In the Eye of the Beholder

Explaining Behavior through Mental State Attribution

This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie).

SIKS Dissertation Series No. 2011-42
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 Wöhrmann Print Service B.V., Zutphen

ISBN 978-90-8570-509-3

In the Eye of the Beholder

Explaining Behavior through Mental State Attribution

In 's aanschouwers oog

Verklaren van gedrag door toeschrijving van een mentale toestand

(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. G. J. van der Zwaan, ingevolge
het besluit van het college voor promoties in het openbaar te verdedigen
op maandag 7 november 2011 des middags te 12.45 uur

DOOR

Michal Peter Sindlar

geboren op 5 oktober 1979 te 's-Hertogenbosch

Promotor:
Prof. dr. J.-J. Ch. Meyer

Co-promotor:
Dr. M. M. Dastani

Contents

Acknowledgements

In creating the work that you are reading I have benefited from the company and advice of many people, and I would like to take this opportunity to express my thanks to them.

First and foremost, I wish to thank my dissertation advisors John-Jules Meyer and Mehdi Dastani. John-Jules, I have greatly enjoyed being your Ph.D. student for the past few years. Your enthousiasm was legendary among students already when I attended your lectures as part of my CKI curriculum, and I consider myself fortunate to have experienced it even more directly during the course of my academic research. Looking back, I believe that meetings with you were always so pleasurable because your sharp wit and genuine interest in my doings ensured that I always left your office wiser than before, whereas your kindhearted and humorous approach ensured a comfortable, even homely, atmosphere. That personal aspect is generally rare in professional life, but all too important; it kept me motivated throughout the whole experience and is something I will definitely miss!

Naturally, Mehdi, I also thank you for the enjoyable cooperation, it has been a true pleasure getting to know you. If I have forgotten the numerous Iranian proverbs you presented us with during meetings, then this is only because my mind was too busy trying to grasp the technical problems which you were bringing to our attention by means of them. This illustrates the aspect of your supervision which I am most grateful for: your unrelentless persistence when it comes to getting the details right. That quality has kept my research on track and eventually managed to bring out the researcher in me; I believe this dissertation would not have been half as good without your guidance. Thank you!

A special thanks goes to Frank Dignum, whom I thank for valuable lessons learned, and for helping shape my ideas in the initial stages of the project. Also, Frank, your efforts in getting me involved with student supervision presented a perfectly timed broadening of horizons, and a pleasant and successful cooperation for all those involved.

Of all the people who have been so kind as to spend some of their time on helping me improve this work, four wise men stand out: my dissertation committee. I thank all of you — prof. dr. Jürgen Dix, prof. dr. Peter Flach, prof. dr. Leon van der Torre, and prof. dr. Cees Witteveen — for your efforts and comments, and look forward to our discussions. At this point I would also like to thank the anonymous reviewers who, over the course of the past few years, have commented (in all sorts of styles and tones) on the work submitted by myself and co-authors, and thus helped improve it.

Closer to home, but still in the scientific domain, I thank my colleagues of the Intelligent Systems Group at the University of Utrecht. Bas, Chris (see you on the courts!), and Nieske, thanks for accepting me straightaway as your new office roommate. It was nice to listen to your talk about pots and pans of different colors, emotions, and dialogue scores, occasionally interrupted by a Flemish speaking iCat; all research-related! Also, it was great to discover so many new beers (although I've somehow forgotten which…) during our infrequent A-119 outings. Thanks goes in equal measure to my subsequent office roommates

Joost and Tom, with whom I shared the room for the latter part of our group's stay in CGN, for giving your unsalted opinions on many topics, for plenty games of ping-pong, and for the good times we had in general. Tom, it's great to still be colleagues! It seems only fitting here to thank you, Sim, as well at this point, for heavily-peppered opinions on everything, and for sharing the ups and downs of our ultimate 'heist' on a not-to-be-named electronics store. Completing this series of roommates, I thank Marieke for making the brief time that we shared the room in BBL together so agreeable!

Apart from office roommates, there have been so many of you at the Intelligent Systems group that I can only extend elaborate thank-yous to a strict subset, lest should the present section compete in length with that of the core of this work. To Nick, for showing me that a professional attitude can go together with being a jolly guy, and for the frequent routine of 'in de bomen hangen'. To Maaike, for a pleasant cooperation and the occasional sharing of thoughts on our related subjects. To Huib for the chats and whisky, as well as getting me involved in organizing ESAW — Virgina, thanks for that as well. To Henry, for your willingness to comment on my first paper, making me realize that even if I hadn't done too bad there was plenty left to learn. To Loïs, for your friendliness and persistence in many respects, certainly on a squash court. To Marco, for inspiring me to become a scientist, and for being so one-of-a-kind. And to other IS(-related) people — André, Bob, Christian, Daniel, Davide, Eric, Fernando, Geert, Gennaro, Gerard, Hado, Jan, Joost, Jurriaan, Lara, Liz, Loris, Marcel, Martin, Max, Paolo, Richard, Susan, and Zheng — thanks as well!

Naast de mensen die op verschillende manieren in professionele zin bij dit werk betrokken zijn geweest, ben ik ook bepaalde mensen erg dankbaar met wie ik in meer persoonlijke sferen heb verkeerd. Allereerst de 'harde kern' van Squash Utrecht, met wie ik zoveel leuke momenten heb gedeeld en die mij de afgelopen jaren lichamelijk, maar zeker ook geestelijk, gezond hebben gehouden. Jan, jij nodigde mij een hele tijd geleden uit bij het 'donderdagavond-groepje' en daarmee ging voor mij het balletje serieus rollen, zo nu en dan ook in de nick! Bedankt voor je jeugdige enthousiasme en het delen van lief en leed. Jorg, ook jij was er die avonden vaak te vinden, en sindsdien zijn we elkaar steeds vaker gaan zien. Ik blijf me altijd verbazen over je aanpassingsvermogen en hoe je altijd in staat bent snel tot de kern van de zaak door te dringen ("gameline!"); ik weet niet hoe ik het volgehouden had zonder regelmatig samen even stoom af te blazen. Hans, fanatiekeling, blij je te hebben leren kennen! Bas, wij kennen elkaar al wat langer maar sinds deze zomer pas écht. Het was erg gezellig, blijf die wilde pruimen zoeken! Jay, ook jou wil ik bedanken voor je lessen op en buiten de baan, alsmede je creatieve vondsten — die ik hier niet zal herhalen — om het één-twee ritme erin te krijgen. Jullie en alle andere leden van mijn 'squashfamilie' — (oud-)teamgenoten, toernooigangers, ladderaars — hoop ik vaak te blijven zien!

Mijn gesport is de laatste tijd vooral beperkt gebleven tot squashen, maar in den beginne was er ook nog zaalvoetbal. Locos, jullie zijn echter meer dan voormalig teamgenoten; betere vrienden kan een mens niet wensen! Paul, altijd klaar met raad, maar bovenal daad. Jos, met jou heb ik één van mijn diepste indrukken van de laatste tijd beleefd — onze nachtelijke omzwerving door hartje IJsland, met als doel bemachtiging van kostbaar gletsjerwater. Deze zijn voor jou inmiddels opgevolgd door nog diepere indrukken met de geboorte van jullie menneke: keimooi! Arthur, ik weet nog dat we naast elkaar zaten bij Biologie op het JRL en het toen altijd al leuk hadden; wie had echter gedacht dat we enkele jaren later goede

vrienden zouden zijn en zulke avonturen zouden beleven, onder meer vele duizenden kilometers afleggend door Noord-Amerika? Ik hoop dat die avonturen blijven komen! Florian, ook jij staat in dit rijtje en ik zou inmiddels bijna gaan denken dat jij ook een 'brabo' bent, maar in hart en nieren blijf je natuurlijk altijd een (Duits-Tsjechische) Hagenees. Bedankt voor de wandeltochten, de muziek, de films en dat ik lekker door mocht slapen, de vakanties, de gesprekken over luchtige en zware zaken; blijf wie je bent, en bedankt dat je mij ook zo accepteert. Ik ben blij en trots jullie mijn vrienden te mogen noemen! Voorts wil ik nog mijn dank uitspreken aan de fijne mensen die ik hier niet bij naam heb genoemd, in Utrecht en daarbuiten, in het bijzonder de pub quizzers en de 'Nijmeegse' vriendinnen en vrienden. Het is fijn dat jullie er allemaal zijn!

Een mens kan niet bestaan zonder familie, besef ik mij maar al te goed, en in dat opzicht voel ik mij bijzonder omdat ik er daar niet één maar twee van heb. Allereerst mijn Nederlandse familie: Anton, Louise, Menno, het voelt alsof ik jullie al mijn hele leven ken, en dat is fijn! Als ik er eens goed over nadenk dan kennen we elkaar in zekere zin zelfs daadwerkelijk al bijna zo lang; de foto van Anton die mij in jaren-'80 zwembroek lesgeeft heb ik in elk geval zo weer op mijn netvlies! Anton, je hebt me het zwemmen prima geleerd, als echte 'landrot' zal ik waarschijnlijk geen makkelijke leerling geweest zijn. En mocht het ooit nodig zijn dan kan Christel me altijd nog wat bijles geven! Louise, jij ook bedankt voor het vele meeleven de laatste jaren, daarnaast ook dank voor de waardevolle momenten van samenzijn bij jou thuis! Menno, natuurlijk ben ik jou niet vergeten. Je had zo in één van de bovenstaande rijtjes kunnen staan: we hebben samen gesquasht in Boxtel en Nijmegen; gezaalvoetbald in Utrecht; de grote reis gemaakt, squirrels achterna in de Rockies; mooie feestjes mogen vieren op de Graafseweg en de Van Lieflandlaan; aan 'extreme hiking' gedaan in Slowakije en Schotland; en meer! We hebben wat meegemaakt de afgelopen jaren. Naast een fijne vriend ben je ook fijne familie en dat is heel wat waard: bedankt!

Milá rodino česká, vy jste mé kořeny: většinou v pozadí, ale současně nepostradatelné. A i když žijeme od sebe daleko, v mém srdci a myšlenkách jste stále na blízku.

Pap en mam, eigenlijk zijn jullie meer dan wie ook mijn Tsjechische familie, maar ik ga jullie toch bedanken in het Nederlands. De afgelopen jaren zijn voor mij niet altijd even makkelijk geweest, maar ik weet dat het juist daarom voor jullie minstens zo zwaar was. Gelukkig waren er ook veel mooie momenten, hopelijk zullen er daar nog vele van volgen! Bedankt voor onaflatende steun, geduld, vertrouwen, en liefde; dit heeft mij gemaakt tot wie ik ben, en in zekere zin is dit proefschrift daarom ook een beetje van jullie: gefeliciteerd!

Lief mopsje, lieve Christel, als ik één persoon zou mogen bedanken in het kader van dit dankwoord dan zou jij dat zijn. Het is, welbeschouwd, zeer bijzonder dat we in dezelfde periode aan onze respectievelijke proefschriften hebben gewerkt en we hebben dat er naar mijn mening niet slecht vanaf gebracht… ik ben erg trots op je en bewonder je inzet en mentaliteit! Als maar een fractie daarvan is afgewreven op mij dan prijs ik me gelukkig. Natuurlijk zijn we meer dan samen "boktorren in spé" — een mooie analogie gezien de hoeveelheid verpulpt hout die we 'gevreten' hebben — en ik ben elke dag weer blij met jou als lieve vriendin; dat alleen al vind ik een goede reden tot groot feest!

