuno and Alberto Mendelzon. On the Difference Between Updating a Knowledge Base and Revising It. In James F. Allen, Richard Fikes, and Erik Sandewall, editors, *KR'91: Principles of Knowledge Representation and Reasoning*, pages 387–394. Morgan Kaufmann, San Mateo, California, 1991.

[Kri63] Saul Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

[KW03] Bryan Kolb and Ian Q. Wishaw. *Fundamentals of Human Neuropsychology*. W. H. Freeman, New York, fifth edition, 2003.

[Lab01] Yannis Labrou. Standardizing Agent Communication. *LNCS*, 2086:74–98, 2001.

[Lat99] Bruno Latour. *Pandora's Hope: Essays on the Reality of Science Studies*. Harvard University Press, Cambridge, Mass, 1999.

[Les95] Victor R. Lesser. Multiagent systems: An emerging subdiscipline of ai. *ACM Computing Surveys*, 27(3):340–342, January 1995.

[Lew02] David Lewis. *Convention: A Philosophical Study*. Blackwell Publishers, 2002. First published 1969 by Harvard University Press.

[Lip64] Seymour Lipschutz. *Set Theory and Related Topics*. Schaum's Outline Series. McGraw-Hill, 1964.

[LWM03a] Henk-Jan Lebbink, Cilia Witteman, and John-Jules Ch. Meyer. Dialogue games for inconsistent and biased information. In Tom Heskes, Peter Lucas, Louis Vuurpijl, and Wim Wiegerinck, editors, *15th*

Belgium-Netherlands Conference on Artificial Intelligence (BNAIC'03), pages 427–428. K.U. Nijmegen, 23–24 October 2003. Extended abstract of [LWM04c].

[LWM03b] Henk-Jan Lebbink, Cilia Witteman, and John-Jules Ch. Meyer. Extended abstract: A dialogue game to agree to disagree. In *First European Workshop on Multi-Agent Systems*, Oxford, UK, December 2003.

[LWM03c] Henk-Jan Lebbink, Cilia Witteman, and John-Jules Ch. Meyer. A Dialogue Game to Agree to Disagree about Inconsistent Information . In Ivana Kruijff-Korbayová and Claudia Kosny, editors, *7th Workshop on the Semantics and Pragmatics of Dialogue (Diabruck'03)*, pages 83–90, Wallerfangen, Germany, September 4-6 2003.

[LWM04a] Henk-Jan Lebbink, Cilia Witteman, and John-Jules Ch. Meyer. A dialogue game to offer an agreement to disagree. In *Second International Workshop for Programming Multi-Agent Systems: Language and Tools (ProMAS'04)*, pages 103–114, New York, NY, USA, 2004. Held with the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04).

[LWM04b] Henk-Jan Lebbink, Cilia Witteman, and John-Jules Ch. Meyer. A dialogue game to offer an agreement to disagree. In *Third international joint conference on Autonomous Agents & Multi-Agent Systems (AAMAS'04)*, pages 1238–1239, New York, NY, USA, July 19–23 2004.

[LWM04c] Henk-Jan Lebbink, Cilia Witteman, and John-Jules Ch. Meyer. Dialogue Games for Inconsistent and Biased Information. *Electronic Lecture Notes of Theoretical Computer Science*, 85(2), 2004.

[LWM05] Henk-Jan Lebbink, Cilia Witteman, and John-Jules Ch. Meyer. A dialogue game for belief revision in multi-agent systems. In *Fourth International Workshop on Agent Communication (AC'05)*, Netherlands, Utrecht, 2005. Held with the Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'05).

[Mar88] S. Marcus, editor. *Automating Knowledge Acquisition for Expert Systems*. Kluwer Academic Publishers, Boston, 1988.

[McC80] John McCarthy. Circumscription–a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980. Reprinted in [McC90].

[McC90] John McCarthy. *Formalization of common sense, papers by John McCarthy edited by V. Lifschitz*. Ablex, 1990.

[McD82] John P. McDermott. R1: A rule-based configurer of computer systems. *Artificial Intelligence*, 19(1):39–88, 1982.

[MCD96] Bernard Moulin and Brahim Chaib-Draa. An overview of distributed artificial intelligence. In Gregory M.P. O'Hare and Nicholas R. Jennings, editors, *Foundations of Distributed Artificial Intelligence*, pages 3–55. John Wiley & Sons Inc., New York, 1996.

[MD80] D. McDermott and J. Doyle. Nonmonotinic logic I. *Artificial Intelligence*, 13:41–72, 1980.

[Men97] Louis Menand. *Pragmatism: A Reader*. Vintage Books, New York, 1997.

[MIG00] Eduardo Mena, Arantza Illarramendi, and Alfredo Goni. Automatic ontology construction for a multiagent-based software gathering service. In Klusch and Kerschberg [KK00], pages 232–243.

[Mou97] Bernard Moulin. The social dimension of interactions in multiagent systems. In Wayne Wobcke, Maurice Pagnucco, and Chengqi Zhang, editors, *Agents and Multi-Agent Systems Formalisms, Methodologies, and Applications*, volume 1441 of *LNCS*, pages 109–123. Springer, 1997.

[MP01a] Peter McBurney and Simon Parsons. Representing Epistemic Uncertainty by means of Dialectical Argumentation. *Annals of Mathematics and Artificial Intelligence, Special Issue on Representations of Uncertainty*, 32(1–4):125–169, 2001.

[MP01b] Peter McBurney and Simon Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language and Information*, 11(3):315–334, 2001. Special Issue on Logic and Games.

[MT91] W. Marek and M. Truszczyski. Autoepistemic logic. *Journal of the ACM*, 38:588–619, 1991.

[MvdH95] John-Jules Ch. Meyer and Wiebe van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, 1995.

[MvEPA03] Peter McBurney, Rogier M. van Eijk, Simon Parsons, and Leila Amgoud. A Dialogue game protocol for agent purchase negotiations. *Journal of Autonomous Agents and Multi-Agent Systems*, 7(3):232–273, 2003.

[NSD$^+$01] Natasha F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and Mark A. Musen. Creating semantic web contents with protege-2000. *IEEE Intelligent Systems*, 16(2):60–71, 2001.

[Orw48] George Orwell. *1984*. Available on `http://www.online-literature.com/orwell/1984/`, 1948.

[Pra00] Henry Prakken. On Dialogue Systems with Speech Acts, Arguments, and Counterarguments. In M. Ojeda-Aciego, M. I. P. de Guzman,

G. Brewka, and L. M. Pereira, editors, *Proceedings of the 7th European Workshop on Logic for Artificial Intelligence (JELIA'00)*, pages 224–238, Berlin, Germany, 2000. LNAI 1919, Springer Verlag.

[PS96] Henry Prakken and Giovanni Sartor. A dialectical model of assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law*, 4:331–368, 1996.

[PS98] Henry Prakken and Giovanni Sartor. Modelling Reasoning with Precedents in a Formal Dialogue Game. *Artificial Intelligence and Law*, 6:231–287, 1998.

[Put81] Hillary Putnam. *Reason, Truth and History*. Cambridge University Press, Cambridge, 1981.

[PV02] Henry Prakken and Gerard Vreeswijk. Logics for defeasible argumentation. In Dov Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, pages 218–319. Kluwer Academic Publishers, Dordrecht, second edition edition, 2002.

[PWA03] Simon Parsons, Michael J. Wooldridge, and Leila Amgoud. Properties and complexity of some formal inter-agent dialogues. *Journal of Logic and Computation*, 13(3):347–376, 2003.

[Ree98] Chris A. Reed. Dialogue frames in agent communication. In Y. Demazeau, editor, *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS 98)*, pages 246–253, Paris, 1998. IEEE Press.

[Rei78] R. Reiter. On closed world data bases. In Gallaire and Minker [GM78], pages 55–76.

[Rei80] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1–2):81–132, 1980.

[Rei87] R. Reiter. Nonmonotonic reasoning. *Annual Review of Computer Science*, 2(2):147–186, 1987.

[Res69] Nicholas Rescher. *Many-valued Logic*. McGraw-Hill, 1969.

[Res73] Nicholas Rescher. *The Coherence Theory of Truth*. Oxford University Press, Oxford, 1973.

[RG98] A. S. Rao and M. P. Georgeff. Decision procedures for bdi logics. *Journal of Logic and Computation*, 8(3):293–344, 1998.

[RN03] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (second edition)*. Prentice Hall, (first edition: 1995) 2003.

[Ros94] Leo Rosten. *Carnival of Wit : From Aristotle to Woody Allen*. Penguin Group, 1994.

[Sch72] Stephen Schiffer. *Meaning*. Oxford University Press, Oxford, 1972.

[Sch96] Andreas Schöter. Evidential Bilattice Logic and Lexical Inference. *Journal of Logic, Language and Information*, 5(1):65–105, April 1996.

[Sea69] John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK, 1969.

[Sea79] John R. Searle. A taxonomy of illocutionary acts. In *Expression and Meaning*. Cambridge University Press, Cambride, England, 1979.

[Sho76] E.H. Shortliffe. Computer based medical consultations: MYCIN. *American Elsevier*, 1976.

[Sin98] Munindar P. Singh. Agent communication languages: Rethinking the principles. *IEEE Computer*, 31(12):40–47, 1998.

[Sin00] Munindar P. Singh. A social semantics for agent communication languages. In Frank Dignum and M. Greaves, editors, *Issues in Agent Communication*, pages 31–45. Springer-Verlag, Heidelberg, Germany, 2000.

[Sin03] Munindar P. Singh. Agent communication languages: Rethinking the principles. In Marc-Philippe Huget, editor, *Communication in Multiagent Systems*, volume 2650 of *LNCS*, pages 37–50. Springer, 2003.

[Sow00] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole Pub Co, 2000.

[Ste95] Mark Steffik. *Introduction to Knowledge Systems*. Morgan Kaufmann Publishers, 1995.

[STT01] F. Sadri, F. Toni, and P. Torroni. Logic agents, dialogues and negotiation: and abductive approach. In M. Schroeder and K. Stathis, editors, *Proceedings of the Symposium on Information Agents for E-Commerce, Artificial Intelligence and the Simulation of Behaviour Conference (AISB 2001)*, York, UK, 2001. AISB.

[Sub97] Peter Suber. Non-contradiction and excluded middle. Available on `http://www.earlham.edu/~peters/courses/logsys/pnc-pem.htm`, 1997.

[SW01] Barry Smith and Chris Welty. Ontology: Towards a new synthesis. In Chris Welty and Barry Smith, editors, *Formal Ontology in Information Systems*, pages iii–x, Ongunquit, Maine, 2001. ACM Press.

[Syc98] Katia P. Sycara. Multiagent systems. *AI Magazine*, 19(2):79–92, 1998.

[Tar56] Alfred Tarski. The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*, pages 152–278. Clarendon Press, 1956. First published in 1933 in Polish.

[Tur50] Alan M. Turing. Computing machinery and intelligence. *MIND (the Journal of the Mind Association)*, LIX(236):433–60, 1950.

[UG96] Mike Uschold and Michael Gruninger. Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, 11(2), June 1996.

[Usc96] Mike Uschold. Building Ontologies: Towards a Unified Methodology. *Expert Systems '96 Conference in Cambridge, UK*, 1996.

[vNM44] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1953 edition, 1944.

[Vre97] Gerard A.W. Vreeswijk. Abstract argumentation systems. *Artificial Intelligence*, 90:225–279, 1997.

[Wei99] Gerhard Weiss. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. The MIT Press, Cambridge, Massachusetts, 1999.

[Wit72] Ludwig Wittgenstein. *On Certainty*. Parper Torchbooks, 1972. Translated by D. Paul and G.E.M Anscobe.

[Wit01] Ludwig Wittgenstein. *Philosophical Investigation*. Blackwell Publishers, UK, 2001. First published in 1953 in German.

[WJ95] Michael J. Wooldridge and Nicholas R. Jennings. Intelligent Agents: Theory and Practice. *The Knowledge Engineering Review*, 10(2):115–52, 1995.

[WJ97] Michael J. Wooldridge and Nicholas R. Jennings. Formalizing the cooperative problem solving process. In Huhns and Singh [HS97b], pages 430–40.

[WK95] Douglas N. Walton and Erik C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY series in Logic and Language. Albany, NY, USA, 1995.

[Woo00] Michael J. Wooldridge. Semantic issues in the verification of agent communication languages. *Agents and Multi-Agent Systems*, 3(1):9–31, 2000.

[Woo02] Michael J. Wooldridge. *An Introduction to Multiagent Systems*. John Wiley & Sons, Chichester, England, 2002.

[WPHT02] Michael Winikoff, Lin Padgham, James Harland, and John Thangarajah. Declarative & Procedural Goals in Intelligent Agent Systems. In *Proceedings of the Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR2002)*, Toulouse, France, April 22–25 2002.

[You02] James O. Young. The Coherence Theory of Truth. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, 2002.

Appendix A

Proofs

A.1 Useful Propositions

Proposition A.1. *For any complete bilattice $\mathcal{B}$, with any $\theta_1, \theta_2, \theta_3 \in \mathcal{B}$ we have:*

1. $\theta_1 \leq_k (\theta_2 \otimes_k \theta_3) \Rightarrow \theta_1 \leq_k \theta_2$
2. $\theta_1 \leq_k \theta_2 \Rightarrow \theta_1 \leq_k (\theta_2 \oplus_k \theta_3)$ *and* $\theta_1 \leq_k \theta_2 \Rightarrow (\theta_1 \oplus_k \theta_3) \leq_k (\theta_2 \oplus_k \theta_3)$
3. $\theta_1 \leq_k (\theta_2 \oplus_k \theta_1)$
4. $\theta_1 \leq_k \theta_2 \Rightarrow \theta_2 = (\theta_1 \oplus_k \theta_2)$

Proposition A.2. *If $\mathcal{T}$ is a complete theory and $p : \theta_1 \in max_k(\mathcal{T})$, then $p : (\theta_1 \oplus_k \theta_2) \in max_k(add(p : \theta_2, \mathcal{T}))$.*

Proof of proposition A.2, page 195. Suppose two complete theories: $\mathcal{T}$ with $p : \theta_1 \in max_k(\mathcal{T})$ and $\mathcal{T}' = add(p : \theta_2, \mathcal{T})$ with $p : \theta'_1 \in max_k(\mathcal{T}')$. To prove: $\theta'_1 = (\theta_1 \oplus_k \theta_2)$. From $p : \theta_1 \in \text{max}_k(\mathcal{T})$ and definition of max_k follows $p : \theta_1 \in \mathcal{T}$. From the definition of addition (def. 4.11) follows $p : \theta_1, p : \theta_2 \in \mathcal{T}'$. From $p : \theta_1, p : \theta_2 \in \mathcal{T}'$ and Cn is closed under rule R3 follows $p : (\theta_1 \oplus_k \theta_2) \in \mathcal{T}'$. From $p : \theta'_1 \in max_k(\mathcal{T}')$ and $p : (\theta_1 \oplus_k \theta_2) \in \mathcal{T}'$ follows $(\theta_1 \oplus_k \theta_2) \leq_k \theta'_1$. From $p : \theta_1 \in max_k(\mathcal{T})$ and $\mathcal{T}' = add(p : \theta_2, \mathcal{T})$ follows $\theta'_1 \leq_k (\theta_1 \oplus_k \theta_2)$. From $(\theta_1 \oplus_k \theta_2) \leq_k \theta'_1$ and $\theta'_1 \leq_k (\theta_1 \oplus_k \theta_2)$ follows $\theta'_1 = (\theta_1 \oplus_k \theta_2)$. ❑

A.2 Proofs from Chapter 4

Proof of proposition 4.1(1), page 44. Let $\mathcal{R} = \{\text{R1}, \text{R2}\}$. Suppose: $\mathcal{T} = Cn(\mathcal{T}, \mathcal{R})$. To prove: $\mathcal{T} = Cn(max_k(\mathcal{T}), \mathcal{R})$. From remark 4.2 follows $max_k(\mathcal{T}) \subseteq \mathcal{T}$. Because Cn is monotone (def. 4.5) and $max_k(\mathcal{T}) \subseteq \mathcal{T}$, it follows that $Cn(max_k(\mathcal{T}), \mathcal{R}) \subseteq Cn(\mathcal{T}, \mathcal{R})$. From $Cn(max_k(\mathcal{T}), \mathcal{R}) \subseteq Cn(\mathcal{T}, \mathcal{R})$ and $\mathcal{T} = Cn(\mathcal{T}, \mathcal{R})$ follows $Cn(max_k(\mathcal{T}), \mathcal{R}) \subseteq \mathcal{T}$.