Utrecht, oktober 2011

Chapter I

Introduction

> "This is merely to say that what is selected as the cause is often the event or condition on which the effect depends (without which the effect would not have occurred) which is, *for one reason or another*, taken to be of primary interest to those doing the describing or explaining. Causal conditions may be out there in the world, but something's status as *the* cause is, it seems, in the eye of the beholder."
>
> Fred Dretske, *Explaining Behavior: Reasons in a World of Causes* (1988, p. 24)

In five words, this dissertation concerns the *explanative attribution of mental states*. That is to say (more verbosely), it is about making the notion concrete that the behavior of others is sometimes explainable by assuming them to have particular goals or beliefs. This 'concreteness' is here realized in a formal, mathematical, way; as such, the work presented in this dissertation lends itself for realizing mental state attribution as an instance of artificial intelligence. As the epigraph to this chapter suggests, this kind of attribution is typically of a subjective nature, in the sense that the one who does the explaining — the 'beholder' — gets to decide which mental state to attribute to the one doing the behaving. In line with this fact, the emphasis of formalization lies on the point of view of the beholder, which throughout this dissertation is referred to as the 'observer' or 'mindreader'; depending on the context.

The present chapter aims to set the stage for the remainder of this dissertation by establishing somewhat of a shared state of mind between author and audience, through the mentioning of ideas and works that have influenced our own.[1] Doing so will lead us to formulate the research questions that this work seeks to answer. As a road map for navigating this dissertation, it is shown in which chapters those questions are answered.

1 Intentionality

The concept of *intentionality* is important for this work, and as such it warrants thorough description. This is all the more so because of the fact that the term 'intentionality' has both a technical and a more colloquial meaning, both of which are relevant from our point of view (Stanford Encyclopedia of Philosophy, 2010; Dennett, 1987).

[1] Going by the title page of this dissertation, it would appear that it has but a single author. Yet, it could not have existed without the help of many people, and to express this fact this work is written in a first person plural form known as the 'pluralis modestiae'. At times, mostly as a part of mathematical discourse, the same form is used in the sense of 'pluralis auctoris'.

1.1 Intentionality of Concepts

The first meaning of 'intentionality' is specialistic, and is mostly encountered in philosophical texts. Quoting Dennett, "intentionality, in philosophical jargon, is — in a word — *aboutness*" (1987, p. 240). This meaning of the term 'intentionality' refers to the fact that certain concepts, such as beliefs and desires, have the property of being *about* something. A more technical way of putting this, is to say that those concepts express 'propositional attitudes'. For illustration, consider the claim that Sally believes that her marble is located in a basket, which is true if (and only if) Sally has a propositional attitude of the belief-type towards the proposition 'the marble is in the basket'. Regardless of how one should go about verifying whether Sally actually has this particular attitude towards that particular proposition, it is noteworthy that the truth of the claim is independent of the truth of the proposition as such: it may occur that the marble is not in the basket, yet Sally (for some reason or other) believes it to be.

Concerning the nature of intentional concepts, Dennett sketches two opposing views that could be called 'realism' and 'interpretationism'. The former, realism, "likens the question of whether a person has a particular belief to the question of whether a person is infected with a particular virus — a perfectly objective internal matter of fact about which an observer can often make educated guesses of great reliability" (1987, p. 14). On the other hand, interpretationism "likens the question of whether a person has a particular belief to the question of whether a person is immoral, or has style, or talent, or would make a good wife" (1987, p. 15). Put otherwise, realism assumes that intentional concepts have a physical instantiation which, in principle, could be measured, to give an objective answer to the question of whether someone has the corresponding propositional attitude. Interpretationists take the view that whether or not someone has a certain belief or desire is only a matter of interpretation. Dennett states that realism and interpretationism are typically regarded as mutually exclusive positions, which he believes to be mistaken, because, although they may be perfectly objective phenomena (a realist claim), mental states can be discerned only by adopting a predictive strategy and assessing its success (an interpretationist claim). Dennett calls this strategy the *intentional stance*, and in his account develops the idea that, apart from oneself wielding concepts that display intentionality, one can also attribute to others the capability to do so. This attribution (in the spirit of interpretationism) is in the eye of the beholder, even if the grounds on which it occurs and can be verified need not be.

1.2 Intentionality of Behavior

The second meaning of 'intentionality' is more colloquial than the first, which was described in the previous section, and refers to the goal-directedness of behavior. That meaning rears its head in, for example, the phrase "Anne has the intention to take the marble from the basket". Given the non-technical interpretation, this sentence refers to the fact that Anne is deliberately 'planning' to do the action of taking the marble from the basket. Interestingly, there is also a hint of intentionality of the kind described in Section 1.1 in this reading, as Anne can be said to have a certain propositional attitude (namely of the intention-type) towards the aforementioned action. The nature of intention (i.e. 'goal-

directedness') as an intentional (i.e. 'aboutness-displaying') concept has been the subject of philosophical debate, and 'intention', given the colloquial meaning discussed in the current section, in this sense also has a technical connotation. This is the subject of the remainder of this section.

A well-known philosophical model of intention and its role in guiding agents' behavior is the Belief-Desire-Intention (BDI) model of practical reasoning developed by Bratman (1987). In contrast to decision-theoretic approaches (Jeffrey, 1983), Bratman's model has the central assumption that agents are resource-bounded: "Deliberation is a process that takes time and uses other resources; this means that there are obvious limits to the extent of deliberation at the time of action. By settling on future-directed intentions, we allow present deliberation to shape later conduct, thereby extending the influence of deliberation and Reason on our lives" (1990, p. 18, summarizing (Bratman, 1987)). In Bratman's account, intentions on which an agent has settled impose certain constraints on its future behavior. Specifically: agents should take action in order to 'satisfy' their current intentions; they should not adopt new intentions that conflict with their current intentions; and agents should furthermore track the success of attempts to satisfy their intentions, reconsidering them if necessary. The BDI model of practical reasoning has been of interest to the artificial intelligence (A.I.) community, given its potential application in developing autonomous, resource-bounded, software.

1.3 Intentional Software Agents

Attempts to formalize 'intention' based on the BDI model of practical reasoning have resulted, amongst others, in the seminal papers by Cohen & Levesque (1990) and Rao & Georgeff (1991). Both those accounts employ modal logic, but provide different formalizations of intention: Cohen & Levesque (1990) model it as a kind of 'persistent goal' in terms of more basic constructs, whereas Rao & Georgeff (1991), who describe intentions as 'partial plans', adopt them as primitives in their formalism. Regardless of specifics, both those accounts have been influential where it comes to the design of autonomous software agents, and it is interesting to see that in this domain also the notion of the intentional stance comes into play. Wooldridge & Jennings define an *agent* as a software-based system that shows autonomy, social ability, reactivity, and proactiveness; furthermore, they state that "being an intentional system seems to be a *necessary* condition for agenthood" and "an agent is a system that is most conveniently described by the intentional stance" (1995, p. 120). In that context the intentional stance is adopted as an abstraction, that has the user or designer of a system treat it *as if* it were capable of displaying intentionality, by attributing to it mental qualities and reasoning about its behavior accordingly (as advocated also by McCarthy (1990)). This abstraction is taken yet a step further by BDI-based agent programming languages (Bordini et al., 2009), where the agent is not only treated — in the course of design or operation — as an intentional system, but is actually equipped with computational counterparts of intentional concepts. These programming languages thus typically provide constructs that represent (derivatives of) BDI primitives, so that a software agent can be directly programmed in terms of, e.g., a plan it should adopt as its intention if it has a particular belief and goal.

I.2

2 Research Questions

At the beginning of this chapter it was made clear that this dissertation is about formalizing the explanative attribution of mental states. Our motivation for doing so stems from a goal to enable the creation of 'better' virtual characters for applications like (serious) computer games and virtual training environments, whereby it is assumed that BDI-based agent programming is in principle a suitable paradigm for implementing virtual characters. This assumption is supported by the literature, which indicates that the abstraction level of BDI concepts makes this paradigm suitable for modeling 'human-like' behavior in a natural way (Norling & Sonenberg, 2004). Moreover, the fact that the BDI model has the assumption of agents' resource-boundedness at its core makes it interesting for games, where such bounds are given by the need for balancing resources used for A.I. with those required for other tasks. In general, the gaming domain is considered to be a potential catalyst for the development of human-like A.I. (Laird & van Lent, 2001; Nareyek, 2007), and recent activity on the development of toolkits for interfacing with state-of-the-art games (Kadlec et al., 2009) has opened up this domain as a testing and proving ground for agent programming platforms (Hindriks et al., 2011). The absorption into game A.I. (Funge, 2004; Millington, 2006) of classical A.I. planning techniques (Orkin, 2006) furthermore indicates the willingness of commercial game developers to expand their horizons.

A genre of games that fit the methaphors of multi-agent systems rather nicely is that of *role-playing games* (RPGs), where players play out roles and interact with richly detailed non-player characters (NPCs), for the ultimate fulfillment of short-term and long-term quest goals. Exemplified by classic series such as ULTIMA, MIGHT AND MAGIC, and EYE OF THE BEHOLDER (sic!), of which the simple graphics and static game flow was reminiscent of the original pen-and-paper variants, state-of-the-art RPGs like recent installments of THE ELDER SCROLLS require the latest hardware to provide environments with a high level of multimedia realism. However, in spite of their overwhelming audiovisual capabilities, research indicates that the social believability of NPCs in those kinds of games is found disappointing by players. Features in which those characters are found lacking include personality, proactiveness, relationships with the player, and relationships with other NPCs (Afonso & Prada, 2008). The importance of those features is supported by research that finds them to be determinant for the believability of virtual characters (Loyall, 1997). Employing BDI-based agent programming can facilitate the development of virtual characters that appear to be proactive and to have a personality (having, e.g., distinct preferences and persisting beliefs). For realizing features relating to social behavior, like forming relationships with other NPCs and players, BDI-based agent programming can also be a potential solution, seeing that decision-making based on beliefs and desires attributed to others is taken to be crucial for this kind of activity (Baron-Cohen, 1995; Dennett, 1987). Apart from the motivation from the game domain, applications in human-computer interaction can furthermore benefit from formal approaches to forming and maintaining a model of the mental states of others. This leads to our overall research question.

Overall Research Question.
How can explanative attribution of mental states by BDI-based agents be realized?

This overall research question is split up into more specific 'sub-questions', which are introduced in the remainder of this section. Of each question it is then, briefly, mentioned how we have proceeded to answer it, and which results have been obtained. Given that our interest is in conclusions that can be drawn a priori — that is, before having observed any of the agent's behavior — our approach employs formal logic instead of probabilistic methods. This choice is furthermore justified by the fact that it allows for a relatively straightforward translation to logic programming, which is the predominant paradigm for instantiating the reasoning mechanisms of BDI-based agents.

If NPCs in a virtual environment are implemented as BDI-based agents, then it may be possible to give them awareness of each other's mental states by allowing them to directly access the beliefs and goals of others. However, it is recognized that such 'cheating' techniques are error-prone because they require a fine balancing act in order to be successful (Lidén, 2002), given that it in many cases it may not be believable for agents to be aware of the mental internals of others. A more 'sensory honest' (Isla & Blumberg, 2002) approach is to have agents *explain* the (observable) behavior of others in terms of their (unobservable) mental states. This approach has a long tradition in A.I., under the umbrella terms of *intention recognition* and *plan recognition* (Sadri, 2011; Carberry, 2001). There is little work in this area that is specific to BDI-based agent programming, though, which leads us to formulate our first (specific) research question.

Research Question 1.
How can explanation of the behavior of BDI-based software agents be formalized?

Explanation is generally considered in logic to be a problem of *abduction* (Aliseda-Llera, 1997), which is sometimes equated with 'inference to the best explanation' (Lipton, 2004; Campos, 2011). Abduction is a logical reasoning form that allows inference of a (defeasible) explanatory conclusion, given a background theory and an observation, and is generally considered complementary to *deduction* and *induction*. The philosopher Charles Sanders Peirce characterized it with this syllogism (Hartshorne et al., 1958):

> *The surprising fact, C, is observed.*
> *But if A were true, C would be a matter of course.*
> *Hence, there is reason to suspect that A is true.*

In answering Research Question 1, the explanation of a BDI-based agent's behavior is modeled as a case of abduction using a ground fragment of predicate logic. The assumption is hereby made, as usual in plan/intention recognition (Sadri, 2011), that the rules of agents are known to the beholder (referred to in this context as the 'observer') that does the explaining. Care is taken that the approach is robust (Carberry, 2001), in the sense that it can handle missing observations. This abductive account is then reformulated in terms of explanatory functions, of which it is shown that they succinctly express the inference of abductive explanations that can be given skeptical and credulous interpretations, and furthermore have useful computational properties.

The abductive approach to explaining observed behavior of BDI-based agents takes into account the defeasible nature of explanations, as well as the fact that actions can be failed

to be observed. However, the formalism employed is not very suitable for expressing the *dynamics* involved in attribution of beliefs and goals on grounds of observed actions. Specifically, it should be taken into account that the mentalistic explanations inferred by means of the abductive explanatory functions, as described in the previous paragraph, represent beliefs and goals that an agent (possibly) had *before* it performed the actions observed by the beholder, a consideration which is also important if one is to obtain a notion of how this agent's (presumed) beliefs and goals evolved as a result of the plan it executed. This leads us to the second research question.

Research Question 2.
How can the dynamics involved in explanative attribution of mental states be modeled?

For answering Research Question 2 we employ propositional dynamic logic (PDL) (Harel et al., 2000) as a modeling tool. In order to render this formalism suitable for our purpose it is equipped with special atomic propositions that pertain to the observation of actions, enabling us to formalize attribution of mental states to BDI-based agents under conditions of both complete and incomplete observation. To express this attribution we employ operators for belief and goal ascription, which are likewise interpreted in terms of specially 'marked' (labeled) atomic propositions. This formal machinery is used to determine classes of PDL models that represent interpretations of the abductive approach which also capture dynamics, taking into account the fact that observation may be incomplete, and that ascription of mental states should be in accordance with the presumption that the agent is executing a particular plan.

An underlying assumption made in answering Research Questions 1 and 2 is that the rules of the observed BDI-based agents are known to the beholder. This assumption is reasonable in contexts where the internals of agents are accessible, yet where it is not desirable for the sake of believability to give knowledge of all those internals to other agents (beholders); explanation of behavior based on the knowledge of their rules then has the benefit of yielding plausible, yet possibly erroneous, explanations. However, in some situations the assumption of agents' rules being known to the beholder cannot be maintained, such as when the internals of others are not accessible by design or by nature. Moreover, even if rules of agents are available, attribution of mental states on grounds other than those rules can be used in complement to rule-based attribution. In psychology, reasoning to a conclusion of attributed mental states is referred to as *mindreading* (Nichols & Stich, 2003; Baron-Cohen, 1995), and our third research question is therefore as follows.

Research Question 3.
How can mindreading be formalized?

To answer this research question, two psychological models of mindreading (also referred to as 'theory of mind', 'mentalizing', and 'folk psychology') are discussed, focusing on the importance of (shared) perception in mindreading, and the embedding of this activity in a general cognitive architecture. PDL is also employed in answering Research Question 3, to preserve compatibility with our approach to the previous research question, and because it is a well-known tool for reasoning about actions with translations to other domains (Zhang

& Foo, 2005). The formalism is extended to handle scenarios where multiple agents are involved, and used to express logical schemata called 'mindreading patterns'. Those schemata provide a formal grasp on mindreading by stating the form of regularities presumed by the beholder (referred to as 'mindreader' in this context) to hold between certain facts and attributed mental states. The use of these schemata is illustrated in modeling the beholder in a *false-belief task*, which is generally considered to be an important task for testing mindreading abilities (Dennett, 1978; Wimmer & Perner, 1983; Baron-Cohen et al., 1985; Bloom & German, 2000).

Seeing that our main research question concerns the *realization* of explanative attribution of mental states by BDI-based agents, it is desirable to show how our approach can be implemented. In doing so, we focus on the answer to Research Question 1, i.e. the explanation of other BDI-based agents' behavior given knowledge of their rules, which has been termed 'mental state abduction'. Given that most current agent programming platforms employ some form of logic programming, it is useful to conceive the implementation as a logic program; this has the added benefit that it allows for further reasoning with inferred explanations. Also, the implementation should account for the nonmonotonicity inherent to mental state abduction; last but not least, considering our intended application in the (serious) gaming domain, it should employ a programming paradigm that is actively developed and has good performance results. In light of those desiderata, our fourth research question becomes as follows.

Research Question 4.
How can mental state abduction be implemented?

Mental state abduction involves nonmonotonicity, because it is known that the abduced explanations are defeasible and can turn out to be wrong. The state-of-the-art approach to nonmonotonic logic programming is *answer set programming* (ASP), which is actively developed, and has been used in industrial settings. Tools from the Potsdam Answer Set Solving Collection (Gebser et al., 2007) have been top performers in recent ASP competitions so that it is deemed legitimate to consider them as foundation for our approach to implementation. Specifically, a combination of GRINGO grounder and CLASP solver is selected as the programming framework, for which a specification for implementing mental state abduction is presented that takes into account the nonmonotonic properties of this activity, in answer of Research Question 4. It deserves mention here that our approach does *not* concern the implementation of a specific system; it concerns rules to obtain the implementation of such a system from the abductive framework. Hereby the typical 'generate-and-test' ASP programming methodology (Lifschitz, 2008) is followed, and the resulting implementation is formally shown to be correct in regard to the abductive framework.

As stated earlier, plan/intention recognition has been an active research area in A.I., and similarly has there been quite some work that pertains to mindreading, in some way or other. Because of this fact, we consider it worthwhile to clearly position our approach in context of related work, to point out, amongst others, in which way our framework could complement approaches from related work, or vice versa. This endeavor is formulated below as the final research question, which, given the motivation as just sketched, falls within the scope of the overall question.

Research Question 5.
How does our approach stand in comparison to related work?

Focusing mostly on accounts that, like ours, are logic-based, we have proceeded to answer Research Question 5 by comparing our work to key related approaches, taking cue from recent literature surveys in this field (Carberry, 2001; Sadri, 2011). In doing so, we have attempted to identify their pros and cons, and to give our view on how these works could complement each other.

3 The Structure of this Dissertation

Having stated in Section 2 the research questions this dissertation intends to address, the current section sketches, per chapter, how it is structured to do so. It is hereby also pointed out which of our earlier work those chapters rely upon.

- Research Question 1 is answered in Chapter II. This chapter relies mostly on our work on the functional approach to mental state abduction and its nonmonotonic formalization (Sindlar et al., 2008, 2011), also incorporating results on certain formal properties that were presented separately (Sindlar et al., 2009a).
- Research Question 2 is answered in Chapter III, which is a rather more extensive reformulation of our earlier work on the same topic (Sindlar et al., 2010b).
- Research Question 3 is answered in Chapter IV. The logical part of this chapter was presented in rudimentary form at a workshop (Sindlar et al., 2010a), whereas the examples of modeling and implementation of the false-belief task are inspired by an earlier attempt in that direction (Sindlar et al., 2009b).
- Research Question 4 is answered in Chapter V, encompassing a generalization to cases of incomplete observation, as well as a minor reformulation, of earlier work on the use of answer set programming (Sindlar et al., 2011).
- Research Question 5 is answered in Chapter VI. Its contents have not as such appeared elsewhere, although fragments of it can be found throughout related work discussions that have appeared as part of our various contributions to this domain (Sindlar et al., 2008, 2009a,b, 2010a,b; Sindlar & Meyer, 2010; Sindlar et al., 2011).

As this overview shows, the structure of this dissertation is quite straightforward, tackling one research question per chapter. The dissertation is concluded with Chapter VII, in which our efforts are reflected upon.

Chapter II

Mental State Abduction

This chapter formalizes explanatory abductive reasoning about an agent's mental state on grounds of its observed actions. First of all, in Section 1 logical abductive reasoning is discussed, and in Section 2 a basic agent programming language called MyAPL is introduced. Section 3 then proceeds to present a general discussion of reasoning about observed agents' behavior under different perceptory conditions, paving the way for a formal treatment that involves the notion of abduction. This logical treatment has the benefit of being formulated in terms of well-known concepts for defeasible reasoning, at the expense of being somewhat unwieldy and verbose. It is therefore reformulated in terms of functions that operate on a meta-level in regard to the logical entities of interest, an approach which has its benefits although it does demand some concessions, as is the topic of Section 4. In conclusion of this chapter, Section 5 reflects on the material presented there.

1 Logical Abduction

The term 'abduction' was introduced into classical logic in the late nineteenth century by the American philosopher Charles Sanders Peirce (Hartshorne et al., 1958), who divided logical inference into three categories: *deduction*, *induction*, and *abduction*.[1] Given two facts and a rule of implication in which one fact is the antecedent and the other the consequent, those notions can be characterized as follows: deduction is inference of the consequent from the antecedent and the rule, induction infers the rule from the antecedent and the consequent, and in abduction the antecedent is the conclusion which is inferred given the rule and the consequent as premises. Deduction is a truth-preserving form of inference, in the sense that if the premises are true, so must be the conclusion. Induction and abduction, on the other hand, are conjectural, meaning that truth of the premises does not guarantee truth of the conclusion. Unfortunately, scientific terminology on this topic is not so unambiguous as one would like it to be, given that the term 'induction' is sometimes used in reference to both induction and abduction, as described above. This latter use of the term 'induction' is regularly encountered in the literature (Flach, 1995; Bessant, 1996; Lipton, 2004). Further confusion is stirred by the fact that some (Aliseda-Llera, 1997) take the term 'abduction' to principally denote the inference of any fact along the format given at the start of this paragraph, whereas others (Harman, 1965) equate it with inferring the 'preferred' conclusion.

Abduction also has an explanatory connotation, which shows through if the inference takes place with respect to some body of information that does not account for an observed fact as such, but would do so if some other fact were incorporated. Under this

[1] For the sake of simplicity, the view on Peirce presented in the running text is somewhat eclectic; see Aliseda-Llera (1997) for a more detailed exposition.

interpretation the observed fact is called the *explanandum* (from Latin: 'that which is to be explained'), the to-be-incorporated fact the *explanans* ('that which does the explaining'), and the body of information is called the *theory* .[2] Both Harman (1965) and Lipton (2004) use the term 'inference to the best explanation' (IBE) in relation to abduction in its explanatory form. Recently, though, it has been argued by Campos (2011) that abduction should not be conflated with IBE and that the two should be evaluated separately, on their own merit. That topic is beyond the scope of this introduction to abduction, though, so that we distance ourselves it and adhere to the view on explanatory abduction presented in Section 1.1.

II.1

1.1 Explanatory Abduction

In order to do justice to the explanatory nature of abduction, several requirements are typically imposed on formal relations that capture the notion of *abductive explanation*. Those requirements are not uniform; for example, in some contexts it may be feasible to define a preference ordering on candidate hypotheses which explain potential observations such that selecting the 'best' explanation is an option, whereas in other contexts it is not. The requirements for the ternary abductive explanation relation $\mid\!\approx$ presented below are quite typical, though; an elaborate account is given by Aliseda-Llera (1997). Note that the semantic connotation of the symbol '$\mid\!\approx$' is not coincidental, as abductive explanation is defined in terms of semantic entailment (denoted '$\models$').