Proof by way of contradiction; assume $\mathcal{T} \not\subseteq Cn(max_k(\mathcal{T}), \mathcal{R})$. From $\mathcal{T} \not\subseteq Cn(max_k(\mathcal{T}), \mathcal{R})$ follows that a $p : \theta$ exists with $p : \theta \in \mathcal{T}$ and $p : \theta \notin Cn(max_k(\mathcal{T}), \mathcal{R})$. From the definition of max_k (def. 4.8) and $p : \theta \in \mathcal{T}$ follows $p : \theta' \in max_k(\mathcal{T})$ with $\theta \leq_k \theta'$. From $\mathcal{T} \subseteq Cn(\mathcal{T}, \mathcal{R})$ and $p{:}\theta' \in max_k(\mathcal{T})$ follows $p{:}\theta' \in Cn(max_k(\mathcal{T}), \mathcal{R})$. From $\theta \leq_k \theta', p{:}\theta' \in Cn(max_k(\mathcal{T}), \mathcal{R})$ and Cn is closed under rule R2 follows $p{:}\theta \in Cn(max_k(\mathcal{T}), \mathcal{R})$. From $p{:}\theta \in Cn(max_k(\mathcal{T}, \mathcal{R})$ and $p{:}\theta \notin Cn(max_k(\mathcal{T}), \mathcal{R})$ follows a contradiction, thus assumption $\mathcal{T} \not\subseteq Cn(max_k(\mathcal{T}), \mathcal{R})$ is not valid, thus $\mathcal{T} \subseteq Cn(max_k(\mathcal{T}), \mathcal{R})$. From $Cn(max_k(\mathcal{T}), \mathcal{R}) \subseteq \mathcal{T}$ and $\mathcal{T} \subseteq Cn(max_k(\mathcal{T}), \mathcal{R})$ follows $\mathcal{T} = Cn(max_k(\mathcal{T}), \mathcal{R})$. ❑

Proof of proposition 4.1(2), page 44. Let $\mathcal{R} = \{\text{R1}, \text{R2}, \text{R3}\}$, we assume $\mathcal{T} = Cn(\mathcal{T}, \mathcal{R})$ and then prove $\mathcal{T} = Cn(max_k(\mathcal{T}), \mathcal{R})$. The proof is equal with $\mathcal{R} = \{\text{R1}, \text{R2}\}$. ❑

Proof of proposition 4.1(3), page 44. Let $\mathcal{R} = \{\text{R1d}, \text{R2d}\}$. Suppose: $\mathcal{T} = Cn(\mathcal{T}, \mathcal{R})$. To prove: $\mathcal{T} = Cn(min_k(\mathcal{T}), \mathcal{R})$. From remark 4.2 follows $min_k(\mathcal{T}) \subseteq \mathcal{T}$. Because Cn is monotone (def. 4.5) and $min_k(\mathcal{T}) \subseteq \mathcal{T}$, it follows that $Cn(min_k(\mathcal{T}), \mathcal{R}) \subseteq Cn(\mathcal{T}, \mathcal{R})$. From $Cn(min_k(\mathcal{T}), \mathcal{R}) \subseteq Cn(\mathcal{T}, \mathcal{R})$ and $\mathcal{T} = Cn(\mathcal{T}, \mathcal{R})$ follows $Cn(min_k(\mathcal{T}), \mathcal{R}) \subseteq \mathcal{T}$. Proof by way of contradiction; assume $\mathcal{T} \not\subseteq Cn(min_k(\mathcal{T}), \mathcal{R})$. From $\mathcal{T} \not\subseteq Cn(min_k(\mathcal{T}), \mathcal{R})$ follows that a $p : \theta$ exists with $p : \theta \in \mathcal{T}$ and $p : \theta \notin Cn(min_k(\mathcal{T}), \mathcal{R})$. From the definition of min_k (def. 4.9) and $p : \theta \in \mathcal{T}$ follows $p : \theta' \in min_k(\mathcal{T})$ with $\theta' \leq_k \theta$. From $\mathcal{T} \subseteq Cn(\mathcal{T}, \mathcal{R})$ and $p{:}\theta' \in min_k(\mathcal{T})$ follows $p{:}\theta' \in Cn(min_k(\mathcal{T}), \mathcal{R})$. From $\theta' \leq_k \theta, p{:}\theta' \in Cn(min_k(\mathcal{T}), \mathcal{R})$ and Cn is closed under rule R2d follows $p{:}\theta \in Cn(min_k(\mathcal{T}), \mathcal{R})$. From $p{:}\theta \in Cn(min_k(\mathcal{T}, \mathcal{R})$ and $p{:}\theta \notin Cn(min_k(\mathcal{T}), \mathcal{R})$ follows a contradiction, thus assumption $\mathcal{T} \not\subseteq Cn(min_k(\mathcal{T}), \mathcal{R})$ is not valid, thus $\mathcal{T} \subseteq Cn(min_k(\mathcal{T}), \mathcal{R})$. From $Cn(min_k(\mathcal{T}), \mathcal{R}) \subseteq \mathcal{T}$ and $\mathcal{T} \subseteq Cn(min_k(\mathcal{T}), \mathcal{R})$ follows $\mathcal{T} = Cn(min_k(\mathcal{T}), \mathcal{R})$. ❑

Proof of proposition 4.2, page 46. To prove: $add(\Psi_1, add(\Psi_2, \Psi_3)) = add(add(\Psi_1, \Psi_2), \Psi_3)$, i.e. addition is commutative. Instead of $Cn(\Psi, \mathcal{R})$ we will write $Cn(\Psi)$. (Special thanks to Albert Visser for his help on this proof.) From $(\Psi_1 \cup \Psi_2) \subseteq (\Psi_1 \cup \Psi_2 \cup \Psi_3)$ and Cn is monotone follows $Cn(\Psi_1 \cup \Psi_2) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$, and from the definition of addition (def. 4.11) follows $add(\Psi_1 \cup \Psi_2) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$. From $\Psi_3 \subseteq (\Psi_1 \cup \Psi_2 \cup \Psi_3)$ and Cn is increasing follows $\Psi_3 \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$. From $add(\Psi_1 \cup \Psi_2) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$ and $\Psi_3 \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$ follows $(add(\Psi_1, \Psi_2) \cup \Psi_3) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$. From $(add(\Psi_1, \Psi_2) \cup \Psi_3) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$ and Cn is monotone follows $Cn(add(\Psi_1, \Psi_2) \cup \Psi_3) \subseteq Cn(Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3))$. From $Cn(add(\Psi_1, \Psi_2) \cup \Psi_3) \subseteq Cn(Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3))$ and Cn is idempotent follows $Cn(add(\Psi_1, \Psi_2) \cup \Psi_3) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$. From $Cn(add(\Psi_1, \Psi_2) \cup \Psi_3) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$ and definition of addition follows $add(add(\Psi_1, \Psi_2), \Psi_3) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$. From Cn is increasing follows $(\Psi_1 \cup \Psi_2) \subseteq Cn(\Psi_1 \cup \Psi_2)$. From $(\Psi_1 \cup \Psi_2) \subseteq Cn(\Psi_1 \cup \Psi_2)$ and definition of addition follows $Cn(\Psi_1, \Psi_2) = add(\Psi_1, \Psi_2)$ follows $(\Psi_1 \cup \Psi_2) \subseteq add(\Psi_1, \Psi_2)$. From $(\Psi_1 \cup \Psi_2) \subseteq add(\Psi_1, \Psi_2)$ follows $(\Psi_1 \cup \Psi_2 \cup \Psi_3) \subseteq (add(\Psi_1, \Psi_2) \cup \Psi_3)$. From Cn is monotone and $(\Psi_1 \cup \Psi_2 \cup \Psi_3) \subseteq (add(\Psi_1, \Psi_2) \cup \Psi_3)$ follows $Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3) \subseteq Cn(add(\Psi_1, \Psi_2) \cup \Psi_3)$. From definition of addition follows $Cn(add(\Psi_1, \Psi_2) \cup \Psi_3) = add(add(\Psi_1, \Psi_2), \Psi_3)$. From $Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3) \subseteq Cn(add(\Psi_1, \Psi_2) \cup \Psi_3)$ and $Cn(add(\Psi_1, \Psi_2) \cup \Psi_3) = add(add(\Psi_1, \Psi_2), \Psi_3)$ follows $Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3) \subseteq add(add(\Psi_1, \Psi_2), \Psi_3)$. From $add(add(\Psi_1, \Psi_2), \Psi_3) \subseteq Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$ and $Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3) \subseteq$

$add(add(\Psi_1, \Psi_2), \Psi_3)$ follows $add(add(\Psi_1, \Psi_2), \Psi_3) = Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3)$. Thus we have: $add(add(\Psi_1, \Psi_2), \Psi_3) = Cn(\Psi_1 \cup \Psi_2 \cup \Psi_3) = Cn(\Psi_2 \cup \Psi_3 \cup \Psi_1) = add(add(\Psi_2, \Psi_3), \Psi_1) = add(\Psi_1, add(\Psi_2, \Psi_3))$. ❑

Proof of proposition 4.3, page 48. To prove: $retr(\Psi_1, retr(\Psi_2, \Psi_3)) = retr(retr(\Psi_1, \Psi_2), \Psi_3)$, i.e. retraction is commutative. (Special thanks to Albert Visser for his help on this proof.) Define $\theta_1 \ominus_k \theta_2 = \bigoplus_k(\{\theta' \in \mathcal{B} \mid (\theta' \leq_k \theta_1) \wedge (\theta_2 \not\leq_k \theta')\}$. This definition corresponds to the definition of retraction for complete theories on page 47 (with as difference that we ignore 'p' and that we allow infinite suprema). We have $\bigoplus_k(\{\theta' \in \mathcal{B} \mid (\theta' \leq_k \theta_1) \wedge (\theta_2 \not\leq_k \theta')\} = \bigoplus_k(\{\theta_1 \otimes \theta' \in \mathcal{B} \mid \theta_2 \not\leq_k \theta'\})$. With distributivity (remark 4.1) we have: $\bigoplus_k(\{\theta_1 \otimes_k \theta' \in \mathcal{B} \mid \theta_2 \not\leq_k \theta'\}) = \theta_1 \otimes_k \bigoplus_k(\{\theta' \in \mathcal{B} \mid \theta_2 \not\leq_k \theta'\})$, i.e. $\theta_1 \ominus_k \theta_2 = \theta_1 \otimes_k \bigoplus_k(\{\theta' \in \mathcal{B} \mid \theta_2 \not\leq_k \theta'\})$. With definition of $\ominus_k$ we have: $\theta_1 \ominus_k (\theta_2 \ominus_k \theta_3) = \theta_1 \ominus_k (\theta_2 \otimes_k \bigoplus_k(\{\theta'' \mid (\theta_3 \not\leq_k \theta''\})) = \theta_1 \otimes_k \bigoplus_k(\{\theta' \in \mathcal{B} \mid (\theta_2 \otimes_k \bigoplus_k(\{\theta'' \in \mathcal{B} \mid (\theta_3 \not\leq_k \theta''\})) \not\leq_k \theta'\})$. With distributivity we have: $\theta_1 \ominus_k (\theta_2 \otimes_k \bigoplus_k(\{\theta'' \in \mathcal{B} \mid (\theta_3 \not\leq_k \theta''\})) = \theta_1 \otimes_k \bigoplus_k(\{\theta' \in \mathcal{B} \mid (\theta_2 \otimes_k \bigoplus_k(\{\theta'' \mid (\theta_3 \not\leq_k \theta''\})) \not\leq_k \theta'\})) = \theta_1 \otimes_k \bigoplus_k(\{\theta' \in \mathcal{B} \mid \theta_2 \not\leq_k \theta'\}) \otimes_k \bigoplus_k(\{\theta'' \in \mathcal{B} \mid \theta_3 \not\leq_k \theta''\}))$. With definition of $\ominus_k$ we have: $(\theta_1 \otimes_k \bigoplus_k(\{\theta' \in \mathcal{B} \mid \theta_2 \not\leq_k \theta'\})) \otimes_k \bigoplus_k(\{\theta'' \in \mathcal{B} \mid \theta_3 \not\leq_k \theta''\}) = (\theta_1 \ominus_k \theta_2) \otimes_k \bigoplus_k(\{\theta'' \in \mathcal{B} \mid \theta_3 \not\leq_k \theta''\}) = (\theta_1 \ominus_k \theta_2) \ominus_k \theta_3$. Thus we have: $(\theta_1 \ominus_k \theta_2) \ominus_k \theta_3 = \theta_1 \ominus_k (\theta_2 \ominus_k \theta_3)$. ❑

A.3 Proofs from Chapter 5

Proof of proposition 5.1, page 75. If $\mathcal{M}$ is a coherent mental state structure (def. 5.10), then $add_{ms}(\psi, \mathcal{T})(\mathcal{M})$ (def. 5.13) yields a coherent mental state structure, for any $\mathcal{T} \in MSN$.