Definition II.1 (abductive explanation $\mid\!\approx$)**.** *Let $\mathscr{L}$ be a classical logical language, $\Theta \subseteq \mathscr{L}$ a logical theory, $\Phi \subseteq \mathscr{L}$ an explanans and $\psi \in \mathscr{L}$ the explanandum.*

$\Theta, \psi \mid\!\approx \Phi$ *if and only if*

$\Theta \cup \Phi \models \psi$	&	(the explanans accounts for the explanandum)
$\Theta \not\models \psi$	&	(unexpectedness of the explanandum)
$\Phi \not\models \psi$	&	(non-triviality of the explanans)
$\Theta \cup \Phi \not\models \bot$		(consistency of the explanans)

It should be noted about Definition II.1 that it mentions a classical logical language, which in the context of this dissertation refers specifically to propositional logic and the ground fragment of first-order logic. Also note that explanantes (plural of 'explanans') are sets of clauses, whereas explananda (plural of 'explanandum') are facts; this is non-essential, but natural in regard to how the criteria for '$\mid\!\approx$' are stated in terms of '$\models$'. To illustrate abductive explanation we present a classical example, but first the consequence operator 'Th' is defined, as follows, where Φ is a set of logical expressions.

$$\mathsf{Th}(\Phi) = \{\phi \mid \Phi \models \phi\}$$

For the example, let the atomic propositions r, s, w respectively denote that it rains, that the sprinklers are on, and that the grass is wet. Furthermore, let $\Theta = \mathsf{Th}(\{r \rightarrow w, s \rightarrow w\})$

[2] In this work the term 'theory' is used to refer to a set of formulas closed under logical consequence.

be the logical theory describing that the grass is wet both if it rains and if the sprinklers are on. Given the observation w, observe that it is the case that $\Theta, w \mid\!\approx \{r\}$ and $\Theta, w \mid\!\approx \{s\}$, such that the observation of the grass being wet is abductively explained by the fact that it rains, as well as the fact that the sprinklers are on.

As Definition II.1 shows, abductive explanation can occur only if the observed fact ψ (the explanandum) is *unexpected*, meaning that it is not accounted for by the background theory (Aliseda-Llera, 1997). Furthermore, it is required that the explanans be *non-trivial*, showing in the fact that the explanandum itself is not considered a valid explanation. Another typical requirement is that the explanans should be *consistent* with the theory; after all, in classical logic anything can be derived from falsity ('ex falso quodlibet'), including the explanandum! Other criteria have also been proposed (Cox & Pietrzykowski, 1986; Aliseda-Llera, 1997); for example, explanations can be required to be *basic*, such that given explanations cannot be explained in terms of other explanations, and it is sometimes required that an abduced explanation is the *best* explanation according to some preference ordering, as stated earlier. Furthermore, explanations are sometimes required to be *minimal*, such that no fact subsumed by an explanation is itself an explanation. In regard to the example presented in the previous paragraph, note that our definition of abductive explanation allows for $\Theta, w \mid\!\approx \{r, s\}$, so that the fact that the sprinklers are on while it rains is also considered a valid explanation for the observation that the grass is wet. In order to restrict the search space for explanations and to ensure that particular constraints are respected, computational approaches to abduction typically require explananda to stem from a pre-specified set of *abducibles* (Kakas et al., 1998); an approach which we also take in later sections of this chapter (Section 3.5.1, to be precise).

The above exposition focuses on the requirements for abductive explanation, and by necessity is restricted in scope and depth. For further reading on abductive reasoning see Flach (1995) and Aliseda-Llera (1997), and also Kakas et al. (1998) who discuss the role of abduction in logic programming in relation to (amongst others) negation-as-failure and default reasoning. In regard to the outcome of a process of abductive reasoning, observe that if an explanandum is to be accounted for by the theory after the conclusion of that process, then the theory should have been updated (i.e. expanded or revised) with an explanans (Boutilier, 1996; Aliseda-Llera, 1997).

1.2 Nonmonotonicity

Abductive reasoning is *defeasible reasoning*, which means that conclusions reached on grounds of abduction can turn out to be false, even if their premises were true. Correspondingly, the relation $\mid\!\approx$ is nonmonotonic, a concept which is perhaps best understood in terms of its complement; the concept of *monotonicity*.

Definition II.2 (monotonicity)**.** *Let $\rightsquigarrow$ be a consequence relation on $\mathcal{L}$. Given the premises $\Phi, \Psi \subseteq \mathcal{L}$ and conclusion $\phi \in \mathcal{L}$, the relation $\rightsquigarrow$ is monotonic if and only if*

$$\forall \Phi, \Psi \subseteq \mathcal{L}\ \forall \phi \in \mathcal{L}: \\ (\Phi \rightsquigarrow \phi \quad \& \quad \Phi \subseteq \Psi) \qquad \Longrightarrow \qquad \Psi \rightsquigarrow \phi$$

That the relation $\not\approx$ of Definition II.1 is not monotonic (i.e. is nonmonotonic) can be illustrated by means of the earlier example, where given $\Theta = \mathsf{Th}(\{r \rightarrow w, s \rightarrow w\})$ and $\Theta' = \mathsf{Th}(\{r, r \rightarrow w, s \rightarrow w\})$ holds $\Theta \subseteq \Theta'$ but $\Theta', w \not\approx \{s\}$, because $\Theta' \models w$.

A central concept in nonmonotonic reasoning is that of *extensions* of a logical theory (Brewka et al., 2008). This can be understood by seeing that if there is need for nonmonotonic reasoning with respect to some background theory, then this theory apparently is incomplete in its description of the domain. In such a case, defeasible information inferred on grounds of nonmonotonic reasoning can be used to augment the theory. For abduction it holds that a theory can be seen as a representation of all its possible *abductive extensions*, i.e. completions of the theory with any of the hypotheses which can be abduced to account for particular explananda. Accordingly, abductive entailment on grounds of a theory can be defined by deductive entailment in the abductive extensions of that theory (Flach & Kakas, 2000). The notion of extensions is also useful in defining properties of a consequence relation. In that respect, *credulous inference*, where a fact can be inferred if it holds in some extension, is typically distinguished from *skeptical inference*, where a fact can be inferred if it holds in all extensions (Brewka et al., 2008). Those notions are further explored in Section 3, which formally approaches the topic of explaining observed agents' behavior, but first it is considered how the behavior of those agents can be generated.

2 A Simple Agent Programming Language

In this section we introduce MYAPL, a propositional BDI-based agent programming language with rules similar to the 'planning goal rules' (or 'PG-rules') of 2APL (Dastani, 2008) and SimpleAPL (Alechina et al., 2010). An agent implemented in this language has a library of behavioral rules at its disposition, which are applicable if certain conditions are satisfied. Specifically, the configuration of a MYAPL agent is taken to comprise sets of facts that constitute that agent's *belief base* and *goal base*, and a rule is taken to be applicable if the goal-related precondition of this rule is entailed by the goal base, and the belief-related precondition is entailed by the belief base. The grammar of behavioral rules is presented in Definition II.3. The basic elements of MYAPL program rules are atomic propositions (represented in the BNF by ⟨*atom*⟩) and primitive actions (represented by ⟨*primaction*⟩). Literals are composed of the set of atoms in union with their literal negation symbolized by '–'. Rules in a library defined by this grammar have the form '$n : \gamma <- \beta \mid \pi$', where '$n$' is a rule's numerical identifier, 'γ' represents the goal-related precondition, 'β' the belief-related precondition, and 'π' is a behavioral recipe (plan). Numerical identifiers of rules are used for technical simplicity in reference to sets of rules; without loss of generality, a library can be treated as a list of rules (like in, e.g., 2APL).

As stated before, a MYAPL rule is applicable if its goal condition is entailed by the agent's goal base, and its belief condition is entailed by the agent's belief base. It is assumed for technical simplicity that the entailment relations governing both the goal base and the belief base are defined in terms of set membership for literals; queries consisting of literals composed by means of 'and' and 'or' are taken to be interpreted as usual.

Definition II.3 (BNF grammar of MYAPL rule libraries)**.**

⟨*library*⟩ ::= ⟨*rule*⟩+
⟨*rule*⟩ ::= ⟨*id*⟩":"⟨*goalquery*⟩" <−"⟨*query*⟩" | "⟨*plan*⟩
⟨*literal*⟩ ::= ⟨*atom*⟩ | "−"⟨*atom*⟩
⟨*goalquery*⟩ ::= ⟨*literal*⟩ | ⟨*goalquery*⟩"and"⟨*goalquery*⟩
⟨*query*⟩ ::= ⟨*literal*⟩ | ⟨*query*⟩"and"⟨*query*⟩ | ⟨*query*⟩"or"⟨*query*⟩
⟨*plan*⟩ ::= ⟨*primaction*⟩ | ⟨*testaction*⟩ | ⟨*seqplan*⟩ | ⟨*ifplan*⟩ | ⟨*whileplan*⟩
⟨*testaction*⟩ ::= "B("⟨*query*⟩")" | "G("⟨*query*⟩")"
⟨*seqplan*⟩ ::= ⟨*plan*⟩";"⟨*plan*⟩;
⟨*ifplan*⟩ ::= "if"⟨*testaction*⟩"then {"⟨*plan*⟩"} else {"⟨*plan*⟩"}"
⟨*whileplan*⟩ ::= "while"⟨*testaction*⟩"do {"⟨*plan*⟩"}"

It should be noted that not all applicable rules are necessarily applied by the agent, and that whether a rule is actually applied can depend on various conditions. For example, the fact that a different rule for the same goal has already been applied can prevent application of a rule which would be applicable if only goals and beliefs were considered. Irrespective of the conditions governing rule application: if it is the case that some rule $n : \gamma <-\beta \mid \pi$ is successfully applied, then the agent selects the accompanying plan π for execution. This discussion of MYAPL semantics is continued in the next section with focus on plan execution, but first an example is given in terms of the program given in Listing II.1. The agent in that program has two rules, the first of which is applicable for achieving the goal that the grass is wet, given the belief that there is no rain coming. The plan accompanying this rule consists of simply turning on the sprinklers. The agent's second rule is applicable if it has the goal to have saved energy, and if it believes the sprinklers to be on while at the same time rain is coming down; in that case the agent turns the sprinklers off.

2.1 Semantics of Plan Execution

As defined in Definition II.3, the basic elements of plans are atomic actions and test actions, which can be composed by constructs for sequential composition (denoted ';'), conditional composition ('if-then-else' construct), and iterative composition ('while-do' construct). Those constructs are common in computer programming languages and the reader may very well be familiar with their meaning, but nevertheless an informal semantics is

```
1:  grass_wet     <-   -rain_coming   |
    { turn_sprinklers_on  }
2:  energy_saved  <-   sprinklers_on and raining     |
    { turn_sprinklers_off }
```

Listing II.1

presented below for the present agent programming context. It is assumed that the plan delimiters '{' and '}' may be omitted if no ambiguity arises.

Sequential composition: The expression '$\pi_1; \pi_2$' means that the agent is to execute plan π_1 first, followed by the plan π_2.

Conditional composition: The expression 'if ϕ then π_1 else π_2' means that if the agent tests on ϕ successfully it should execute the plan π_1, otherwise it should execute the plan π_2.

Iterative composition: The expression 'while ϕ do π' means that the agent should test on ϕ and subsequently execute π if the test succeeds, and continue this process of testing on ϕ and executing π, until the test on ϕ fails.