1. $add_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.8) holds:
   - If $\mathcal{T} = B_s B_j$ then $\mathcal{M}' \models \psi \in B_s B_j \wedge \psi \notin B_s \widetilde{B}_j$ (def. 5.13(1)).
   - If $\mathcal{T} = B_s \widetilde{B}_j$ then $\mathcal{M}' \models \psi \in B_s \widetilde{B}_j \wedge \psi \notin B_s B_j$ (def. 5.13(1)).
   - If $\mathcal{T} = B_s D_j B_a$ then $\mathcal{M}' \models \psi \in B_s D_j B_a \wedge \psi \notin B_s \widetilde{D}_j B_a$ (def. 5.13(5)).
   - If $\mathcal{T} = B_s D_j \widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_s D_j \widetilde{B}_a \wedge \psi \notin B_s \widetilde{D}_j \widetilde{B}_a$ (def. 5.13(5)).
   - If $\mathcal{T} = B_s \widetilde{D}_j B_a$ then $\mathcal{M}' \models \psi \in B_s \widetilde{D}_j B_a \wedge \psi \notin B_s D_j B_a$ (def. 5.13(6)).
   - If $\mathcal{T} = B_s \widetilde{D}_j \widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_s \widetilde{D}_j \widetilde{B}_a \wedge \psi \notin B_s D_j \widetilde{B}_a$ (def. 5.13(6)).
2. $add_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.9) holds:
   - If $\mathcal{T} = B_s B_j B_s$ then $\mathcal{M}' \models \psi \in B_s B_j B_s \wedge \psi \in B_s \widetilde{B}_j \widetilde{B}_s$ (def. 5.13(2)).
   - If $\mathcal{T} = B_s B_j D_s B_a$ then $\mathcal{M}' \models \psi \in B_s B_j D_s B_a \wedge \psi \in B_s \widetilde{B}_j \widetilde{D}_s B_a$ (def. 5.13(9)).
   - If $\mathcal{T} = B_s B_j D_s \widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_s B_j D_s \widetilde{B}_a \wedge \psi \in B_s \widetilde{B}_j \widetilde{D}_s \widetilde{B}_a$ (def. 5.13(10)).
3. $add_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.10) holds:

- If $\mathcal{T} = B_sB_j\widetilde{B}_s$ then $\mathcal{M}' \models \psi \in B_sB_j\widetilde{B}_s \land \psi \in B_s\widetilde{B}_jB_s$ (def. 5.13(3)).
- If $\mathcal{T} = B_sB_j\widetilde{D}_sB_a$ then $\mathcal{M}' \models \psi \in B_sB_j\widetilde{D}_sB_a \land \psi \in B_s\widetilde{B}_jD_sB_a$ (def. 5.13(11)).
- If $\mathcal{T} = B_sB_j\widetilde{D}_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_sB_j\widetilde{D}_s\widetilde{B}_a \land \psi \in B_s\widetilde{B}_jD_s\widetilde{B}_a$ (def. 5.13(11)).

4. $add_{ms}(\psi,\mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.11) holds:

   - If $\mathcal{T} = B_s\widetilde{B}_j\widetilde{B}_s$ then $\mathcal{M}' \models \psi \in B_s\widetilde{B}_j\widetilde{B}_s \land \psi \notin B_sB_j\widetilde{B}_s$ (def. 5.13(7)).
   - If $\mathcal{T} = B_sB_j\widetilde{B}_s$ then $\mathcal{M}' \models \psi \in B_sB_j\widetilde{B}_s \land \psi \notin B_s\widetilde{B}_j\widetilde{B}_s$ (def. 5.13(3)).
   - If $\mathcal{T} = B_s\widetilde{B}_j\widetilde{D}_sB_a$ then $\mathcal{M}' \models \psi \in B_s\widetilde{B}_j\widetilde{D}_sB_a \land \psi \notin B_sB_j\widetilde{D}_sB_a$ (def. 5.13(7)).
   - If $\mathcal{T} = B_sB_j\widetilde{D}_sB_a$ then $\mathcal{M}' \models \psi \in B_sB_j\widetilde{D}_sB_a \land \psi \notin B_s\widetilde{B}_j\widetilde{D}_sB_a$ (def. 5.13(11)).
   - If $\mathcal{T} = B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a \land \psi \notin B_sB_j\widetilde{D}_s\widetilde{B}_a$ (def. 5.13(7)).
   - If $\mathcal{T} = B_sB_j\widetilde{D}_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_sB_j\widetilde{D}_s\widetilde{B}_a \land \psi \notin B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a$ (def. 5.13(11)).

5. $add_{ms}(\psi,\mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.12) holds:

   - If $\mathcal{T} = B_s\widetilde{B}_jB_s$ then $\mathcal{M}' \models \psi \in B_s\widetilde{B}_jB_s \land \psi \notin B_sB_jB_s$ (def. 5.13(8)).
   - If $\mathcal{T} = B_sB_jB_s$ then $\mathcal{M}' \models \psi \in B_sB_jB_s \land \psi \notin B_s\widetilde{B}_jB_s$ (def. 5.13(2)).
   - If $\mathcal{T} = B_s\widetilde{B}_jD_sB_a$ then $\mathcal{M}' \models \psi \in B_s\widetilde{B}_jD_sB_a \land \psi \notin B_sB_jD_sB_a$ (def. 5.13(8)).
   - If $\mathcal{T} = B_sB_jD_sB_a$ then $\mathcal{M}' \models \psi \in B_sB_jD_sB_a \land \psi \notin B_s\widetilde{B}_jD_sB_a$ (def. 5.13(9)).
   - If $\mathcal{T} = B_s\widetilde{B}_jD_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_s\widetilde{B}_jD_s\widetilde{B}_a \land \psi \notin B_sB_jD_s\widetilde{B}_a$ (def. 5.13(8)).
   - If $\mathcal{T} = B_sB_jD_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_sB_jD_s\widetilde{B}_a \land \psi \notin B_s\widetilde{B}_jD_s\widetilde{B}_a$ (def. 5.13(10)).

6. $add_{ms}(\psi,\mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which assumption 5.1 holds:

   - If $\mathcal{T} = D_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \in D_s\widetilde{B}_a \land \psi \notin D_sB_a$ (def. 5.13(4)).
   - If $\mathcal{T} = D_sB_a$ then $\mathcal{M}' \models \psi \in D_sB_a \land \psi \notin D_s\widetilde{B}_a$ (def. 5.13(4)).

7. $add_{ms}(\psi,\mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.13) holds:

   - If $\mathcal{T} = B_sD_jB_a$ then $\mathcal{M}' \models \psi \in B_sD_jB_a \land \psi \notin B_sD_j\widetilde{B}_a$ (def. 5.13(5)).
   - If $\mathcal{T} = B_sD_j\widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_sD_j\widetilde{B}_a \land \psi \notin B_sD_jB_a$ (def. 5.13(5)).

8. $add_{ms}(\psi,\mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.14) holds:

   - If $\mathcal{T} = B_sB_jD_sB_a$ then $\mathcal{M}' \models \psi \in B_sB_jD_sB_a \land \psi \in B_s\widetilde{B}_jD_s\widetilde{B}_a$ (def. 5.13(9)).

9. $add_{ms}(\psi,\mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.15) holds:

   - If $\mathcal{T} = B_sB_jD_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \in B_sB_jD_s\widetilde{B}_a \land \psi \in B_s\widetilde{B}_jD_sB_a$ (def. 5.13(10)).

From 1 to 9 follows that $\mathcal{M}'$ is a coherent mental state structure. ❏

Proof of proposition 5.2, page 75. If $\mathcal{M}$ is a coherent mental state structure (def. 5.10), then $retr_{ms}(\psi,\mathcal{T})(\mathcal{M})$ (def. 5.14) yields a coherent mental state structure, for any $\mathcal{T} \in MSN$.

1. $retr_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.9) holds:
   - if $\mathcal{T} = B_s\widetilde{B}_j\widetilde{B}_s$ then $\mathcal{M}' \models \psi \notin B_s\widetilde{B}_j\widetilde{B}_s \wedge \psi \notin B_sB_jB_s$ (def. 5.14(1)).
   - if $\mathcal{T} = B_s\widetilde{B}_j\widetilde{D}_sB_a$ then $\mathcal{M}' \models \psi \notin B_s\widetilde{B}_j\widetilde{D}_sB_a \wedge \psi \notin B_sB_jD_sB_a$ (def. 5.14(1)).
   - if $\mathcal{T} = B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \notin B_s\widetilde{B}_j\widetilde{D}_s\widetilde{B}_a \wedge \psi \notin B_sB_jD_s\widetilde{B}_a$ (def. 5.14(1)).
2. $retr_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.10) holds:
   - if $\mathcal{T} = B_s\widetilde{B}_jB_s$ then $\mathcal{M}' \models \psi \notin B_s\widetilde{B}_jB_s \wedge \psi \notin B_sB_j\widetilde{B}_s$ (def. 5.14(2)).
   - if $\mathcal{T} = B_s\widetilde{B}_jD_sB_a$ then $\mathcal{M}' \models \psi \notin B_s\widetilde{B}_jD_sB_a \wedge \psi \notin B_sB_j\widetilde{D}_sB_a$ (def. 5.14(4)).
   - if $\mathcal{T} = B_s\widetilde{B}_jD_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \notin B_s\widetilde{B}_jD_s\widetilde{B}_a \wedge \psi \notin B_sB_j\widetilde{D}_s\widetilde{B}_a$ (def. 5.14(3)).
3. $retr_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.14) holds:
   - if $\mathcal{T} = B_s\widetilde{B}_jD_s\widetilde{B}_a$ then $\mathcal{M}' \models \psi \notin B_s\widetilde{B}_jD_s\widetilde{B}_a \wedge \psi \notin B_sB_jD_sB_a$ (def. 5.14(3)).
4. $retr_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a $\mathcal{M}'$ in which equation (5.15) holds:
   - if $\mathcal{T} = B_s\widetilde{B}_jD_sB_a$ then $\mathcal{M}' \models \psi \notin B_s\widetilde{B}_jD_sB_a \wedge \psi \notin B_sB_jD_s\widetilde{B}_a$ (def. 5.14(4)).

From 1 to 4 follows that $\mathcal{M}'$ is a coherent mental state structure. ❏

Proof of proposition 5.3, page 76. If $\mathcal{M}$ is a coherent mental state structure (def. 5.10), then $update_s(\Pi, \mathcal{M})$ yields a coherent mental state structure, for any coherent subset of $\mathcal{L}'_{MS}$ Π. For all $\pi \in \Pi$ we have: either $\pi = \psi \in \mathcal{T}$ with $\mathcal{T} \in MSN$ and $\psi \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$, then $add_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a coherent mental state structure (proposition 5.1), or $\pi = \psi \notin \mathcal{T}$ with $\mathcal{T} \in MSN$ and $\psi \in \mathcal{L}_{\mathcal{B},\mathcal{F}}$, then $retr_{ms}(\psi, \mathcal{T})(\mathcal{M})$ yields a coherent mental state structure (proposition 5.2). ❏

A.4 Proofs from Chapter 7

Proof of proposition 7.1, page 144. Suppose: $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_sB_jB_s \in \mathcal{M}_s$ and $B_jB_s \in \mathcal{M}_j$ with $p : \theta_4 \in max_k(B_sB_jB_s)$ and $p : \theta_5 \in max_k(B_jB_s)$. Also suppose that Sarah has uttered $\lambda(s, j, p : \theta')$, then, after update of both Sarah's and John's mental state structures, we have $\mathcal{M}'_s$ and $\mathcal{M}'_j$ with $B_sB_jB'_s \in \mathcal{M}'_s$ and $B_jB'_s \in \mathcal{M}'_j$ with $p : \theta'_4 \in max_k(B_sB_jB'_s)$ and $p : \theta'_5 \in max_k(B_jB'_s)$. Assume: $\theta_5 \leq_k \theta_4$, i.e. $B_sB_jB_s \subseteq B_jB_s$. To prove: $\theta'_5 \leq_k \theta'_4$, i.e. $B_sB_jB'_s \subseteq B_jB'_s$.

1. $\lambda = qba_1$. From $post(\lambda(s, j, p : \theta'))$ follows $p : \theta' \in B_j\widetilde{B}_s$, i.e. $\mathcal{M}'_j = add_{ms}(p : \theta', B_j\widetilde{B}_s)(\mathcal{M}_j)$. From definition 5.13(2) follows $p : \theta' \notin B_jB_s$, thus $\theta' \nleq_k \theta'_5$. Because nothing is added to B_jB_s, it follows that $\theta'_5 \leq_k \theta_5$. From $\theta'_5 \leq_k \theta_5$ and $\theta_5 \leq_k \theta_4$ follows $\theta'_5 \leq_k \theta_4$. Because $B_sB_jB_s$ is unaltered, $\theta_4 = \theta'_4$. From $\theta'_5 \leq_k \theta_4$ and $\theta_4 = \theta'_4$ follows $\theta'_5 \leq_k \theta'_4$.
2. $\lambda = qba_2$. Proof is equal to 1.

3. $\lambda = gqba_1$. From $post(\lambda(s,j,p:\theta'))$ follows $\mathcal{M}'_j = add_{ms}(p:\theta', B_jB_s)(\mathcal{M}_j)$ and $\mathcal{M}'_s = add_{ms}(p:\theta', B_sB_jB_s)(\mathcal{M}_s)$. From proposition A.2 follows $\theta'_4 = (\theta_4 \oplus_k \theta')$. From proposition A.2 follows $\theta'_5 = (\theta_5 \oplus_k \theta')$. From proposition A.1(2) and $\theta_5 \leq_k \theta_4$ follows $(\theta_5 \oplus_k \theta') \leq_k (\theta_4 \oplus_k \theta')$. From $(\theta_5 \oplus_k \theta') \leq_k (\theta_4 \oplus_k \theta')$, $\theta'_4 = (\theta_4 \oplus_k \theta')$ and $\theta'_5 = (\theta_5 \oplus_k \theta')$ follows $\theta'_5 \leq_k \theta'_4$.

For other λ, the proof goes analogously. ❑

Proof of proposition 7.2, page 145. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_jB_s \in \mathcal{M}_s$ with $p:\theta_1 \in max_k(B_s)$ and $p:\theta_4 \in max_k(B_sB_jB_s)$. Assume: $\mathcal{M}_s \not\models pre(is(s,j,p:\theta',\widetilde{B}_s))$. To prove: $\mathcal{M}_s \models p:\theta' \in B_sB_jB_s \Rightarrow \mathcal{M}_s \models p:\theta' \in B_s$, i.e. $\theta' \leq_k \theta_4 \Rightarrow \theta' \leq_k \theta_1$, i.e. $\theta_4 \leq_k \theta_1$. From $\mathcal{M}_s \not\models pre(is(s,j,\psi,\widetilde{B}_s))$ follows that $\theta_4 \leq_k \theta_1$. ❑

Proof of proposition 7.3, page 145. From proposition 7.1 follows $\mathcal{M}_j \models \psi \in B_jB_s \Rightarrow \mathcal{M}_s \models B_sB_jB_s$. From proposition 7.2 follows $\mathcal{M}_s \models \psi \in B_sB_jB_s \Rightarrow \mathcal{M}_s \models \psi \in B_s$. From $\mathcal{M}_j \models \psi \in B_jB_s \Rightarrow \mathcal{M}_s \models B_sB_jB_s$ and $\mathcal{M}_s \models \psi \in B_sB_jB_s \Rightarrow \mathcal{M}_s \models \psi \in B_s$ follows $\mathcal{M}_j \models \psi \in B_jB_s \Rightarrow \mathcal{M}_s \models \psi \in B_s$. ❑

Proof of proposition 7.4, page 145. Suppose: $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_sB_j\widetilde{B}_s \in \mathcal{M}_s$ and $B_j\widetilde{B}_s \in \mathcal{M}_j$. Also suppose that Sarah has uttered $\lambda(s,j,p:\theta')$, then, after update of both Sarah's and John's mental state structures, we have $\mathcal{M}'_s$ and $\mathcal{M}'_j$ with $B_sB_j\widetilde{B}'_s \in \mathcal{M}'_s$ and $B_j\widetilde{B}'_s \in \mathcal{M}'_j$. Assume: for all $p:\tau_3 \in B_sB_j\widetilde{B}_s$ follows $p:\tau_3 \in B_j\widetilde{B}_s$, i.e. $B_sB_j\widetilde{B}_s \subseteq B_j\widetilde{B}_s$. To prove: for all $p:\tau'_3 \in B_sB_j\widetilde{B}'_s$ follows $p:\tau'_3 \in B_j\widetilde{B}'_s$, i.e. $B_sB_j\widetilde{B}'_s \subseteq B_j\widetilde{B}'_s$. For all λ, the proof goes analogously to the proof of proposition 7.1. ❑

Proof of proposition 7.5, page 145. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j\widetilde{B}_s \in \mathcal{M}_s$ with $p:\theta_1 \in max_k(B_s)$ and $p:\tau_3 \in B_sB_jB_s$. Assume: $\mathcal{M}_s \not\models pre(is(s,j,p:\theta',B_s))$. To prove: $\mathcal{M}_s \models p:\theta' \in B_sB_jB_s \Rightarrow \mathcal{M}_s \models p:\theta' \notin B_s$, i.e. $\tau_3 \leq_k \theta' \Rightarrow \theta' \nleq_k \theta_1$, i.e. $\tau_3 \nleq_k \theta_1$. From $\mathcal{M}_s \not\models pre(is(s,j,\psi,B_s))$ follows that $\tau_3 \nleq_k \theta_1$. ❑

Proof of proposition 7.6, page 145. From proposition 7.4 follows $\mathcal{M}_j \models \psi \in B_j\widetilde{B}_s \Rightarrow \mathcal{M}_s \models B_sB_jB_s$. From proposition 7.5 follows $\mathcal{M}_s \models \psi \in B_sB_j\widetilde{B}_s \Rightarrow \mathcal{M}_s \models \psi \notin B_s$. From $\mathcal{M}_j \models \psi \in B_j\widetilde{B}_s \Rightarrow \mathcal{M}_s \models B_sB_j\widetilde{B}_s$ and $\mathcal{M}_s \models \psi \in B_sB_j\widetilde{B}_s \Rightarrow \mathcal{M}_s \models \psi \notin B_s$ follows $\mathcal{M}_j \models \psi \in B_j\widetilde{B}_s \Rightarrow \mathcal{M}_s \models \psi \notin B_s$. ❑

A.5 Proofs from Chapter 8

Proof of proposition 8.1, page 152. Suppose: $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_s \in \mathcal{M}_s$ and $B_j \in \mathcal{M}_j$ with $p:\theta_1 \in max_k(B_s)$ and $p:\theta_2 \in max_k(B_j)$. Also suppose: $\mathcal{M}'_s = add_{ms}(p:\xi_1, B_s)(\mathcal{M}_s)$ with $B'_s \in \mathcal{M}'_s$ with $p:\theta'_1 \in max_k(B'_s)$. To prove: $\theta'_1 \geqslant_k \theta_2$. From proposition A.2 follows $\theta'_1 = (\theta_1 \oplus_k \xi_1)$. Proof by way of contradiction; assume $\theta_2 \nleq_k (\theta_1 \oplus_k \xi_1)$. By definition of ξ_1 (eq. (8.2)), $\xi_1 \equiv \otimes_k(\{\theta' \in \mathcal{B} \mid (\theta_1 \oplus_k \theta') = (\theta_1 \oplus_k \theta_2)\})$ follows $(\theta_1 \oplus_k \xi_1) = (\theta_1 \oplus_k \theta_2)$. From $\theta_2 \nleq_k (\theta_1 \oplus_k \xi_1)$ and $(\theta_1 \oplus_k \xi_1) = (\theta_1 \oplus_k \theta_2)$ follows $\theta_2 \nleq_k (\theta_1 \oplus_k \theta_2)$. From proposition A.1(3) and $\theta_2 \nleq_k (\theta_1 \oplus_k \theta_2)$ follows a contradiction, thus assumption

$\theta_2 \not\leq_k (\theta_1 \oplus_k \xi_1)$ is not valid, thus $\theta_2 \leq_k (\theta_1 \oplus_k \xi_1)$. From $\theta_2 \leq_k (\theta_1 \oplus_k \xi_1)$ and $\theta'_1 = (\theta_1 \oplus_k \xi_1)$ follows $\theta_2 \leq_k \theta'_1$. From $\theta_2 \leq_k \theta'_1$ follows $\theta'_1 \gtrless_k \theta_2$. ❏

Proof of proposition 8.2, page 153. Suppose: $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_s \in \mathcal{M}_s$ and $B_j \in \mathcal{M}_j$ with $p : \theta_1 \in max_k(B_s)$ and $p : \theta_2 \in max_k(B_j)$. Also suppose: $\mathcal{M}'_j = add_{ms}(p : \xi_2, B_j)(\mathcal{M}_j)$ with $B'_j \in \mathcal{M}'_j$ with $p : \theta'_2 \in max_k(B'_j)$. To prove: $\theta'_2 \gtrless_k \theta_1$. From proposition A.2 follows $\theta'_2 = (\theta_2 \oplus_k \xi_2)$. Proof by way of contradiction; assume $\theta_1 \not\leq_k (\theta_2 \oplus_k \xi_2)$. By definition of ξ_2 (eq. (8.3)), $\xi_2 \equiv \bigotimes_k(\{\theta' \in \mathcal{B} \mid (\theta_2 \oplus_k \theta') = (\theta_1 \oplus_k \theta_2)\})$ follows $(\theta_2 \oplus_k \xi_2) = (\theta_1 \oplus_k \theta_2)$. From $\theta_1 \not\leq_k (\theta_2 \oplus_k \xi_2)$ and $(\theta_2 \oplus_k \xi_2) = (\theta_1 \oplus_k \theta_2)$ follows $\theta_1 \not\leq_k (\theta_1 \oplus_k \theta_2)$. From proposition A.1(3) and $\theta_1 \not\leq_k (\theta_1 \oplus_k \theta_2)$ follows a contradiction, thus assumption $\theta_1 \not\leq_k (\theta_2 \oplus_k \xi_2)$ is not valid, thus $\theta_1 \leq_k (\theta_2 \oplus_k \xi_2)$. From $\theta_1 \leq_k (\theta_2 \oplus_k \xi_2)$ and $\theta'_2 = (\theta_2 \oplus_k \xi_2)$ follows $\theta_1 \leq_k \theta'_2$. From $\theta_1 \leq_k \theta'_2$ follows $\theta'_2 \gtrless_k \theta_1$. ❏

Proof of proposition 8.3, page 153. Suppose: $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_s \in \mathcal{M}_s$ and $B_j \in \mathcal{M}_j$ with $p : \theta_1 \in max_k(B_s)$ and $p : \theta_2 \in max_k(B_j)$. Also suppose: $\mathcal{M}'_s = retr_{ms}(p : \eta_1, B_s)(\mathcal{M}_s)$ with $B'_s \in \mathcal{M}'_s$ with $p : \theta'_1 \in max_k(B'_s)$. To prove: $\theta'_1 \gtrless_k \theta_2$. From the definition of retraction follows $p : \eta_1 \notin B'_s$, thus $\eta_1 \not\leq_k \theta'_1$. Because nothing is added to B_s, it follows that $\theta'_1 \leq_k \theta_1$. Proof by way of contradiction; assume $\theta'_1 \not\leq_k \theta_2$. From the contraposition of proposition A.1(1) follows that from $\theta'_1 \not\leq_k \theta_2$ follows $\theta'_1 \not\leq_k (\theta_1 \otimes_k \theta_2)$. From $\theta'_1 \leq_k \theta_1$ and $\theta'_1 \not\leq_k (\theta_1 \otimes_k \theta_2)$, and the definition of η_1 (eq. (8.4)), $\eta_1 \equiv \bigotimes_k(\{\theta' \in \mathcal{B} \mid (\theta' \leq_k \theta_1) \wedge (\theta' \not\leq_k (\theta_1 \otimes_k \theta_2))\})$, it follows that $\eta_1 = (\eta_1 \otimes_k \theta'_1)$, i.e. $\eta_1 \leq_k \theta'_1$. From $\eta_1 \leq_k \theta'_1$ and $\eta_1 \not\leq_k \theta'_1$ follows a contradiction, thus assumption $\theta'_1 \not\leq_k \theta_2$ is not valid, thus $\theta'_1 \leq_k \theta_2$. From $\theta'_1 \leq_k \theta_2$ follows $\theta'_1 \gtrless_k \theta_2$. ❏

Proof of proposition 8.4, page 154. Suppose: $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_s \in \mathcal{M}_s$ and $B_j \in \mathcal{M}_j$ with $p : \theta_1 \in max_k(B_s)$ and $p : \theta_2 \in max_k(B_j)$. Also suppose: $\mathcal{M}'_j = retr_{ms}(p : \eta_2, B_j)(\mathcal{M}_j)$ with $B'_j \in \mathcal{M}'_j$ with $p : \theta'_2 \in max_k(B'_j)$. To prove: $\theta'_2 \gtrless_k \theta_1$. From the definition of retraction follows $p : \eta_2 \notin B'_j$, thus $\eta_2 \not\leq_k \theta'_2$. Because nothing is added to B_j, it follows that $\theta'_2 \leq_k \theta_2$. Proof by way of contradiction; assume $\theta'_2 \not\leq_k \theta_1$. From the contraposition of proposition A.1(1) follows that from $\theta'_2 \not\leq_k \theta_1$ follows $\theta'_2 \not\leq_k (\theta_1 \otimes_k \theta_2)$. From $\theta'_2 \leq_k \theta_2$ and $\theta'_2 \not\leq_k (\theta_1 \otimes_k \theta_2)$, and the definition of η_2 (eq. (8.4)), $\eta_2 \equiv \bigotimes_k(\{\theta' \in \mathcal{B} \mid (\theta' \leq_k \theta_2) \wedge (\theta' \not\leq_k (\theta_1 \otimes_k \theta_2))\})$, it follows that $\eta_2 = (\eta_2 \otimes_k \theta'_2)$, i.e. $\eta_2 \leq_k \theta'_2$. From $\eta_2 \leq_k \theta'_2$ and $\eta_2 \not\leq_k \theta'_2$ follows a contradiction, thus assumption $\theta'_2 \not\leq_k \theta_1$ is not valid, thus $\theta'_2 \leq_k \theta_1$. From $\theta'_2 \leq_k \theta_1$ follows $\theta'_2 \gtrless_k \theta_1$. ❏

Proof of proposition 8.5, page 155. Suppose: $\mathcal{M}_s$ and $\mathcal{M}_j$ with $B_s, B_sB_j, B_s\widetilde{B_s} \in \mathcal{M}_s$ and $B_j \in \mathcal{M}_j$ with $p : \theta_1 \in max_k(B_s)$, $p : \theta_2 \in max_k(B_j)$, $p : \theta_3 \in max_k(B_sB_j)$ and $p : \tau_2 \in B_s\widetilde{B_j}$. Assume: $\theta_1 \not\gtrless_k \theta_3$ and $\tau_2 \leq_k (\theta_1 \oplus_k \theta_3)$, i.e. $\mathcal{M}_s \models$ equation (8.6). To prove: $\theta_1 \not\gtrless_k \theta_2$, i.e. Sarah and John disagree about the truth value of p. From proposition 7.1 follows $\theta_3 \leq_k \theta_2$. From proposition 7.4 follows $\tau_2 \not\leq_k \theta_2$. Proof by way of contradiction; assume $\theta_1 \leq_k \theta_2$. From proposition A.1(2) and $\theta_1 \leq_k \theta_2$ follows $(\theta_1 \oplus_k \theta_3) \leq_k (\theta_2 \oplus_k \theta_3)$. From $\tau_2 \leq_k (\theta_1 \oplus_k \theta_3)$ and $(\theta_1 \oplus_k \theta_3) \leq_k (\theta_2 \oplus_k \theta_3)$ follows $\tau_2 \leq_k (\theta_2 \oplus_k \theta_3)$. From proposition A.1(4) and $\theta_3 \leq_k \theta_2$ follows $\theta_2 = (\theta_2 \oplus_k \theta_3)$. From $\tau_2 \leq_k (\theta_2 \oplus_k \theta_3)$ and $\theta_2 = (\theta_2 \oplus_k \theta_3)$ follows $\tau_2 \leq_k \theta_2$. From $\tau_2 \leq_k \theta_2$ and $\tau_2 \not\leq_k \theta_2$ follows a contradiction, thus assumption $\theta_1 \leq_k \theta_2$ is not valid, thus $\theta_1 \nleq_k \theta_2$. Proof by way of contradiction; assume $\theta_2 \leq_k \theta_1$. From $\theta_3 \leq_k \theta_2$ and $\theta_2 \leq_k \theta_1$ follows $\theta_3 \leq_k \theta_1$. From $\theta_1 \not\gtrless_k \theta_3$ follows $\theta_3 \nleq_k \theta_1$. From $\theta_3 \leq_k \theta_1$ and $\theta_3 \nleq_k \theta_1$ follows a contradiction, thus assumption $\theta_2 \leq_k \theta_1$ is not valid, thus $\theta_2 \nleq_k \theta_1$. From $\theta_1 \nleq_k \theta_2$ and $\theta_2 \nleq_k \theta_1$ follows $\theta_1 \not\gtrless_k \theta_2$. ❑