The above descriptions abstract from the intentional connotation of tests in order to avoid confusion. For example, the test action $\mathsf{B}(\phi)$ could be described as a test on whether the agent *believes* ϕ. If, however, the agent should have perfect information (possibly by design), then describing a successful test on $\mathsf{B}(\phi)$ as 'the agent knows ϕ' is perhaps more appropriate. This latter description typically applies to test actions of the form $\mathsf{G}(\psi)$, which represent goal reflection and — given the assumption that the agent is aware of its own goals — are best described by as tests on whether the 'agent knows it has the goal ψ'. In any case, such questions are avoided by the terminology used in the above descriptions. It could be argued that terms such as 'belief', 'desire', and 'intention' are best reserved for describing human agents, but in this dissertation (as is common in the literature on BDI-based software agents) our usage of those terms is more liberal, and they are used in reference to artificial agents as well.

As stated earlier, various conditions may determine whether an applicable rule is actually applied by an agent. Those conditions differ across agent programming languages: in 2APL (Dastani, 2008), for example, rules are processed in order, such that the first applicable rule is always applied and the corresponding plan selected. As example of a contrasting approach, early versions of the GOAL agent language (Hindriks, 2001) can be considered. There, so-called 'action rules' were utilized, each of which was considered for application, and in case multiple action rules applied then the GOAL interpreter arbitrarily chose and applied one of those applicable rules. What exactly occurs when a rule is applied also differs per agent programming language. In order to allow us to be specific on this topic without providing formal semantics of the interpretation of the agent programming language MYAPL used in this chapter, the following general characteristics are here distinguished.

Execution strategy: Every agent has a *list* of plans it has selected in order to achieve its goals. This list is referred to as the agent's *intention base*, and its interpretation is loosely based on the notion of an 'intention stack' as described by Rao (1996). It is assumed for plans in the intention base that the agent has recorded the goal for which they were selected, corresponding to plan selection in 2APL (Dastani, 2008). Of the plans in its intention base, the agent executes only the first action of the first plan per execution step. Agents can have either an *interleaving* or a *non-interleaving* execution strategy, in terms of Alechina et al. (2007). An agent that has an interleaving strategy places the plan of which it has just executed the first action last in the list

of plans, whereas an agent that has a non-interleaving strategy does not rotate plans after executing actions.

Types of commitment: Plans that an agent has selected as a result of the application of goal-directed rules are regarded as its intentions, which lead to states of affairs the agent has chosen and committed to (Cohen & Levesque, 1990). Rao & Georgeff (1991) give a formal account of commitment in different degrees, to wit: *blind*, *single-minded*, and *open-minded* commitment. Paraphrasing that work to fit our terminology, a blindly committed agent maintains an intention until it believes the relevant goal to have been achieved, a single-minded agent maintains an intention until the goal has been achieved or is deemed unachievable, and an open-minded agent maintains an intention as long it still has the relevant goal.

Instead of providing formal semantics of agent operation, the following is assumed:

- MYAPL agents have a non-interleaving execution strategy.
- MYAPL agents can be blindly, single-mindedly, or open-mindedly committed.

2.2 Remarks

Assuming the non-interleaving execution strategy instead of allowing for any execution strategy is done for technical simplicity, at the expense of the approach sketched in this chapter not being generally applicable (in any case not if agents do not follow a non-interleaving execution strategy). At first glance this may seem like a severe restriction, but it should be noted that in certain cases this restriction can be alleviated. Specifically, if the required factors are under control of the system designer (for example in games or other virtual environments, where our approach could be utilized to implement believable virtual characters), then the behavior of agents that do in reality follow a non-interleaving strategy can be simplified for the sake of explanation by presenting annotated actions to the observer of the agent's actions.[3] A second remark concerns the handling of plans, where it is assumed that an agent which is executing some plan does not switch to another plan until its commitment type allows it to do so, at which point the initial (un)finished plan is either finished, or dropped.

3 Explaining the Observed Behavior of BDI-Based Agents

An observer's reasons for explaining the behavior of observed entities may be manifold, and may include its goal to respond to that behavior in a cooperative or obstructive way. Whatever the motivation, explanation of behavior is an activity humans practice frequently, and

[3] For example, an agent might interleave the actions of the plan 'action 1 followed by action 2' with those of the plan 'action 3 followed by action 4', possibly resulting in the agent performing the action sequence '1', '3', '2', '4'. If it is unknown whether or not the agent interleaves its plans then an observer attempting to explain the agent's actions must attempt to disentangle the observed sequence of actions. However, by annotating the actions '1' and '2' as stemming from the same plan and doing likewise for the actions '3' and '4', the observer is spared this effort and our approach (which assumes a non-interleaving execution strategy) can still be applied.

a possible outcome of that explanative process are references to the observed entity's presumed mental state (Baron-Cohen, 1995; Dennett, 1987). Sometimes people even refer in mentalistic terms to entities to which they would *not* grant the possession of actual mental states. Examples are commonplace: the clerk behind the desk exclaiming that the system does not "wish to cooperate" today; or a repairman examining the presumedly broken thermostat, saying "It seems to believe it's hot in here, but it's actually freezing!", at the same time being firmly convinced of the fact that a thermostat does not, in reality, have beliefs. This anecdotal evidence goes to illustrate that the mentalistic level of description is at times considered useful by humans even if it is at the same time not considered ontologically accurate, something which has also been recognized in the A.I. literature (McCarthy, 1990).

II.3

Software agents based on the BDI paradigm are an interesting case when referred to in mentalistic terms, because those terms are *grounded in computation*. This means that there exists some computer-based representation that can be identified as referent of the mentalistic expression; e.g., a particular data structure representing the goal or belief of the agent. It is interesting to compare this matter to the philosophical discussion concerning intentionality in Chapter I. There it was stated that some philosophers (called 'realists' by Dennett) believe that intentional concepts such as 'belief' are objective and can be reduced to physical substrates, whereas others ('interpretationists') are convinced that this is not necessarily the case because those concepts are in principle subjective. Fortunately, things are more straightforward when it comes to software agents of which one knows the design, because in that case there exists no uncertainty about the representation underlying agents' beliefs or goals, nor about their involvement in producing behavior. In 2APL, for example, beliefs are instantiated by a set of Prolog clauses, whereas goals are instantiated as lists (Dastani, 2008). In comparison, the Jadex belief base stores strings that identify specific beliefs, which are allowed to be arbitrary Java objects, and the Jadex goal base distinguishes four different types of goals, each with specific attributes (Pokahr et al., 2005). In any of those cases, the correctness of references to the agents' internals can in principle be empirically verified (assuming, of course, that inspection of the agent is possible), which is opposed to explanations referring to the internals of, say, humans.

Abductively explaining observed behavior of agents, along the lines of explanatory abduction, assumes the existence of a logical theory that pertains to the behavior of the agent. Assume Θ is such a theory, so that if explanations are given in reference to the internal constitution of the agent, this looks as follows in analogy with Definition II.1.

$$\Theta,\ \textit{'external behavior'} \mid\!\approx\ \textit{'internal condition'}$$

In the following sections the semi-formal notion presented above is made explicit, with the perspective in mind of a beholder referred to as the *observer*, whose single function it is to explain the behavior of observed agents. This perspective serves the purpose of separating concerns, and to avoid confounding the reader by questions which might be raised if the observing party were presented as another agent.

3.1 Framing the Abductive Theory

In order to formalize abductive explanation of a BDI-based agent's behavior, a good starting point is to formulate a logical background theory describing that behavior in relation to the configuration of the agent. Also, it must be made explicit which aspect of the agent's behavior is considered observable, and which facts are considered as possible abductive hypotheses for observed behavior. The triplet 'background theory'/'observables'/'abducibles' is referred to here as 'abductive theory', which will be formally explicated in the following sections. In reference to Section 1, it holds that the explanantes stem from the observables, and explananda from the abducibles.

The BDI-based agents whose behavior is reasoned about are assumed to be programmed in the MYAPL language presented in Section 2, and to operate in line with the informal semantics sketched in that section. Recall that such agents have goal-directed rules of the form '$n : \gamma <- \beta \mid \pi$', which are in principle applicable if 'γ' follows from that agent's goals and 'β' from its beliefs. Also recall that various conditions determine rule application, such that the fact that a rule is applicable does not necessarily mean that this rule is actually applied (although rules which are not applicable are never applied). In formulating an abductive theory with regard to the agent's operation, a conditional relation must be established that reflects rule applicability, and which allows for abductively inferring particulars regarding the agent's configuration on grounds of behavior it is observed to perform.

The behavior that an agent programmed in the MYAPL language performs, based on the plan it is executing, consists principally of primitive actions and test actions — denoted in the BNF of Definition II.3 by ⟨*primaction*⟩ and ⟨*testaction*⟩ elements, respectively — which can be composed by means of the various programming constructs mentioned in Section 2. Of those basic actions, only the *primitive actions* of the agent are considered to be (in principle) observable to the observer that attempts to explain the agent's actions. In order to abduce the rule an agent has applied, the observable behavior which is generated on grounds of plans must be somehow extracted. This topic is tackled formally in Section 3.3, but first some notice is taken of perceptory conditions that may influence observation.

3.2 Perceptory Conditions

Essentially, as noted in Chapter I, relating an agent's observed behavior to observable behavior generated by its plans is a case of *intention recognition* (in the sense that the plan which the agent has selected, and from which observed actions originate, is being inferred) or *plan recognition* (in the sense that a subset of plans from the agent's library is being inferred). In our view, which is consistent with the literature on this topic (Sadri, 2011; Carberry, 2001), a system in which plan recognition takes place consists of one or more *agents*, an *environment*, and an *observer*. Each of these three components of a plan recognition system can influence the nature of the observations available to the observer, and here we present our take on the role of those components in the recognition process in order to set the stage for presenting our formalization.

In regard to the agent component of a plan recognition system, Cohen et al. (1981) identify a distinction between *keyhole* recognition and *intended* recognition. In keyhole

recognition, as the term suggests, the agent is unaware of the fact that it is being observed and does not interfere with the recognition process. Intended recognition is the setting in which the agent is aware of observation, and structures its activities in order to aid the recognition process. Geib & Goldman (2001) have identified *adversarial* recognition as a third class, and state it to be the natural complement of intended recognition as it concerns the case in which the agent does *not* want its plan to be recognized, and acts in order to thwart recognition. In our approach keyhole recognition is assumed, although adversarial recognition can be handled if adversarity is restricted to the agent *hiding* its actions from the observer's view. Specifically, it is assumed that adversarial recognition does not include misleading behavior of the agent, a scenario which is classified as the recognition of *diversionary* intentions by Sadri (2011).

The environment component is understood here as the medium through which perceptory data regarding the agent's actions 'travels' before it reaches the observer. This is taken to include any factor which may influence the observer's perception; i.e. its representation of the event that is presumed to be generated by the agent having performed an action. The environment is assumed to be either *noise-free* or *noisy*. In our particular approach, this means that either the environment does not at all influence whether or not the observer perceives an action, or it does. In case that the environment does indeed influence the observer's perception, then this occurs in the sense that some actions are perceived by the observer, whereas others are not. A more fine-grained approach could also involve uncertainty on part of the observer, in the sense of logical disjunction (e.g. "action α or α' was observed") or probabilities (e.g. "40% certainty that action α was observed, 60% certainty that it was α'").

A third component of any plan recognition system is the observer; the entity which perceives the actions of the agent. In our view, an observer may be able to influence its own perception of an agent's actions by directing the focus of its attention; a claim which makes all the more sense if the observer is itself 'just another agent' with limited (cognitive) resources. Regardless of specifics of how the notion of attention direction could be realized in practice, it is here assumed that the observer can have either *absolute*, *late*, or *intermittent* attention for the agent's behavior. The aforementioned components of a plan recognition system can influence the perception of the observer, and can essentially be broken down to the following three *perceptory conditions*:

Complete observation: In the case of complete observation it is assumed that the observer perceives every observable action of the agent.

Late observation: In the case of late observation it is assumed that the observer perceives every observable action of the agent from the moment observation starts, but has possibly failed to see some of the agent's initial actions.

Partial observation: In the case of partial observation it is assumed that the observer perceives some of the actions of the agent, but does not perceive others.

These perceptory conditions are intrinsically related to characteristics of the agent, the observer, and the environment. The ideal perceptory condition (from an observer's point of view) is, of course, that of complete observation, because in this case the observer has the maximum amount of information regarding the actions performed by the agent. Things

are often not ideal, though, and it can be easily seen that if some of the factors influencing the perceptory conditions vary, then late or partial observation quickly rears its head as being the actual perceptory condition; or, at best, the worst-case condition which may not actually occur but which the observer should take into consideration. To see this, consider an exemplary ideal scenario in which keyhole or intended recognition occurs, the observer's attention is absolute, and the environment noise-free. In such a case observation is complete. However, if the agent becomes adversarial, or the observer's attention wavers, or the environment becomes noisy, then it *possibly* occurs that the observer fails to see some of the agent's actions so that complete observation can only be regarded as a best-case condition, and the observer also has to consider late or partial observation.

It is important for approaches to plan recognition to handle incomplete observation (Sadri, 2011; Carberry, 2001), if only because it may be undesirable to always have complete observation; think of virtual environments such as games or training applications, where it might be possible to make an observer's perception complete, but at the same time it might not be desirable to do so because this would be detrimental to believability. For example, a virtual character involved in a dangerous situation might, for the sake of believability, not be given complete observation of the actions of another character, even if this is technically feasible and it is in principle believable that this character would 'prefer' to have absolute attention for the other's behavior.

3.3 Extracting Observable Behavior from Plans

In this section, a formal approach is taken to defining the observables of the abductive theory pertaining to the behavior of BDI-based agents. Plans in the language MYAPL, as defined in Definition II.3 through ⟨*plan*⟩-elements in the BNF grammar, correspond to the class of while-programs (Harel et al., 2000) in terms of expressiveness. In turn, while-programs are a subclass of regular programs adhering to certain syntactic constraints, and it is recognized, based on the fact that programs of aforementioned classes are simply procedural descriptions, that process languages (Baeten & Weijland, 1990) can be employed to express such programs succinctly. In the following we define a process language called the *plan language* which is sufficiently expressive to describe MYAPL plans, as defined in Definition II.3 through ⟨*plan*⟩-elements. This plan language should be sufficiently expressive to capture MYAPL plans, and for that reason employs a construct ';' for sequential composition, '+' for non-deterministic choice, and superscript '$*$' for iteration. The plan language is defined inductively as follows, on top of the basic propositional logical language $\mathscr{L}_0$ of which the definition is left for a later stage.

Definition II.4 (plan language $\mathscr{L}_\Pi$)**.** *Let* Act *comprise the primitive* MYAPL *actions, and let* $\mathscr{L}_0$ *be the logical counterpart of* MYAPL *test (or query) expressions. The language* $\mathscr{L}_\Pi$ *with typical element* π *is then the smallest set closed under the following clauses.*

- *If* $\alpha \in \mathsf{Act}$ *and* $\phi \in \mathscr{L}_0$ *then* $\alpha, \mathbf{B}(\phi)?, \mathbf{G}(\phi)?, \neg\mathbf{B}(\phi)?, \neg\mathbf{G}(\phi)? \in \mathscr{L}_\Pi$.
- *If* $\pi, \pi' \in \mathscr{L}_\Pi$ *then* $\pi;\pi', \pi + \pi', \pi^* \in \mathscr{L}_\Pi$.

To express MYAPL plans in the language $\mathscr{L}_\Pi$ of Definition II.4, a translation function τ_p is defined below, therein following Harel et al. (2000) with regard to the translation of

conditional and iterative composition. Note that the exact form of tests on elements of $\mathscr{L}_0$ is not relevant at present but will be specified in more detail further on in this thesis. Here, it suffices to know that $\mathscr{L}_0$ consists of propositional expressions describing MYAPL queries, such that the action $\mathbf{B}(\phi)?$ is the $\mathscr{L}_\Pi$-counterpart of the program expression $\mathsf{B}(\psi)$ if ϕ corresponds to the translation of ψ to $\mathscr{L}_0$. This translation is done by the function τ_q, mentioned here for completeness and defined in full in the next chapter (Definition III.8).

II.3

Definition II.5 (MYAPL plan translation function τ_p)**.** *Let* Act *correspond to the set of primitive* MYAPL *actions. The function* τ_p *maps* ⟨*plan*⟩ *elements into* $\mathscr{L}_\Pi$*, as follows.*

$$
\begin{aligned}
&\tau_p(\alpha) && = \alpha \quad \textit{if} \quad \alpha \in \mathsf{Act}\\
&\tau_p(\mathsf{B}(\phi)) && = \mathbf{B}(\tau_q(\phi))?\\
&\tau_p(\mathsf{G}(\phi)) && = \mathbf{G}(\tau_q(\phi))?\\
&\tau_p(\pi;\pi') && = \tau_p(\pi);\tau_p(\pi')\\
&\tau_p(\mathsf{if}\ \phi\ \mathsf{then}\ \{\pi\}\ \mathsf{else}\ \{\pi'\}) && = (\tau_p(\phi);\tau_p(\pi)) + (\neg\tau_p(\phi);\tau_p(\pi'))\\
&\tau_p(\mathsf{while}\ \phi\ \mathsf{do}\ \{\pi\}) && = (\tau_p(\phi);\tau_p(\pi))^*;\neg\tau_p(\phi)
\end{aligned}
$$

Primitive actions in the set 'Act' constitute the observables of the agent's behavior, and are known to stem from plans that are part of the set of rules the agent has at its disposition. Plans do not provide constructs for parallel execution, nor do the semantics of rule interpretation given in Section 2.1. In the absence of parallelism, primitive actions are by definition executed in sequence, and, accordingly, observable sequences are simply sequences of primitive actions, as defined by the following *percept language* $\mathscr{L}_\Delta$. It should be noted that $\mathscr{L}_\Delta \subseteq \mathscr{L}_\Pi$ for technical convenience, and also that the explananda (i.e. observations which are to be explained) considered in later sections are described by this language of observables.

Definition II.6 (percept language $\mathscr{L}_\Delta$)**.** *Let* Act *be a set of primitive actions. The percept language* $\mathscr{L}_\Delta$ *with typical element* δ *is then the smallest set closed under the following clauses.*

- *If* $\alpha \in \mathsf{Act}$ *then* $\alpha \in \mathscr{L}_\Delta$.
- *If* $\delta, \delta' \in \mathscr{L}_\Delta$ *then* $\delta;\delta' \in \mathscr{L}_\Delta$.

In order to extract the possible observable sequences of primitive actions generated by some plan $\pi \in \mathscr{L}_\Pi$, the function *OS* is defined in Definition II.7 that translates expressions from the plan language $\mathscr{L}_\Pi$ to the percept language $\mathscr{L}_\Delta$. In doing so, this function deals with the various compositional constructs appropriately, and it is especially noteworthy that test actions on propositions of the type $\mathscr{L}_0$ are filtered out by this function, in line with the rationale that only (external) primitive actions of the agent can be perceived by the observer and (internal) test actions cannot. Note that $\mathbb{N}_0$ denotes the natural numbers including zero, whereas $\mathbb{N}_1$ denotes $\mathbb{N}_0 \setminus \{0\}$.

Definition II.7 (observability function *OS*)**.** *The function* $OS : \mathscr{L}_\Pi \longrightarrow \wp(\mathscr{L}_\Delta)$*, which translates plans to sets of observable sequences, is defined as follows, given* $\alpha \in \mathsf{Act}$ *and* $\phi \in$

$\{\mathbf{B}(\psi), \neg\mathbf{B}(\psi), \mathbf{G}(\psi), \neg\mathbf{G}(\psi) \mid \psi \in \mathscr{L}_0\}$. *For convenience, assume existence of a* 'skip' *action.*

$$
\begin{aligned}
OS(\alpha) &= \{\alpha\} \\
OS(\phi?) &= \emptyset \\
OS(\pi;\pi') &= OS(\pi) \bullet OS(\pi') \\
OS(\pi + \pi') &= OS(\pi) \cup OS(\pi') \\
OS(\pi^*) &= \bigcup_{n \in \mathbb{N}_0} OS(\pi^n)
\end{aligned}
$$

where $\pi^n = \pi;\pi^{n-1}$, $\pi^0 = \mathsf{skip}$, *and* $OS(\mathsf{skip}) = \emptyset$

The composition operator $\bullet : \wp(\mathscr{L}_\Delta) \times \wp(\mathscr{L}_\Delta) \longrightarrow \wp(\mathscr{L}_\Delta)$ *is defined as follows.*

$$
\begin{aligned}
\Delta \bullet \Delta' &= \{\delta;\delta' \mid \delta \in \Delta, \delta' \in \Delta'\} && \text{if } \Delta \neq \emptyset \text{ and } \Delta' \neq \emptyset \\
\Delta \bullet \Delta' &= \Delta && \text{if } \Delta \neq \emptyset \text{ and } \Delta' = \emptyset \\
\Delta \bullet \Delta' &= \Delta' && \text{if } \Delta = \emptyset \text{ and } \Delta' \neq \emptyset \\
\Delta \bullet \Delta' &= \emptyset && \text{otherwise}
\end{aligned}
$$

It is of note that the function *OS*, which extracts observable sequences, shows similarity to the function *CS* defined by Harel et al. (2000), which extracts computation sequences. This similarity is not coincidental as those functions fulfill a similar role; the difference being that *OS* filters out actions that are deemed unobservable. Given the formal machinery introduced up to this point, the pieces are now in place for introducing the background theory employed by the observer in reasoning about agents' behavior.

3.4 Formulating the Background Theory

In this section a logical theory is presented that describes the operation of a MYAPL agent, based on its set of behavioral rules. It was stated in Section 3.1 that the MYAPL language is propositional; the language that describes the operation of the agent is in turn a (ground) predicate language $\mathscr{L}_1$, defined as follows.

Definition II.8 (predicate language $\mathscr{L}_1$)**.** *Let* Act *be a set of primitive actions,* Atom *a set of atomic* MYAPL *propositions, and* $\mathsf{Lit} = \{p, -p \mid p \in \mathsf{Atom}\}$. *The language* $\mathscr{L}_1$ *is then the smallest set closed under the following clauses.*

- *If* $\phi \in \mathsf{Lit}$ *then* $\mathsf{belief}(\phi), \mathsf{goal}(\phi) \in \mathscr{L}_1$.
- *If* $\alpha \in \mathsf{Act}$ *and* $n \in \mathbb{N}_1$ *then* $\mathsf{obs}(\alpha, n), \mathsf{seen}(\alpha, n), \mathsf{rule}(n), \mathsf{seq}(n) \in \mathscr{L}_1$.
- *If* $\phi, \psi \in \mathscr{L}_1$ *then* $\neg\phi, \phi \vee \psi \in \mathscr{L}_1$.

The logical connectives $\wedge, \rightarrow, \leftrightarrow$ can be derived from the above as usual. The informal interpretation of the predicates of $\mathscr{L}_1$ is given below; note hereby that the literal MYAPL propositions have been reified as arguments to the predicates of $\mathscr{L}_1$, so that this language allows for making statements *about* MYAPL agents, as follows.