Proof of proposition 8.6, page 157. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j, B_s\widetilde{B}_s, B_sB_jB_s, B_sB_j\widetilde{B}_{\underline{s}} \in \mathcal{M}_s$ with $p : \theta_1 \in \underset{\sim}{m}ax_k(B_s)$, $p : \theta_3 \in max_k(B_sB_j)$, $p : \theta_4 \in max_k(B_sB_jB_s)$, $p : \tau_2 \in B_s\widetilde{B}_j$ and $p : \tau_3 \in B_sB_j\widetilde{B}_s$. Assume: $\theta_3 \not\gtrless_k \theta_4$, $\tau_2 \leq_k (\theta_3 \oplus_k \theta_4)$ and $\tau_3 \leq_k (\theta_3 \oplus_k \theta_4)$, i.e. $\mathcal{M}_s \models$ equation (8.7). To prove: $\theta_1 \not\gtrless_k \theta_3$ and $\tau_2 \leq_k (\theta_3 \oplus_k \theta_1)$, i.e. $\mathcal{M}_s \models$ equation (8.6). From proposition 7.2 follows $\theta_4 \leq_k \theta_1$. From proposition 7.5 follows $\tau_3 \nleq_k \theta_1$. From the coherence relation as given in equation (5.3) follows $\tau_2 \nleq_k \theta_3$. Proof by way of contradiction; assume $\theta_1 \leq_k \theta_3$. From $\theta_4 \leq_k \theta_1$ and $\theta_1 \leq_k \theta_3$ follows $\theta_4 \leq_k \theta_3$. From proposition A.1(4) and $\theta_4 \leq_k \theta_3$ follows $\theta_3 = (\theta_3 \oplus_k \theta_4)$. From $\tau_2 \leq_k (\theta_3 \oplus_k \theta_4)$ and $\theta_3 = (\theta_3 \oplus_k \theta_4)$ follows $\tau_2 \leq_k \theta_3$. From $\tau_2 \leq_k \theta_3$ and $\tau_2 \nleq_k \theta_3$ follows a contradiction, thus assumption $\theta_1 \leq_k \theta_3$ is not valid, thus $\theta_1 \nleq_k \theta_3$. Proof by way of contradiction; assume $\theta_3 \leq_k \theta_1$. From $\theta_3 \leq_k \theta_1$ follows $(\theta_3 \oplus_k \theta_4) \leq_k (\theta_1 \oplus_k \theta_4)$. From $\tau_3 \leq_k (\theta_3 \oplus_k \theta_4)$ and $(\theta_3 \oplus_k \theta_4) \leq_k (\theta_1 \oplus_k \theta_4)$ follows $\tau_3 \leq_k (\theta_1 \oplus_k \theta_4)$. From $\theta_4 \leq_k \theta_1$ follows $\theta_1 = (\theta_1 \oplus_k \theta_4)$. From $\tau_3 \leq_k (\theta_1 \oplus_k \theta_4)$ and $\theta_1 = (\theta_1 \oplus_k \theta_4)$ follows $\tau_3 \leq_k \theta_1$. From $\tau_3 \leq_k \theta_1$ and $\tau_3 \nleq_k \theta_1$ follows a contradiction, thus assumption $\theta_3 \leq_k \theta_1$ is not valid, thus $\theta_3 \nleq_k \theta_1$. From $\theta_1 \nleq_k \theta_3$ and $\theta_3 \nleq_k \theta_1$ follows $\theta_1 \not\gtrless_k \theta_3$. Proof $\tau_2 \leq_k (\theta_3 \oplus_k \theta_1)$. From proposition A.1(2) and $\theta_4 \leq_k \theta_1$ follows $(\theta_3 \oplus_k \theta_4) \leq_k (\theta_3 \oplus_k \theta_1)$. From $\tau_2 \leq_k (\theta_3 \oplus_k \theta_4)$ and $(\theta_3 \oplus_k \theta_4) \leq_k (\theta_3 \oplus_k \theta_1)$ follows $\tau_2 \leq_k (\theta_3 \oplus_k \theta_1)$. ❑

Proof of proposition 8.7, page 158. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j \in \mathcal{M}_s$ with $p : \theta_1 \in max_k(B_s)$ and $p : \theta_3 \in max_k(B_sB_j)$. Also suppose $\mathcal{M}'_s = add_{ms}(p : \xi_3, B_s)(\mathcal{M}_s)$ with $B'_s \in \mathcal{M}'_s$ with $p : \theta'_1 \in max_k(B'_s)$. To prove: $\theta'_1 \gtrless_k \theta_3$. From proposition A.2 follows $\theta'_1 = (\theta_1 \oplus_k \xi_3)$. Proof by way of contradiction; assume $\theta_3 \nleq_k (\theta_1 \oplus_k \xi_3)$. By definition of ξ_3 (eq. (8.8)), $\xi_3 \equiv \bigotimes_k(\{\theta' \in \mathcal{B} \mid (\theta_1 \oplus_k \theta') = (\theta_1 \oplus_k \theta_3)\})$ follows $(\theta_1 \oplus_k \xi_3) = (\theta_1 \oplus_k \theta_3)$. From $\theta_3 \nleq_k (\theta_1 \oplus_k \xi_3)$ and $(\theta_1 \oplus_k \xi_3) = (\theta_1 \oplus_k \theta_3)$ follows $\theta_3 \nleq_k (\theta_1 \oplus_k \theta_3)$. From proposition A.1(3) and $\theta_3 \nleq_k (\theta_1 \oplus_k \theta_3)$ follows a contradiction, thus assumption $\theta_3 \nleq_k (\theta_1 \oplus_k \xi_3)$ is not valid, thus $\theta_3 \leq_k (\theta_1 \oplus_k \xi_3)$. From $(\theta_1 \oplus_k \xi_3) = (\theta_1 \oplus_k \theta_3)$ and $\theta_3 \leq_k (\theta_1 \oplus_k \xi_3)$ follows $\theta_3 \leq_k (\theta_1 \oplus_k \theta_3)$. From $\theta_3 \leq_k (\theta_1 \oplus_k \theta_3)$ and $\theta'_1 = (\theta_1 \oplus_k \xi_3)$ follows $\theta_3 \leq_k \theta'_1$. From $\theta_3 \leq_k \theta'_1$ follows $\theta'_1 \gtrless_k \theta_3$. ❑

Proof of proposition 8.8, page 158. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j, B_sB_jB_s \in \mathcal{M}_s$ with $p : \theta_1 \in max_k(B_s)$, $p : \theta_3 \in max_k(B_sB_j)$ and $p : \theta_4 \in max_k(B_sB_jB_s)$. Also suppose $\mathcal{M}'_s = add_{ms}(p : \xi_3, B_sB_jB_s)(\mathcal{M}_s)$ with $B_sB_jB'_s \in \mathcal{M}'_s$ with $p : \theta'_4 \in max_k(B_sB_jB'_s)$. To prove: $\theta'_4 \gtrless_k \theta_3$. From proposition 7.2 follows $\theta_4 \leq_k \theta_1$. From proposition A.2 follows $\theta'_4 = (\theta_4 \oplus_k \xi_3)$. By definition of ξ_3 (eq. (8.8)), $\xi_3 \equiv \bigotimes_k(\{\theta' \in \mathcal{B} \mid (\theta_1 \oplus_k \theta') = (\theta_1 \oplus_k \theta_3)\})$ follows $(\theta_1 \oplus_k \xi_3) = (\theta_1 \oplus_k \theta_3)$. From proposition A.1(2) and $\theta_4 \leq_k \theta_1$ follows $(\theta_4 \oplus_k \xi_3) \leq_k (\theta_1 \oplus_k \xi_3)$. Proof by way of contradiction; assume $\theta_3 \nleq_k (\theta_4 \oplus_k \xi_3)$. From $\theta_3 \nleq_k (\theta_4 \oplus_k \xi_3)$ and $(\theta_4 \oplus_k \xi_3) \leq_k (\theta_1 \oplus_k \xi_3)$ follows $\theta_3 \nleq_k (\theta_1 \oplus_k \xi_3)$. From $\theta_3 \nleq_k (\theta_1 \oplus_k \xi_3)$ and $(\theta_1 \oplus_k \xi_3) = (\theta_1 \oplus_k \theta_3)$ follows $\theta_3 \nleq_k (\theta_1 \oplus_k \theta_3)$. From proposition A.1(3) and $\theta_3 \nleq_k (\theta_1 \oplus_k \theta_3)$ follows a contradiction, thus assumption $\theta_3 \nleq_k (\theta_4 \oplus_k \xi_3)$ is not valid, thus $\theta_3 \leq_k (\theta_4 \oplus_k \xi_3)$. From $\theta_3 \leq_k (\theta_4 \oplus_k \xi_3)$ and $\theta'_4 = (\theta_4 \oplus_k \xi_3)$ follows $\theta_3 \leq_k \theta'_4$. From $\theta_3 \leq_k \theta'_4$ follows $\theta'_4 \gtrless_k \theta_3$. ❑

Proof of proposition 8.9, page 158. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j, B_s\widetilde{B}_j \in \mathcal{M}_s$ with $p:\theta_1 \in max_k(B_s)$, $p{:}\theta_3 \in max_k(B_sB_j)$ and $p{:}\tau_2 \in B_s\widetilde{B}_j$. Also suppose $\mathcal{M}'_s = add_{ms}(p{:}\xi_2, B_sB_j)(\mathcal{M}_s)$ with $B_sB'_j \in \mathcal{M}'_s$ with $p:\theta'_3 \in max_k(B_sB'_j)$. Assume: $\theta_1 \not\gtrless_k \theta_3$ and $\tau_2 \leq_k (\theta_1 \oplus_k \theta_3)$, i.e. $\mathcal{M}_s \models$ equation (8.6). To prove: $\theta_1 \gtrless_k \theta'_3$. Suppose $\theta_1 = b_1 \times d_1$, $\theta_2 = b_2 \times d_2$, $\theta_3 = b_3 \times d_3$, $\theta'_3 = b'_3 \times d'_3$ and $\xi_2 = b_\xi \times d_\xi$. From proposition A.2 follows $\theta'_3 = (\theta_3 \oplus_k \xi_2)$. From equation 4.6 follows $(\theta_3 \oplus_k \xi_2) = (b_3 \sqcup_b b_\xi \times d_3 \sqcup_d d_\xi)$. From $\theta_1 \not\gtrless_k \theta_3$, i.e. $\theta_1 \not\leq_k \theta_3$ and $\theta_3 \not\leq_k \theta_1$, follows $\theta_1 \leq_t \theta_3$ or $\theta_3 \leq_t \theta_1$. From $\theta_1 \leq_t \theta_3$ or $\theta_3 \leq_t \theta_1$ and equation 4.2 follows ($b_1 \leq_b b_3$ and $d_3 \leq_d d_1$) or ($b_3 \leq_b b_1$ and $d_1 \leq_d d_3$). Assume $b_1 \leq_b b_3$ and $d_3 \leq_d d_1$. From proposition 7.6 follows $\theta_3 \leq_k \theta_2$. From equation 4.1 and $\theta_3 \leq_k \theta_2$ follows $b_3 \leq_b b_2$ and $d_3 \leq_d d_2$. From definition of ξ_2 (eq. (8.3)), $\xi_2 \equiv \bigotimes_k(\{\theta' \in \mathcal{B} \mid (\theta_2 \oplus_k \theta') = (\theta_1 \oplus_k \theta_2)\})$ follows $(\theta_2 \oplus_k \xi_2) = (\theta_1 \oplus_k \theta_2)$. From $(\theta_2 \oplus_k \xi_2) = (\theta_1 \oplus_k \theta_2)$ and equation 4.6 follows $(b_2 \sqcup_b b_\xi \times d_2 \sqcup_d d_\xi) = (b_1 \sqcup_b b_2 \times d_1 \sqcup_d d_2)$. From $(b_2 \sqcup_b b_\xi \times d_2 \sqcup_d d_\xi) = (b_1 \sqcup_b b_2 \times d_1 \sqcup_d d_2)$ follows $b_2 \sqcup_b b_\xi = b_2 \sqcup_b b_1$ and $d_2 \sqcup_d d_\xi = d_2 \sqcup_d d_1$. Proof by way of contradiction; assume $d_2 \not\leq_d d_\xi$. From $d_2 \sqcup_d d_\xi = d_2 \sqcup_d d_1$ and $d_2 \not\leq_d d_\xi$ follows $d_2 = d_2 \sqcup_d d_1$. From $d_2 = d_2 \sqcup_d d_1$ follows $d_1 \leq_d d_2$. From $b_1 \leq_b b_3$ and $b_3 \leq_b b_2$ follows $b_1 \leq_b b_2$. From $b_1 \leq_b b_2$ and $d_1 \leq_d d_2$ follows $b_1 \times d_1 \leq_k b_2 \times d_2$, i.e. $\theta_1 \leq_k \theta_2$. From $\theta_1 \leq_k \theta_2$, $\theta_3 \leq_k \theta_2$, and equation 4.1 follows $(\theta_1 \oplus_k \theta_3) \leq_k \theta_2$. From $\tau_2 \leq_k (\theta_1 \oplus_k \theta_3)$ and $(\theta_1 \oplus_k \theta_3) \leq_k \theta_2$ follows $\tau_2 \leq_k \theta_2$. From proposition 7.4 follows $\tau_2 \not\leq_k \theta_2$. From $\tau_2 \not\leq_k \theta_2$ and $\tau_2 \leq_k \theta_2$ follows a contradiction, thus assumption $d_2 \not\leq_d d_\xi$ is not valid, thus $d_2 \leq_d d_\xi$. From $d_2 \sqcup_d d_\xi = d_2 \sqcup_d d_1$ and $d_2 \leq_d d_\xi$ follows $d_\xi = d_1$. From $d_\xi = d_1$ follows $d_1 \leq_d d_\xi$. From $d_1 \leq_d d_\xi$ follows $d_1 \leq_d d_3 \sqcup_d d_\xi$. From $b_1 \leq_b b_3$ follows $b_1 \leq_b b_3 \sqcup_b b_\xi$. From $b_1 \leq_b b_3 \sqcup_b b_\xi$, $d_1 \leq_d d_3 \sqcup_d d_\xi$ and equation 4.1 follows $b_1 \times d_1 \leq_k (b_3 \sqcup_b b_\xi \times d_3 \sqcup_d d_\xi)$, i.e. $\theta_1 \leq_k \theta_3$. From $\theta_1 \leq_k \theta_3$ follows $\theta_1 \gtrless_k \theta_3$. ❑