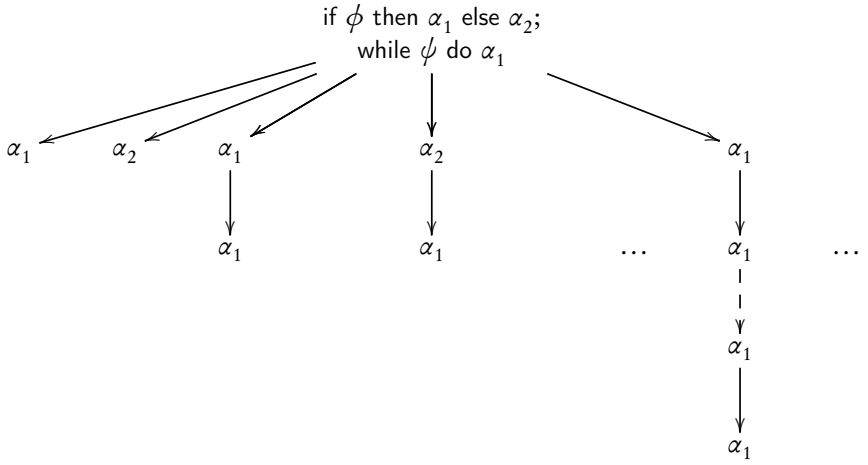


Figure II.1: Tree-like representation of a MYAPL plan.

- rule(n): 'the agent has applied rule n'
- belief(ϕ): 'ϕ followed from the agent's beliefs at the time it applied its rule'
- goal(ϕ): 'ϕ followed from the agent's goals at the time it applied its rule'
- seq(n): 'the observer expects to see sequence n'
- obs(α, n): 'action α is observable as the n'th action'
- seen(α, n): 'action α is observed as the n'th action'

Given the above predicates, conditional relations can be formulated as part of a theory that describes operation of the agent. Those conditionals state that if the agent has applied a particular rule then, specifically, those actions are observable which constitute an observable sequence of the plan selected on grounds of that rule. In this respect it may help to think of a plan as a tree that represents the plan's possible executions, of which the agent 'selects' a single branch. The primitive actions occurring in that branch then constitute the resulting observable sequence, and all observable sequences are in the output of the *OS* function applied to that plan. As Definition II.5 shows, choice points in plans occur on grounds of the if-then-else construct, and — with the tree analogy in mind — those correspond to points where different branches are grown. Furthermore, note that on grounds of the while-do construct there may be a countably infinite number of branches of finite depth, as illustrated by Figure II.1.

The branches of a plan can be enumerated, but in the case of defining a theory of observability it suffices to enumerate the distinct observable sequences of a plan. The function $\iota_\delta : \mathscr{L}_\Delta \longrightarrow \mathbb{N}_1$ is therefore defined to assign numerical identifiers to action sequences. Using the tools given in preceding sections, a part of the theory describing the behavior of agents can now be defined. In the following, the symbol '$\mathscr{R}$' is used to refer to a set of

MyAPL rules. Without loss of generality, it is assumed that this set of rules belongs to a single agent whose behavior is perceived by the observer; in case that multiple agents should be observed (which is allowed for but which is not treated explicitly in this chapter) then the set of rules can be indexed with an identifier of the agent to whom it belongs, and the same can be done with percepts.

Definition II.9 (observability clauses $C_{\mathscr{R}}$)**.** *Let $\mathscr{R}$ be a set of* MyAPL *rules, and OS the function defined in Definition II.7. The set of logical clauses $C_{\mathscr{R}} \subseteq \mathscr{L}_1$ concerning the observability of the behavior of an agent with those rules is then defined as follows.*

$$\forall(n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}\ \forall(\alpha_1;\ldots;\alpha_i) \in OS(\pi):$$
$$(\mathsf{rule}(n) \wedge \mathsf{seq}(\iota_\delta(\alpha_1;\ldots;\alpha_i))) \rightarrow (\mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)) \quad \in C_{\mathscr{R}}$$

The implicatory relation expressed by the clauses in $C_{\mathscr{R}}$ defines the actions of observable sequences as being individually observable if the agent has applied some rule and is executing a particular branch of the accompanying plan, which in this case is referred to by identifying (using ι_δ) the single observable sequence that stems from that branch. The above relation holds if all actions of the agent are expected to be observed by the observer, but, as stated before, in cases of incomplete observation it may be the case that only some of the agent's actions are. In the following subsections the theory is extended to handle cases of incomplete observation as well.

3.4.1 Incremental Observation and the Perceptory Conditions

The conditional relation sketched in Definition II.9 posits rule application and the expectation of seeing a (complete) observable sequence as premise to a sequence of actions which are considered individually observable at particular positions. A typical desideratum for an approach to plan recognition (Sadri, 2011; Carberry, 2001) is that it should be able to deal with *incremental observation*, so that a plan can be recognized while the agent is executing it and not just after the agent has completed its execution. To this extent, structural relations can be defined that allow for relating percepts to particular observable sequences under different perceptory conditions. Apart from enabling recognition with incremental observation, those structural relations can also be employed to deal with the fact that observation might be incomplete.

In Section 3.2 the three perceptory conditions of complete, late, and partial observation were put forward, and are briefly recalled here. Complete observation is the case in which every action performed by the agent is also perceived by the observer. This means that if observation is incremental, an observed sequence must be the uninterrupted initial segment of some observable sequence. In the case of late observation, where the observer perceives every action but might have missed initial actions, an observed sequence must be an uninterrupted segment occurring somewhere in an observable sequence. Finally, partial observation entails that the observer does perceive some actions but may fail to perceive others, which means that the observed actions must appear somewhere throughout an observable sequence in the order of their observation, possibly with other actions occurring

in between. The perceptory conditions of complete, late, and partial observation are captured by partial orders on $\mathscr{L}_\Delta$ referred to as *prefix*, *substring*, and *dilution*, respectively, and denoted $\underline{\blacktriangleright}$, $\underline{\blacktriangledown}$, and $\underline{\odot}$.[4]

Definition II.10 (prefix $\underline{\blacktriangleright}$, substring $\underline{\blacktriangledown}$, and dilution $\underline{\odot}$ relations)**.** *The structural relations* $\underline{\blacktriangleright}$, $\underline{\blacktriangledown}$, *and* $\underline{\odot}$ *are partial orders on the language* $\mathscr{L}_\Delta$*, and defined as follows.*

II.3

$$\underline{\blacktriangleright} = \{(\delta,\delta),(\delta,\delta;\delta') \mid \delta,\delta' \in \mathscr{L}_\Delta\}$$
$$\underline{\blacktriangledown} = \{(\delta,\delta),(\delta,\delta;\delta'),(\delta,\delta';\delta),(\delta,\delta';\delta;\delta'') \mid \delta,\delta',\delta'' \in \mathscr{L}_\Delta\}$$
$$\underline{\odot} = \bigcup_{n\in\mathbb{N}_0} \mathsf{dil}_n(\underline{\blacktriangledown})$$

The operator dil_n*, used in definition of* $\underline{\odot}$*, is defined as follows.*

$$\mathsf{dil}_n(\underline{\blacktriangledown}) = \mathsf{dil}_{n-1}(\underline{\blacktriangledown}) \cup \{(\delta;\delta'',\delta';\delta''') \mid (\delta,\delta'),(\delta'',\delta''') \in \mathsf{dil}_{n-1}(\underline{\blacktriangledown})\}$$
for $n > 0$*, where* $\mathsf{dil}_0(\underline{\blacktriangledown}) = \underline{\blacktriangledown}$

The above relations on the percept language $\mathscr{L}_\Delta$ are also related to each other, as the following proposition shows.

Proposition II.1. *It holds for the structural relations of Definition II.10 that*

$$(\underline{\blacktriangleright}) \subseteq (\underline{\blacktriangledown}) \quad \& \quad (\underline{\blacktriangledown}) \subseteq (\underline{\odot})$$

Proof. The fact that $(\underline{\blacktriangleright}) \subseteq (\underline{\blacktriangledown})$ follows straightforwardly from the set comprehension notation used in the definition of $\underline{\blacktriangledown}$. Furthermore, $\underline{\odot}$ is a superset of $\mathsf{dil}_0(\underline{\blacktriangledown}) = \underline{\blacktriangledown}$, and $(\underline{\blacktriangledown}) \subseteq (\underline{\odot})$ therefore follows immediately as well. □

Considering Definition II.9 it turns out that the prefix relation is already captured by the theory based on the clauses of $C_{\mathscr{R}}$, as is shown by the following proposition.

Proposition II.2. *Let* $C_{\mathscr{R}}$ *be the smallest set closed under Definition II.9 with respect to a set of* MyAPL *rules* $\mathscr{R}$*, let and* $\Theta = \mathsf{Th}(C_{\mathscr{R}})$*. It then holds that*

$$\forall(n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}\ \forall\delta \in OS(\pi)\ \forall\alpha_1;\ldots;\alpha_i \in \mathscr{L}_\Delta :$$
$$\alpha_1;\ldots;\alpha_i \underline{\blacktriangleright} \delta \implies \exists\phi \in \mathscr{L}_1 : \quad (\{\phi\} \not\models \mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)$$
$$\& \quad \Theta \cup \{\phi\} \models \mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i))$$

[4] In earlier work (Sindlar et al., 2008) we used the term 'subsequence' for $\underline{\blacktriangledown}$, which is more commonly used in reference to the relation $\underline{\odot}$. In order to conform to standard terminology but avoid confusion with respect to our earlier work, use of the term 'subsequence' is omitted and replaced by 'substring', but 'dilution' is retained in favor of 'subsequence'.

Proof. Take any $(n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}$. Given $\delta = \alpha_1;\ldots;\alpha_j$, if $\delta \in OS(\pi)$ then on grounds of Definition II.9 holds $(\mathsf{rule}(n) \wedge \mathsf{seq}(\iota_\delta(\delta)) \rightarrow (\mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_j,j)) \in \Theta$, where it should be noted that if $\phi = (\mathsf{rule}(n) \wedge \mathsf{seq}(\iota_\delta(\delta)))$ then $\phi \in \mathscr{L}_1$ and $\{\phi\} \not\models \mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)$. Furthermore, observe that if $\alpha_1;\ldots;\alpha_i \blacktriangleright \delta$ then $\delta = \alpha_1;\ldots;\alpha_i;\ldots;\alpha_j$, where $j > i$, or $\delta = \alpha_1;\ldots;\alpha_i$. In either case $\Theta \cup \{\phi\} \models \mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)$, which is straightforward in the latter case and in the former follows from the fact that $\{\mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i) \wedge \ldots \mathsf{obs}(\alpha_j,j)\} \models \mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)$. □

It should be noted that the above proposition mentions the existence of some elements of $\mathscr{L}_1$ which can be employed in order to account for the observability of actions, and in doing so provides a basis for identifying abducibles. Also, it shows how Definition II.9 lays the foundation for explanation of an agent's observed behavior with incremental observation, because by considering every individual action of an entire observable sequence to be observable the prefix condition is automatically met, as observability of the entire sequence entails observability of its prefixes. A similar relation can be found in cases of incomplete observation, but it should be noted that it then holds that an action which occurs at some position in an observable sequence can actually be observed at an 'earlier' position if preceding actions have been missed. To capture this fact in a formal theory pertaining to an observed agent's behavior, that theory should incorporate the appropriate structural relations.

For the perceptory condition of late observation, remember that its underlying rationale is that an observer possibly fails to have seen initial actions, but 'expects' to see all actions from the point it starts observing the agent's behavior. Thus, an appropriate relation underlying a logical formulation of this condition is the *suffix* relation on observable sequences, which is denoted '$\underline{\blacktriangleleft}$' and defined as follows — noting $(\underline{\blacktriangleleft}) \neq (\blacktriangleright^{-1})$.