Proof of proposition 8.10, page 159. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j, B_sB_jB_s \in \mathcal{M}_s$ with $p:\theta_1 \in max_k(B_s)$, $p:\theta_3 \in max_k(B_sB_j)$ and $p:\theta_4 \in max_k(B_sB_jB_s)$. Also suppose $\mathcal{M}'_s = add_{ms}(p:\xi_4, B_sB_j)(\mathcal{M}_s)$ with $B_sB'_j \in \mathcal{M}'_s$ with $p:\theta'_3 \in max_k(B_sB'_j)$. To prove: $\theta_4 \gtrless_k \theta'_3$. From proposition A.2 follows $\theta'_3 = (\theta_3 \oplus_k \xi_4)$. Proof by way of contradiction; assume $\theta_4 \not\leq_k (\theta_3 \oplus_k \xi_4)$. By definition of ξ_4 (eq. (8.9)), $\xi_4 \equiv \bigotimes_k(\{\theta' \in \mathcal{B} \mid (\theta_3 \oplus_k \theta') = (\theta_4 \oplus_k \theta_3)\})$ follows $(\theta_3 \oplus_k \xi_4) = (\theta_4 \oplus_k \theta_3)$. From $\theta_4 \not\leq_k (\theta_3 \oplus_k \xi_4)$ and $(\theta_3 \oplus_k \xi_4) = (\theta_4 \oplus_k \theta_3)$ follows $\theta_4 \not\leq_k (\theta_4 \oplus_k \theta_3)$. From proposition A.1(3) and $\theta_4 \not\leq_k (\theta_4 \oplus_k \theta_3)$ follows a contradiction, thus assumption $\theta_4 \not\leq_k (\theta_3 \oplus_k \xi_4)$ is not valid, thus $\theta_4 \leq_k (\theta_3 \oplus_k \xi_4)$. From $\theta'_3 = (\theta_3 \oplus_k \xi_4)$ and $\theta_4 \leq_k (\theta_3 \oplus_k \xi_4)$ follows $\theta_4 \leq_k \theta'_3$. From $\theta_4 \leq_k \theta'_3$ follows $\theta_4 \gtrless_k \theta'_3$. ❑

Proof of proposition 8.11, page 160. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j \in \mathcal{M}_s$ with $p:\theta_1 \in max_k(B_s)$ and $p:\theta_3 \in max_k(B_sB_j)$. Also suppose $\mathcal{M}'_s = retr_{ms}(p:\eta_3, B_s)(\mathcal{M}_s)$ with $B'_s \in \mathcal{M}'_s$ with $p:\theta'_1 \in max_k(B'_s)$. To prove: $\theta'_1 \gtrless_k \theta_3$. From the definition of retraction follows $p:\eta_3 \notin B'_s$, thus $\eta_3 \not\leq_k \theta'_1$. Because nothing is added to B'_s, it follows that $\theta'_1 \leq_k \theta_1$. Proof by way of contradiction; assume $\theta'_1 \not\leq_k \theta_3$. From the contraposition of proposition A.1(1) follows that from $\theta'_1 \not\leq_k \theta_3$ follows $\theta'_1 \not\leq_k (\theta_1 \otimes_k \theta_3)$. From $\theta'_1 \leq_k \theta_1$ and $\theta'_1 \not\leq_k (\theta_1 \otimes_k \theta_3)$, and the definition of η_3

(eq. (8.10)), $\eta_3 \equiv \bigotimes_k(\{\theta' \in \mathscr{B} \mid (\theta' \leq_k \theta_1) \wedge (\theta' \not\leq_k (\theta_1 \otimes_k \theta_3))\})$, it follows that $\eta_3 = (\eta_3 \otimes_k \theta'_1)$, i.e. $\eta_3 \leq_k \theta'_1$. From $\eta_3 \leq_k \theta'_1$ and $\eta_3 \not\leq_k \theta'_1$ follows a contradiction, thus assumption $\theta'_1 \not\leq_k \theta_3$ is not valid, thus $\theta'_1 \leq_k \theta_3$. From $\theta'_1 \leq_k \theta_3$ follows $\theta'_1 \gtrless_k \theta_3$. ❏

Proof of proposition 8.12, page 160. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j, B_sB_jB_s \in \mathcal{M}_s$ with $p{:}\theta_1 \in max_k(B_s)$, $p : \theta_3 \in max_k(B_sB_j)$ and $p : \theta_4 \in max_k(B_sB_jB_s)$. Also suppose $\mathcal{M}'_s = retr_{ms}(p : \eta_3, B_sB_jB_s)(\mathcal{M}_s)$ with $B_sB_jB'_s \in \mathcal{M}'_s$ with $p{:}\theta'_4 \in max_k(B_sB_jB'_s)$. To prove: $\theta'_4 \gtrless_k \theta_3$. From the definition of retraction follows $p : \eta_3 \notin B_sB_jB'_s$, thus $\eta_3 \not\leq_k \theta'_4$. Because nothing is added to $B_sB_jB'_s$, it follows that $\theta'_4 \leq_k \theta_4$. From proposition 7.2 follows $\theta_4 \leq_k \theta_1$. From $\theta'_4 \leq_k \theta_4$ and $\theta_4 \leq_k \theta_1$ follows $\theta'_4 \leq_k \theta_1$. Proof by way of contradiction; assume $\theta'_4 \not\leq_k \theta_3$. From the contraposition of proposition A.1(1) follows that from $\theta'_4 \not\leq_k \theta_3$ follows $\theta'_4 \not\leq_k (\theta_1 \otimes_k \theta_3)$. From $\theta'_4 \leq_k \theta_1$ and $\theta'_4 \not\leq_k (\theta_1 \otimes_k \theta_3)$, and the definition of η_3 (eq. (8.10)), $\eta_3 \equiv \bigotimes_k(\{\theta' \in \mathscr{B} \mid (\theta' \leq_k \theta_1) \wedge (\theta' \not\leq_k (\theta_1 \otimes_k \theta_3))\})$, it follows that $\eta_3 = (\eta_3 \otimes_k \theta'_4)$, i.e. $\eta_3 \leq_k \theta'_4$. From $\eta_3 \leq_k \theta'_4$ and $\eta_3 \not\leq_k \theta'_4$ follows a contradiction, thus assumption $\theta'_4 \not\leq_k \theta_3$ is not valid, thus $\theta'_4 \leq_k \theta_3$. From $\theta'_4 \leq_k \theta_3$ follows $\theta'_4 \gtrless_k \theta_3$. ❏

Proof of proposition 8.13, page 161. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j \in \mathcal{M}_s$ with $p : \theta_1 \in max_k(B_s)$ and $p : \theta_3 \in max_k(B_sB_j)$. Also suppose $\mathcal{M}'_s = retr_{ms}(p : \eta_2, B_sB_j)(\mathcal{M}_s)$ with $B_sB'_j \in \mathcal{M}'_s$ with $p : \theta'_3 \in max_k(B_sB'_j)$. To prove: $\theta_1 \gtrless_k \theta'_3$. From the definition of retraction follows $p : \eta_2 \notin B_sB_j$, thus $\eta_2 \not\leq_k \theta'_3$. Because nothing is added to $B_sB'_j$, it follows that $\theta'_3 \leq_k \theta_3$. From proposition 7.1 follows $\theta_3 \leq_k \theta_2$. From $\theta'_3 \leq_k \theta_3$ and $\theta_3 \leq_k \theta_2$ follows $\theta'_3 \leq_k \theta_2$. Proof by way of contradiction; assume $\theta'_3 \not\leq_k \theta_1$. From the contraposition of proposition A.1(1) follows that from $\theta'_3 \not\leq_k \theta_1$ follows $\theta'_3 \not\leq_k (\theta_1 \otimes_k \theta_2)$. From $\theta'_3 \leq_k \theta_2$ and $\theta'_3 \not\leq_k (\theta_1 \otimes_k \theta_2)$, and the definition of η_2 (eq. (8.5)), $\eta_2 \equiv \bigotimes_k(\{\theta' \in \mathscr{B} \mid (\theta' \leq_k \theta_2) \wedge (\theta' \not\leq_k (\theta_1 \otimes_k \theta_2))\})$, it follows that $\eta_2 = (\eta_2 \otimes_k \theta'_3)$, i.e. $\eta_2 \leq_k \theta'_3$. From $\eta_2 \leq_k \theta'_3$ and $\eta_2 \not\leq_k \theta'_3$ follows a contradiction, thus assumption $\theta'_3 \not\leq_k \theta_1$ is not valid, thus $\theta'_3 \leq_k \theta_1$. From $\theta_1 \leq_k \theta'_3$ follows $\theta_1 \gtrless_k \theta'_3$. ❏

Proof of proposition 8.14, page 161. Suppose: $\mathcal{M}_s$ with $B_s, B_sB_j, B_sB_jB_s \in \mathcal{M}_s$ with $p : \theta_1 \in max_k(B_s)$, $p : \theta_3 \in max_k(B_sB_j)$ and $p : \theta_4 \in max_k(B_sB_jB_s)$. Also suppose $\mathcal{M}'_s = retr_{ms}(p : \eta_4, B_sB_j)(\mathcal{M}_s)$ with $B'_j \in \mathcal{M}'_s$ with $p : \theta'_3 \in max_k(B_sB'_j)$. To prove: $\theta_4 \gtrless_k \theta'_3$. From the definition of retraction follows $p : \eta_4 \notin B_sB'_j$, thus $\eta_4 \not\leq_k \theta'_3$. Because nothing is added to $B_sB'_j$, it follows that $\theta'_3 \leq_k \theta_3$. Proof by way of contradiction; assume $\theta'_3 \not\leq_k \theta_4$. From the contraposition of proposition A.1(1) follows that from $\theta'_3 \not\leq_k \theta_4$ follows $\theta'_3 \not\leq_k (\theta_3 \otimes_k \theta_4)$. From $\theta'_3 \leq_k \theta_3$ and $\theta'_3 \not\leq_k (\theta_3 \otimes_k \theta_4)$, and the definition of η_4 (eq. (8.11)), $\eta_4 \equiv \bigotimes_k(\{\theta' \in \mathscr{B} \mid (\theta' \leq_k \theta_3) \wedge (\theta' \not\leq_k (\theta_3 \otimes_k \theta_4))\})$, it follows that $\eta_4 = (\eta_4 \otimes_k \theta'_3)$, i.e. $\eta_4 \leq_k \theta'_3$. From $\eta_4 \leq_k \theta'_3$ and $\eta_4 \not\leq_k \theta'_3$ follows a contradiction, thus assumption $\theta'_3 \not\leq_k \theta_4$ is not valid, thus $\theta'_3 \leq_k \theta_4$. From $\theta'_3 \leq_k \theta_4$ follows $\theta'_3 \gtrless_k \theta_4$. ❏

Bijlage B

Samenvatting

In dit proefschrift stellen we ons het doel de besluitvorming van agenten en de communicatie tussen deze agenten te beschrijven en te formaliseren opdat de agenten rechtvaardigbare opvattingen hebben en houden. Om deze beschrijvingen en formaliseringen te verkrijgen, behandelen we de volgende onderwerpen.

- In Hoofdstuk 2 beschrijven we dat een bewering de waarheid is voor onze agenten als aan de criteria is voldaan die volgens de subjectieve ervaring van de agenten normatief zijn voor het toetsen van de bewering. Met andere woorden, agenten kennen regels waarvan ze denken dat deze regels de conventies beschrijven wanneer ze mogen zeggen dat iets de waarheid is. Deze conventies gelden alleen binnen de gemeenschap van de agenten, de waarheid is dus subjectief voor een agent en intersubjectief voor een gemeenschap.

  We spreken over beweringen in plaats van uitspraken omdat agenten alleen beweringen kunnen uitdrukken met talige uitspraken. De beweringen die agenten kunnen uiten over cognitieve toestanden noemen we epistemische beweringen (Sectie 2.1). We bespreken vier verschillende waarheidstheorieën (Sectie 2.2). Als eerste bespreken we de correspondentietheorie. In deze theorie is de waarheid van een bewering afhankelijk van het bestaan van een feit in the wereld waarnaar de bewering refereert. Als tweede bespreken we de coherentietheorie. In deze theorie is de waarheid van een bewering afhankelijk van het bestaan van andere beweringen die coherent zijn met deze bewering. Als derde de consensustheorie. In deze theorie is de waarheid van een bewering afhankelijk van de overeenstemming van een groep. Als laatste bespreken we de waarheidstheorie volgens de interpretatie van Ellenbogen van Wittgenstein's *meaning is use* theorie. In deze theorie is de waarheid voor een agent een bewering die net als alle andere beweringen criteria heeft die het correcte gebruik beschrijven. De betekenis van beweringen is gelijk aan de conventies voor hun gebruik dat overeengekomen is door de agenten uit een gemeenschap.

De opvatting van een agent is een voorbeeld van een epistemische bewering. Net als alle andere beweringen heeft een epistemische bewering een betekenis die gelijk is aan de gebruiksconventies (Sectie 2.3). De kennis van onze agenten is gebaseerd op hun acceptatie van de talige conventies van hun gemeenschap. Omdat agenten de criteria van beweringen kennen om ze te kunnen gebruiken, kunnen ze de waarheidscondities van de beweringen toetsen aan hun cognitieve toestanden. Agenten kunnen over kennis beschikken, wetende dat als het gebruik van de beweringen verandert, ze hun kennis herzien.

In dit proefschrift beschrijven we, onafhankelijk van een domein, de criteria wanneer wij veronderstellen dat agenten gerechtvaardigde opvattingen hebben. Met andere woorden, we beschrijven de situaties waarin de cognitieve toestand van agenten expliciete rechtvaardiging geeft voor epistemische beweringen. De criteria zijn gebaseerd op de betekenis van beweringen. Dat wil zeggen, de gebruiksconventies beschrijven de specifieke criteria die moeten gelden voor een agent om gerechtvaardigd te zijn een epistemische bewering te hebben (Sectie 2.4).

- In Hoofdstuk 3 beschrijven we de notie van paraconsistentie in relatie tot de opvattingen van agenten. De samenvoeging van een bewering en de ontkenning van de bewering resulteert in een strijdige bewering. Eerst beschrijven we dat onze agenten die een op gebruik gebaseerde notie van betekenis hanteren in situaties kunnen komen waarin een strijdig wereldbeeld onafwendbaar is (Sectie 3.1). Andere bronnen van strijdigheid zijn bijvoorbeeld de leugenaarszin of sensorfouten (Sectie 3.2). Als een agent geconfronteerd wordt met een strijdige bewering, dan geeft de gebruiksinterpretatie van betekenis een samenhangende semantiek. In tegenstelling tot de correspondentie en de coherentie theorie, deze theorieën geven die geen samenhangende semantiek aan strijdige beweringen.

  Een logische theorie die een bewering en tegelijkertijd de ontkenning van de bewering waar maakt, is een onsamenhangende theorie (Sectie 3.4). Een onsamenhangende theorie maakt een strijdige bewering waar. Als de logische theorie toch het gebrek aan samenhang wil beschrijven, moet de theorie strijdige beweringen betekenis geven op een samenhangende manier. Deze theorieën worden paraconsistente theorieën genoemd.

- In Hoofdstuk 4 definiëren we een meerwaardige logica die we in Hoofdstuk 5 gebruiken bij de definitie van de cognitieve toestand van een agent. Voor onze meerwaardige logica gebruiken we bilattice structuren die in de literatuur worden beschreven als een gegeneraliseerde ruimte met waarheidswaarden met twee ordeningen. Deze ordeningen beschrijven de relaties tussen de waarheid en informatie van de waarheidswaarden (Sectie 4.1). De waarheidswaarden die wij gebruiken beschrijven informatie met een component die positieve steun symboliseert, en een component die negatieve steun symboliseert. De kleinste bilattice structuur heeft vier waarheidswaarden, te weten: waar, onwaar, inconsistent en 'geen informatie'. De waarheidswaarden waar en onwaar komen qua intuïtie overeen met de klassieke waarheidswaarden. De waarheidswaarde inconsistent komt overeen met een strijdige opvatting, en 'geen informatie' komt overeen met de afwezigheid van een klassieke waarheidswaarde. Bilattice structuren kunnen naast deze vier ook de tussenliggende waarheidswaarden hebben: de waarden die neigen naar waar of waarden die neigen naar onwaar.

Een beweringen van onze meerwaardige logica bestaan uit een formule uit een verzameling formules en een waarheidswaarde van een bilattice structuur (Sectie 4.2). Een meerwaardige theorie is een verzameling van deze beweringen die afgesloten is onder verschillende regels (Sectie 4.3). De meerwaardigheid uit zich dat een formule in deze theorieën meerdere waarheidswaarden kan hebben. We onderscheiden drie verschillende typen theorieën: normale, complete en inverse theorieën. We definiëren theoriebeschrijvingen die een compacte en eenduidige waargave geven van theorieën, deze beschrijvingen kunnen eenvoudig worden geprogrammeerd met computertalen (Sectie 4.4). Voor de verschillende typen theorieën worden twee acties gedefinieerd: het toevoegen en het verwijderen van beweringen. Daarnaast geven we procedures voor het uitvoeren van deze acties op theoriebeschrijvingen (Sectie 4.5).

- In Hoofdstuk 5 definiëren we onze cognitieve agenten. De cognitieve toestand van een agent bestaat uit een mentale toestand, een kennisbank en een epistemologie (Sectie 5.1). De mentale toestand bestaat uit een eindig aantal meerwaardige theorieën die de verschillende opvattingen van een agent beschrijven zoals overtuigingen, wensen en opvattingen die betrekking hebben op de opvattingen en wensen van andere agenten. De kennisbank is een verzameling van kennisregels die de kennis beschrijft die een agent als niet gemeenschappelijk beschouwt. De epistemologie is net als de kennisbank een verzameling kennisregels, maar met het verschil dat de agent de kennis wel als gemeenschappelijk beschouwt. De kennisregels beschrijven kennis die afkomstig is van een bepaalde menselijke expert of van een bepaald expertisedomein. Niet alle mogelijkheden van verschillende mentale toestanden zijn samenhangend; we beschrijven daarom de eigenschappen tussen verschillende mentale toestanden van een agent. Een agent kan niet de opvatting hebben dat een andere agent een bepaalde opvatting heeft, en tegelijkertijd de opvatting hebben dat deze agent niet deze opvatting heeft (Sectie 5.2). De twee acties voor het veranderen van meerwaardige theorieën worden uitgebreid met acties voor het veranderen van de mentale toestanden zodat de cognitieve toestand waarin deze mentale toestanden zich bevinden samenhangend blijft.

  In Hoofdstuk 6 en 7 worden beslis- en dialoogregels beschreven. De uitvoering van deze regels heeft naast het maken van een beslissing of het uiten van verbale communicatie het beoogde effect dat de mentale toestand van de agent verandert. De volgorde waarin de verschillende beslis- en dialoogregels de mentale toestanden kunnen veranderen is vastgelegd in de *deliberation cycle* van de agent (Sectie 5.3). In vogelvlucht: als een agent geen beslissingen meer kan maken, dan zal ze de binnengekomen communicatie afhandelen, als ook alle binnengekomen communicatie is afgehandeld, dan mag de agent uitgaande communicatie starten.

- In Hoofdstuk 6 beschrijven we beslisspellen als verzamelingen beslisregels. De situaties waarin correct beslissingen worden gemaakt worden beschreven door de criteria die agenten rechtvaardigt de conclusies van beslissingen te accepteren (Sectie 6.1). We gebruiken de criteria waarmee een agent gerechtvaardigd is een bepaalde cognitieve toestand te hebben, zoals het hebben van opvattingen, als de preconditie van een beslisregel waarmee de agent beslist deze cognitieve toestand te hebben. Daarnaast beschrijft de beslisregel welke condities gelden in de cognitieve toestand van de agent nadat ze een beslissing heeft genomen (Sectie 6.2). We definiëren twee verschillende beslisspellen. Het spel voor het aannemen van opvattingen, dat wil zeggen, het beslisspel beweringen aan de mentale toestand voor opvattingen toe te voegen. Het tweede spel is het duale spel voor het verwerpen van opvattingen, dat wil zeggen, het beslisspel dat beweringen uit de mentale toestand voor opvattingen verwijdert. Beide beslisspellen beschrijven het gebruik van verschillende beslissingen. Met een beslisregel definiëren we het gebruik, dat wil zeggen, de betekenis van afleidingsregels die deel uitmaken van de kennisbank van de agent (Sectie 6.3). Met een andere beslisregel definiëren we wanneer agenten zich mogen conformeren aan de opvattingen van andere agenten (Sectie 6.4). Met de laatste beslisregel definiëren we dat een agent alleen haar opvattingen mag veranderen als ze dat wenselijk acht (Sectie 6.5).

- In Hoofdstuk 7 beschrijven we dialoogspellen (dialogue games) als verzamelingen dialoogregels. Op dezelfde manier dat beslisspellen betekenis geven aan beslissingen, geven dialoogspellen betekenis aan communicatie tussen agenten. Wij beschrijven de criteria wanneer wij veronderstellen dat agenten correct, volgens hun conventies, *speech acts* mogen uiten (Sectie 7.1). We definiëren vijf dialoogspellen. Een spel met vragen voor rechtvaardigingen voor het hebben van een opvatting (Sectie 7.2). Een spel met vragen voor rechtvaardigingen voor het niet hebben van een opvatting (Sectie 7.3). Een spel met verzoeken dat de geadresseerde een opvatting aanneemt (Sectie 7.4). Een spel met verzoeken dat de geadresseerde een opvatting verwerpt (Sectie 7.6). Het laatste dialoogspel beschrijft de betekenis van communicatie waarin agenten elkaar informeren over veranderde opvattingen (Sectie 7.6).

- In Hoofdstuk 8 beschrijven we wanneer onze agenten gerechtvaardigd zijn het met elkaar eens te zijn dat ze het oneens zijn. Twee agenten zijn het oneens over een bewering als de waarheidswaarden van hun opvattingen niet overeenkomen (Sectie 8.1). Als agenten zich bewust worden dat ze het oneens zijn, kunnen ze de vier dialoogspellen uit Hoofdstuk 7 gebruiken om zichzelf en de ander te overtuigen van nieuwe informatie opdat hun opvattingen veranderen en ze het eens worden over de bewering (Sectie 8.2). Als de dialoogspellen niet geholpen hebben, rest de agenten niets anders dan te accepteren dat hun meningsverschil onoplosbaar is. Als agenten zich bewust worden dat hun meningsverschil onoplosbaar is, dan kunnen ze elkaar een voorstel doen het eens te worden dat ze het oneens zijn (Sectie 8.3). De interpretatie van een dergelijke acceptatie van een onoplosbaar meningsverschil is dat de agenten hun opvattingen hebben gebaseerd op informatie die ze gemeenschappelijk noemen, maar dat niet is (Sectie 8.4).

- In Hoofdstuk 9 beschrijven we onze eindconclusie en mogelijke onderwerpen voor vervolg onderzoek. Omdat onze beslisspellen definiëren wanneer agenten correcte beslissingen nemen die conform de conventies van hun gemeenschap zijn, zullen onze agenten alleen nieuwe opvattingen kunnen afleiden op basis van het gangbare handelen.

  Agenten uit bijvoorbeeld de medische wetenschappen kunnen alleen nieuwe opvattingen verkrijgen volgens de conventies van de medische wetenschappen. Hetzelfde geldt voor dialoogspellen: agenten communiceren volgens de gangbare gebruiken van de gemeenschap van medische wetenschappen. Als een agent kennis heeft die rechtstreeks afkomstig is van een medische expert, dan kan de agent alleen een nieuwe opvatting verkrijgen die conform de gebruiken van de medische gemeenschap is. In het geval dat deze gemeenschap niet de opvattingen van een agent deelt, dan is, in principe, de gemeenschap te overtuigen. Dit omdat de agent heeft gehandeld conform de heersende opvattingen.

  Een bruikbare uitbreiding op de vijf dialoogspellen zou het spel zijn waarin betekenis wordt gegeven wanneer een agent mag vragen waarom andere agenten een bepaalde opvatting wel of niet hebben. Met dergelijke uitingen kan een agent informeren naar de rechtvaardigingen die andere agenten hebben gebruikt om bepaalde opvattingen te hebben. Een andere bruikbare uitbreiding zou het dialoogspel zijn waarin de agent vraagt naar de criteria die een andere agent hanteert voor het rechtvaardigen van opvattingen. Met andere woorden, een dialoogspel waarin agenten informeren naar de betekenis, dat wil zeggen, het gebruik van opvattingen.

De besluitvorming van onze agenten en de communicatie tussen deze agenten zijn een afspiegeling van de mogelijke conventies zoals die heersen in de gemeenschap die de agenten vertegenwoordigen. We zijn erin geslaagd een multiagent systeem te beschrijven waarin de opvattingen van agenten conform de conventies zijn en blijven.

Bijlage C

Dankwoord

Soms heb je het druk en blijven er klussen liggen. Even tijd inruimen voor de afwas of de plinten in mijn huis doe ik niet meer. Wellicht doen gezegdes uit grootmoeders tijd nog steeds opgeld. Van uitstel komt afstel bijvoorbeeld. Toch hangt dit tegeltje hier niet, en dat gaan we even bewijzen. Een dankwoord is net een afwasmachine, maar wel met tenminste één groot verschil. Welnu, als ik het voor het zeggen had, dan wordt afwas verboden. Nu wil het feit dat ik hierover niet zoveel te zeggen heb, dus heb ik een afwasmachine aangeschaft. Geen uitstel noch afstel. Nu had ik natuurlijk kunnen weten dat een afwasvoorbeeld nooit een tegenvoorbeeld voor een tegeltje geeft. Daar is groter geschut voor nodig. Ja, jij! De ingrediënten zijn de volgende: een computer, een idee, steun en vier jaar. De combinatie steun en vier jaar typen achter een computer met een-of-ander idee creëren de ideale voedingsbodem voor een proefschrift, maar ook voor uitstel—vier jaar uitstel van een dankbetuiging. Zonder jullie steun geen tegenvoorbeeld en geen proefschrift! Nu de inhaalslag.

Onder de bevlogen leiding van Adriaan Tuiten heb ik de afgelopen vier jaar met plezier bij Emotional Brain gewerkt en ik hoop dat nog lang te doen. Adriaan heeft de buitengewone gave om achter de horizon te kijken; met veel passie deelt hij deze vergezichten met zijn collega's. Samen met Caren Wolferink, Maartje van der Veen en Robert Voorn hebben we ons vaak afgevraagd wat hij precies probeert te beschrijven. Meestal kwamen we er wel uit, maar soms ook niet, dan werd het niet concreter dan een *the A-team* busje. Zeker een leuk vergezicht! Ik hoop er nog een boel te zien. Caren, Maartje en Robert wil ik bedanken voor het zijn van leuke collega's. Dit klinkt wellicht als een afscheid, maar dat is het niet. Casper van Waveren en Jos Bloemers kan ik natuurlijk niet vergeten voor de broodnodige zin en onzin, thanx! Verder natuurlijk Brenda, Cor, Dymph, Ed, Els, Flip, Gert, Inge, Janske, Karen, Liesbeth, Lyna, Mathij, Mazin, Milly, Natascha, Pia en Saskia. Als laatste, maar zeker niet als minste, wil ik Rick Schermer bedanken voor de commerciële injecties en alle andere facetten van technologie, maar vooral voor de prettige samenwerking.

Daarnaast was ik in de luxueuze situatie om over een tweede bureau te beschikken: eentje bij de onderzoeksgroep *Diagnostic Decision Making* van de Radboud Universiteit Nijmegen. Er zijn ongetwijfeld maar weinig promovendi die zoveel uren overleg met hun begeleiders mochten genieten als ik. Met Cilia Witteman heb ik ontelbare NS-reisuren gemaakt. Naast onze artikelen, bespraken we horoscopen, de definities van betekenis, de betekenis van definities, de definities van humor, de humor van onzin, de onzin van definities, etc. Voor al deze creativiteit en het engelengeduld bij het zorgvuldig lezen en corrigeren, heel erg bedankt! Daarnaast wil ik mijn kamergenote, Leontien de Kwaadsteniet, samen met de andere collega's John van den Bercken, Edward van Aarle, Nicole Krol, en Marieke de Vries bedanken voor de prettige tijd die ik Nijmegen had. Omdat ik niet zoveel wist van diagnostiek, was het altijd een mentale uitdaging jullie discussies onder de wekelijkse besprekingen te volgen. Ik heb er iets van opgestoken. Bedankt mensen!

Over een luxueuze situatie gesproken—ik had een derde bureau: eentje bij de vakgroep *Intelligent Systems* van de Universiteit Utrecht. Samen met andere promovendi hadden we veel constructieve en minder constructieve discussies over agenten, ontologieën en de relativiteit van de waarheid. Maar ook hartstochtelijke discussies over volstrekt verwerpelijke zaken zoals Coca Cola en McDonald's. Birna van Riemsdijk, Bob van der Vecht, Cees Pierik, Davide Grossi, Geert Jonker, Huib Aldewereld, Jurriaan van Diggelen en Paul Harrenstein bedankt hiervoor. Wat wij gezamenlijk hadden is John-Jules Meyer. Hij is de hoogleraar met waarschijnlijk 36 begeleidingsuren per dag voor ons ter beschikking voor inzage in zijn indrukwekkende hoeveelheid kennis; daarvoor, en de vele lachbuien, erg bedankt! Andere collega's als Frank Dignum, Gerard Vreeswijk, Henry Prakken, Jan Broersen, Marco Wiering, Martin Caminada en Mehdi Dastani maakten een grote maar toch hechte groep met veel goede ideeën!

Als student van de SIKS-onderzoeksschool mocht ik erg veel collega-studenten ontmoeten. Weken zaten we opgesloten in conferentieoorden waar werd gedoceerd en gediscussieerd op het scherp van de snede. Heel veel ideeën heb ik hier gehoord en onthouden, en ben ik ook weer vergeten. 's Avonds melden Carsten Riggelsen, Danielle Sent, Femke de Jonge, Loes Braun, Ronnie Bathoorn, en vele anderen, inclusief docenten, zich af bij het biljard, bij de bar of blikjesautomaat, voor bord- en kaartspelletjes. Eventueel reden we rondjes door de gangen op elektrische invalidenkarretjes. De SIKS-coördinator, Richard Starmans, wil ik speciaal bedanken om dit allemaal in goede banen te leiden!

Sommige vrienden namen de gelegenheid om 'even' te achterhalen waar ik me de afgelopen vier jaar mee bezig heb gehouden. Speciale vermeldingen verdienen Lotte Bremer, Sanne Verhoef, Rogier Woltjer, maar zeker Bo Zorn. Voor het vinden van de vele typo's in de tekst en definities, thanx Bo!

In mijn vrije tijd genoot ik de mogelijkheid om niet over wetenschap en techniek na te mogen denken. In de weekenden bij Scouting Hoogeveen, en specifiek de OL-stam, heb ik veel prettige kampeerdagen mogen bijschrijven. Daarvoor wil ik graag Bart en Marjolijn Noordam, Epco Dijk, Freek van der Haar, Yanoula, Hans en Gerard Hommes, Janneke Metselaar, Joanne en Lammert Otten, Kim Oosterveen, Marcel Boxem, Margo Doornbos, Nelleke Verhage, Simone Teubel en vele anderen bedanken.

Bedankt! Speciaal wil ik Henri Kats, Alja Bouter en René Slagter, maar zeker niet uitsluitend, voor de hikes en andere waterbikkeltochtjes bedanken. Technologische vooruitgang krijgt alleen betekenis in contrast tot de afwezigheid ervan, bedacht ik me toen jullie de hele dag druk aan het varen waren en ik nog geen vinger had uitgestoken tijdens het lezen van mijn boek. Zeker onvergetelijk!

Daarnaast had ik doordeweeks bij mijn oude huisgenoten op Uilenstede 248 de nodige ontspanning en spanning in de vorm van BBQ's, feesten en zelfs wintersport. Bart Berghuijs, Carola van Eck, Joost Kruimer, Laura Kool, Marc Hamburger, Mark Couwenberg, Martijn Koorn, Martijn van Iterson, Nienke Debats, Sandra de Jong en Saphira Heijens wil ik speciaal bedanken om van een klein kamertje een geweldige thuis te maken. Daarnaast, hoewel geen huisgenoten, wil ik Rogier Woltjer en Jeroen Pijpe bedanken voor de steun en ontspanning in binnen- en buitenland! Soms loop je fout, maar toch loop je nooit teveel met Johan van der Beek, Johanna Plug, Ron Rotteveel, Roos Plaatzer, en natuurlijk Bo Zorn! Nu ben ik waarschijnlijk nog steeds mensen vergeten, jullie ook bedankt!

Speciaal wil ik mijn familie bedanken voor de jarenlange en onuitputtelijke steun op alle denkbare fronten, inclusief de ontelbare aanboden mijn plinten te regelen. Mijn broertje Jasper, mijn vader en moeder Henry en Emilie Lebbink – van Swieten, heel erg bedankt!

Appendix D

SIKS Dissertation Series

1998

Johan van den Akker, *DEGAS - An Active, Temporal Database of Autonomous Objects*, CWI, 1998-1

Floris Wiesman, *Information Retrieval by Graphically Browsing Meta-Information*, UM, 1998-2

Ans Steuten, *A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective*, TUD, 1998-3

Dennis Breuker, *Memory versus Search in Games*, UM, 1998-4

E.W. Oskamp, *Computerondersteuning bij Straftoemeting*, RUL, 1998-5

1999

Mark Sloof, *Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products*, VU, 1999-1

Rob Potharst, *Classification using decision trees and neural nets*, EUR, 1999-2

Don Beal, *The Nature of Minimax Search*, UM, 1999-3

Jacques Penders, *The practical Art of Moving Physical Objects*, UM, 1999-4

Aldo de Moor, *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems*, KUB, 1999-5

Niek J.E. Wijngaards, *Re-design of compositional systems*, VU, 1999-6

David Spelt, *Verification support for object database design*, UT, 1999-7

Jacques H.J. Lenting, *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.*, UM, 1999-8

2000

Frank Niessink, *Perspectives on Improving Software Maintenance*, VU, 2000-1

Koen Holtman, *Prototyping of CMS Storage Management*, TUE, 2000-2

Carolien M.T. Metselaar, *Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief*, UVA, 2000-3

Geert de Haan, *ETAG, A Formal Model of Competence Knowledge for User Interface Design*, VU, 2000-4

Ruud van der Pol, *Knowledge-based Query Formulation in Information Retrieval*, UM, 2000-5

Rogier van Eijk, *Programming Languages for Agent Communication*, UU, 2000-6

Niels Peek, *Decision-theoretic Planning of Clinical Patient Management*, UU, 2000-7

Veerle Coup, *Sensitivity Analyis of Decision-Theoretic Networks*, EUR, 2000-8

SIKS Dissertation Series

Florian Waas, *Principles of Probabilistic Query Optimization*, CWI, 2000-9

Niels Nes, *Image Database Management System Design Considerations, Algorithms and Architecture*, CWI, 2000-10

Jonas Karlsson, *Scalable Distributed Data Structures for Database Management*, CWI, 2000-11

2001

Silja Renooij, *Qualitative Approaches to Quantifying Probabilistic Networks*, UU, 2001-1

Koen Hindriks, *Agent Programming Languages: Programming with Mental Models*, UU, 2001-2

Maarten van Someren, *Learning as problem solving*, UvA, 2001-3

Evgueni Smirnov, *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets*, UM, 2001-4

Jacco van Ossenbruggen, *Processing Structured Hypermedia: A Matter of Style*, VU, 2001-5

Martijn van Welie, *Task-based User Interface Design*, VU, 2001-6

Bastiaan Schonhage, *Diva: Architectural Perspectives on Information Visualization*, VU, 2001-7

Pascal van Eck, *A Compositional Semantic Structure for Multi-Agent Systems Dynamics*, VU, 2001-8

Pieter Jan 't Hoen, *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes*, RUL, 2001-9

Maarten Sierhuis, *Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design*, UvA, 2001-10

Tom M. van Engers, *Knowledge Management: The Role of Mental Models in Business Systems Design*, VUA, 2001-11

2002

Nico Lassing, *Architecture-Level Modifiability Analysis*, VU, 2002-01

Roelof van Zwol, *Modelling and searching web-based document collections*, UT, 2002-02

Henk Ernst Blok, *Database Optimization Aspects for Information Retrieval*, UT, 2002-03

Juan Roberto Castelo Valdueza, *The Discrete Acyclic Digraph Markov Model in Data Mining*, UU, 2002-04

Radu Serban, *The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents*, VU, 2002-05

Laurens Mommers, *Applied legal epistemology; Building a knowledge-based ontology of the legal domain*, UL, 2002-06

Peter Boncz, *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications*, CWI, 2002-07

Jaap Gordijn, *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas*, VU, 2002-08

Willem-Jan van den Heuvel, *Integrating Modern Business Applications with Objectified Legacy Systems*, KUB, 2002-09

Brian Sheppard, *Towards Perfect Play of Scrabble*, UM, 2002-10

Wouter C.A. Wijngaards, *Agent Based Modelling of Dynamics: Biological and Organisational Applications*, VU, 2002-11

Albrecht Schmidt, *Processing XML in Database Systems*, UVA, 2002-12

Hongjing Wu, *A Reference Architecture for Adaptive Hypermedia Applications*, TUE, 2002-13

Wieke de Vries, *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems*, UU, 2002-14

Rik Eshuis, *Semantics and Verification of UML Activity Diagrams for Workflow Modelling*, UT, 2002-15

Pieter van Langen, *The Anatomy of Design: Foundations, Models and Applications*, VU, 2002-16

Stefan Manegold, *Understanding, Modeling, and Improving Main-Memory Database Performance*, UVA, 2002-17

2003

Heiner Stuckenschmidt, *Onotology-Based Information Sharing In Weakly Structured Environments*, VU, 2003-1

Jan Broersen, *Modal Action Logics for Reasoning About Reactive Systems*, VU, 2003-02

Martijn Schuemie, *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy*, TUD, 2003-03

Milan Petkovic, *Content-Based Video Retrieval Supported by Database Technology*, UT, 2003-04

Jos Lehmann, *Causation in Artificial Intelligence and Law - A modelling approach*, UVA, 2003-05

Boris van Schooten, *Development and specification of virtual environments*, UT, 2003-06

Machiel Jansen, *Formal Explorations of Knowledge Intensive Tasks*, UvA, 2003-07

Yongping Ran, *Repair Based Scheduling*, UM, 2003-08

Rens Kortmann, *The resolution of visually guided behaviour*, UM, 2003-09

Andreas Lincke, *Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture*, UvT, 2003-10

Simon Keizer, *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*, UT, 2003-11

Roeland Ordelman, *Dutch speech recognition in multimedia information retrieval*, UT, 2003-12

Jeroen Donkers, *Nosce Hostem - Searching with Opponent Models*, UM, 2003-13

Stijn Hoppenbrouwers, *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations*, KUN, 2003-14

Mathijs de Weerdt, *Plan Merging in Multi-Agent Systems*, TUD, 2003-15

Menzo Windhouwer, *Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses*, CWI, 2003-16

David Jansen, *Extensions of Statecharts with Probability, Time, and Stochastic Timing*, UT, 2003-17

Levente Kocsis, *Learning Search Decisions*, UM, 2003-18

2004

Virginia Dignum, *A Model for Organizational Interaction: Based on Agents, Founded in Logic*, UU, 2004-01

Lai Xu, *Monitoring Multi-party Contracts for E-business*, UvT, 2004-02

Perry Groot, *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving*, VU, 2004-03

Chris van Aart, *Organizational Principles for Multi-Agent Architectures*, UVA, 2004-04

Viara Popova, *Knowledge discovery and monotonicity*, EUR, 2004-05

Bart-Jan Hommes, *The Evaluation of Business Process Modeling Techniques*, TUD, 2004-06

Elise Boltjes, *Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes*, UM, 2004-07

Joop Verbeek, *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise*, UM, 2004-08

Martin Caminada, *For the Sake of the Argument; explorations into argument-based reasoning*, VU, 2004-09

Suzanne Kabel, *Knowledge-rich indexing of learning-objects*, UVA, 2004-10

Michel Klein, *Change Management for Distributed Ontologies*, VU, 2004-11

The Duy Bui, *Creating emotions and facial expressions for embodied agents*, UT, 2004-12

Wojciech Jamroga, *Using Multiple Models of Reality: On Agents who Know how to Play*, UT, 2004-13

Paul Harrenstein, *Logic in Conflict. Logical Explorations in Strategic Equilibrium*, UU, 2004-14

Arno Knobbe, *Multi-Relational Data Mining*, UU, 2004-15

Federico Divina, *Hybrid Genetic Relational Search for Inductive Learning*, VU, 2004-16

Mark Winands, *Informed Search in Complex Games*, UM, 2004-17

Vania Bessa Machado, *Supporting the Construction of Qualitative Knowledge Models*, UvA, 2004-18

Thijs Westerveld, *Using generative probabilistic models for multimedia retrieval*, UT, 2004-19

Madelon Evers, *Learning from Design: facilitating multidisciplinary design teams*, Nyenrode, 2004-20

2005

Floor Verdenius, *Methodological Aspects of Designing Induction-Based Applications*, UVA, 2005-01

Erik van der Werf, *AI techniques for the game of Go*, UM, 2005-02

Franc Grootjen, *A Pragmatic Approach to the Conceptualisation of Language*, RUN, 2005-03

Nirvana Meratnia, *Towards Database Support for Moving Object data*, UT, 2005-04

Gabriel Infante-Lopez, *Two-Level Probabilistic Grammars for Natural Language Parsing*, UVA, 2005-05

Pieter Spronck, *Adaptive Game AI*, UM, 2005-06

Flavius Frasincar, *Hypermedia Presentation Generation for Semantic Web Information Systems*, TUE, 2005-07

Richard Vdovjak, *A Model-driven Approach for Building Distributed Ontology-based Web Applications*, TUE, 2005-08

Jeen Broekstra, *Storage, Querying and Inferencing for Semantic Web Languages*, VU, 2005-09

Anders Bouwer, *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments*, UVA, 2005-10

Elth Ogston, *Agent Based Matchmaking and Clustering - A Decentralized Approach to Search*, VU, 2005-11

Csaba Boer, *Distributed Simulation in Industry*, EUR, 2005-12

Fred Hamburg, *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen*, UL, 2005-13

Borys Omelayenko, *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics*, VU, 2005-14

Tibor Bosse, *Analysis of the Dynamics of Cognitive Processes*, VU, 2005-15

Joris Graaumans, *Usability of XML Query Languages*, UU, 2005-16

Boris Shishkov, *Software Specification Based on Re-usable Business Components*, TUD, 2005-17

Danielle Sent, *Test-selection strategies for probabilistic networks*, UU, 2005-18

Michel van Dartel, *Situated Representation*, UM, 2005-19

Cristina Coteanu, *Cyber Consumer Law, State of the Art and Perspectives*, UL, 2005-20

Wijnand Derks, *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics*, UT, 2005-21

2006

Samuil Angelov, *Foundations of B2B Electronic Contracting*, TUE, 2006-01

Cristina Chisalita, *Contextual issues in the design and use of information technology in organizations*, VU, 2006-02

Noor Christoph, *The role of metacognitive skills in learning to solve problems*, UVA, 2006-03

Marta Sabou, *Building Web Service Ontologies*, VU, 2006-04

Cees Pierik, *Validation Techniques for Object-Oriented Proof Outlines*, UU, 2006-05

Ziv Baida, *Software-aided Service Bundling – Intelligent Methods & Tools for Graphical Service Modeling*, VU, 2006-06

Marko Smiljanic, *XML schema matching – balancing efficiency and effectiveness by means of clustering*, UT, 2006-07

Eelco Herder, *Forward, Back and Home Again – Analyzing User Behavior on the Web*, UT, 2006-08

Mohamed Wahdan, *Automatic Formulation of the Auditor's Opinion*, UM, 2006-09

Ronny Siebes, *Semantic Routing in Peer-to-Peer Systems*, VU, 2006-10

Joeri van Ruth, *Flattening Queries over Nested Data Types*, UT, 2006-11

Bert Bongers, *Interactivation – Towards an e-cology of people, our technological environment, and the arts*, VU, 2006-12

Henk-Jan Lebbink, *Dialogue and Decision Games for Information Exchanging Agents*, UU, 2006-13