Definition II.11 (suffix relation $\underline{\blacktriangleleft}$)**.** *The structural relation $\underline{\blacktriangleleft}$ is a partial order on $\mathscr{L}_\Delta$, defined as follows.*

$$\underline{\blacktriangleleft} = \{(\delta,\delta),(\delta,\delta';\delta) \mid \delta,\delta' \in \mathscr{L}_\Delta\}$$

Given this structural relation on observable sequences, a theory can be defined which formalizes the logical relation between, on the one hand, the rule that an agent has applied and the sequence of actions the observer expects to see, and, on the other hand, the actions the observer considers to be observable as a result of this. This can be done by assigning an identifier to each suffix and proceeding similarly to how this was done with complete observable sequences in Definition II.9.

Definition II.12 (late observability clauses $L_{\mathscr{R}}$)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, and OS the function of Definition II.7. The set of logical clauses $L_{\mathscr{R}} \subseteq \mathscr{L}_1$ pertaining to late observability of the behavior of an agent with the rules $\mathscr{R}$ is then defined as follows.*

$$\begin{aligned}
&\forall (n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}\ \forall \delta \in OS(\pi)\ \forall \alpha_1;\ldots;\alpha_i \in \mathscr{L}_\Delta : \\
&\quad (\alpha_1;\ldots;\alpha_i \underline{\blacktriangleleft} \delta) \implies \\
&\qquad (\mathsf{rule}(n) \wedge \mathsf{seq}(\iota_\delta(\alpha_1;\ldots;\alpha_i))) \rightarrow (\mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)) \quad \in L_{\mathscr{R}}
\end{aligned}$$

As Proposition II.3 proves, the late observability clauses defined in Definition II.12 capture the substring relation as intended.

Proposition II.3. *Let $L_{\mathscr{R}}$ be the smallest set closed under Definition II.12 with respect to some set of* MyAPL *rules $\mathscr{R}$, and let $\Theta = \mathsf{Th}(L_{\mathscr{R}})$. It then holds that*

$$\begin{aligned}&\forall(n:\gamma <\!\!-\beta \mid \pi)\in\mathscr{R}\ \forall\delta\in OS(\pi)\ \forall\alpha_1;\ldots;\alpha_i\in\mathscr{L}_\Delta:\\ &\quad \alpha_1;\ldots;\alpha_i \mathbin{\underline{\blacktriangledown}} \delta \implies \exists\phi\in\mathscr{L}_1:\ (\{\phi\}\not\models \mathsf{obs}(\alpha_1,1)\wedge\ldots\wedge\mathsf{obs}(\alpha_i,i)\\ &\qquad \&\ \ \Theta\cup\{\phi\}\models \mathsf{obs}(\alpha_1,1)\wedge\ldots\wedge\mathsf{obs}(\alpha_i,i))\end{aligned}$$

II.3

Proof. Take any $(n:\gamma <\!\!-\beta \mid \pi)\in\mathscr{R}$ and let $\delta\in OS(\pi)$. Observe that if $\alpha_1;\ldots;\alpha_i \mathbin{\underline{\blacktriangledown}} \delta$ then $\exists\delta'\in\mathscr{L}_\Delta:(\delta' \mathbin{\underline{\blacktriangleleft}} \delta\ \&\ \alpha_1;\ldots;\alpha_i \mathbin{\underline{\blacktriangleright}} \delta')$. Let $\delta'=\alpha_1;\ldots;\alpha_j$, such that on grounds of Definition II.12 holds $(\mathsf{rule}(n)\wedge\mathsf{seq}(\iota_\delta(\delta'))\rightarrow(\mathsf{obs}(\alpha_1,1)\wedge\ldots\wedge\mathsf{obs}(\alpha_j,j))\in\Theta$. Given $\phi=(\mathsf{rule}(n)\wedge\mathsf{seq}(\iota_\delta(\delta'))$, such that $\phi\in\mathscr{L}_1$, and considering that $\alpha_1;\ldots;\alpha_i \mathbin{\underline{\blacktriangleright}} \delta'$, the proof of Proposition II.2 can be followed in showing that $\{\phi\}\not\models\mathsf{obs}(\alpha_1,1)\wedge\ldots\wedge\mathsf{obs}(\alpha_i,i)$ but $\Theta\cup\{\phi\}\models\mathsf{obs}(\alpha_1,1)\wedge\ldots\wedge\mathsf{obs}(\alpha_i,i)$. □

In the case of partial observation, a situation occurs with respect to actions possibly being missed which is similar to that occurring in the case of late observation, the main difference being that the observer not only takes into account that it might have missed initial actions, but that after perceiving some action it might also fail to see actions occurring later in the process of the agent's plan execution. Specifically, the observer admits the possibility that it will perceive a dilution of the complete sequence of observable actions generated by the agent's plan, as given by $\underline{\odot}$, so that the condition of partial observation is logically formalized as follows.

Definition II.13 (partial observability clauses $P_{\mathscr{R}}$)**.** *Let $\mathscr{R}$ be a set of* MyAPL *rules, and OS the function of Definition II.7. The set of logical expressions $P_{\mathscr{R}}\subseteq\mathscr{L}_1$ pertaining to late observability of the behavior of an agent with the rules $\mathscr{R}$ is then defined as follows.*

$$\begin{aligned}&\forall(n:\gamma <\!\!-\beta \mid \pi)\in\mathscr{R}\ \forall\delta\in OS(\pi)\ \forall\alpha_1;\ldots;\alpha_i\in\mathscr{L}_\Delta:\\ &\quad (\alpha_1;\ldots;\alpha_i \mathbin{\underline{\odot}} \delta)\implies\\ &\qquad (\mathsf{rule}(n)\wedge\mathsf{seq}(\iota_\delta(\alpha_1;\ldots;\alpha_i)))\rightarrow(\mathsf{obs}(\alpha_1,1)\wedge\ldots\wedge\mathsf{obs}(\alpha_i,i))\quad\in P_{\mathscr{R}}\end{aligned}$$

Again, it can be proven that this formalization captures the intended condition by incorporating its characterizing structural relation, which in the case of partial observation is straightforward because its definition is directly based on that relation.

Proposition II.4. *Let $P_{\mathscr{R}}$ be the smallest set closed under Definition II.13 with respect to some set of* MyAPL *rules $\mathscr{R}$, and let $\Theta=\mathsf{Th}(P_{\mathscr{R}})$. It then holds that*

$$\begin{aligned}&\forall(n:\gamma <\!\!-\beta \mid \pi)\in\mathscr{R}\ \forall\delta\in OS(\pi)\ \forall\alpha_1;\ldots;\alpha_i\in\mathscr{L}_\Delta:\\ &\quad \alpha_1;\ldots;\alpha_i \mathbin{\underline{\odot}} \delta \implies \exists\phi\in\mathscr{L}_1:\ (\{\phi\}\not\models \mathsf{obs}(\alpha_1,1)\wedge\ldots\wedge\mathsf{obs}(\alpha_i,i)\\ &\qquad \&\ \ \Theta\cup\{\phi\}\models \mathsf{obs}(\alpha_1,1)\wedge\ldots\wedge\mathsf{obs}(\alpha_i,i))\end{aligned}$$

Proof. Take any $(n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}$ and let $\delta \in OS(\pi)$. Observe that if $\alpha_1;\ldots;\alpha_i \trianglelefteq \delta$ then from Definition II.12, given $\phi = (\mathsf{rule}(n) \wedge \mathsf{seq}(\iota_\delta(\alpha_1;\ldots;\alpha_i))$, directly follows that $\phi \rightarrow (\mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)) \in \Theta$, $\{\phi\} \not\models \mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)$, and $\Theta \cup \{\phi\} \models \mathsf{obs}(\alpha_1,1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i,i)$. □

3.4.2 Reasoning About Unobservables

Apart from stating something about observables (i.e. primitive actions), a logical theory describing the behavior of an agent on grounds of its rules can state something about *unobservables*, like the agent's beliefs and goals. To this extent the functions τ_b and τ_g are defined that, respectively, translate belief and goal queries, occurring as preconditions in rules of a MYAPL program, to the language $\mathscr{L}_1$.

Definition II.14 (MYAPL rule precondition translation functions τ_b and τ_g)**.** *Let* Atom *be the set of* MYAPL *atoms, and* $\mathsf{Lit} = \{p, \neg p \mid p \in \mathsf{Atom}\}$. *Let* ϕ *represent any* MYAPL ⟨*query*⟩ *element, and let* $\psi \in \mathsf{Lit}$. *The translation functions* τ_b, τ_g *that translate rule preconditions into the language* $\mathscr{L}_1$ *are then as follows.*

$$\begin{aligned}
\tau_b(\psi) &= \mathsf{belief}(\psi) & \tau_g(\psi) &= \mathsf{goal}(\psi)\\
\tau_b(\phi \text{ or } \phi') &= (\tau_b(\phi) \vee \tau_b(\phi')) & \tau_g(\phi \text{ or } \phi') &= (\tau_g(\phi) \vee \tau_g(\phi'))\\
\tau_b(\phi \text{ and } \phi') &= (\tau_b(\phi) \wedge \tau_b(\phi')) & \tau_g(\phi \text{ and } \phi') &= (\tau_g(\phi) \wedge \tau_g(\phi'))
\end{aligned}$$

Note that in Definition II.14 negation is 'lost in translation', given that $\mathsf{belief}(p)$ has no inherent relation to $\mathsf{belief}(\neg p)$ in the classical language $\mathscr{L}_1$, and likewise for $\mathsf{goal}/1$. The translation is thus perhaps somewhat unsatisfactory, but it is acceptable given the interpretation of the predicates $\mathsf{belief}/1$ and $\mathsf{goal}/1$ stated in Section 3.4. Also, in later chapters more satisfactory interpretations of MYAPL negation are given.

Knowledge of the rules applied by the agent provides knowledge of the agent's mental state, in the sense that if it is presumed that the agent has applied a particular rule then the preconditions of this rule must have been satisfied with respect to its mental state. This notion is captured formally in the following clauses.

Definition II.15 (mentalistic precondition clauses $M_{\mathscr{R}}$)**.** *Let* $\mathscr{R}$ *be a set of* MYAPL *rules, and* $M_{\mathscr{R}} \subseteq \mathscr{L}_1$ *a set of logical expressions reflecting inference of the mental state of an agent on grounds of those rules.*

$$\begin{aligned}
&\forall (n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R} :\\
&\qquad \mathsf{rule}(n) \rightarrow (\tau_g(\gamma) \wedge \tau_b(\beta)) \quad \in M_{\mathscr{R}}
\end{aligned}$$

Apart from making claims about an agent's mental state, the observer can also reason about the agent's decision-making process. Specifically, the assumption is here made explicit that the agent does not apply different rules simultaneously, as is the assumption that the agent does not simultaneously follow different plan branches and, accordingly, that the observer expects to see only a single sequence as a result thereof.

Definition II.16 (deliberation clauses $S_{\mathscr{R}}$)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules. The set of expressions $S_{\mathscr{R}} \subseteq \mathscr{L}_1$ then reflects the assumption that a single rule is selected and a single action sequence is expected to be seen, and is defined as follows.*

$$\begin{aligned}
&\forall(n:\gamma \mathrel{<\!\!-} \beta \mid \pi),(m:\gamma' \mathrel{<\!\!-} \beta' \mid \pi') \in \mathscr{R} \forall \delta \in OS(\pi) \forall \delta' \in OS(\pi') \forall \delta'', \delta''' \in \mathscr{L}_{\Delta}: \\
&\quad ((n \neq m) \implies \neg(\mathsf{rule}(m) \wedge \mathsf{rule}(n)) \quad \in S_{\mathscr{R}}) \quad \& \\
&\quad ((\delta'' \underline{\odot} \delta \;\&\; \delta''' \underline{\odot} \delta' \;\&\; \iota_{\delta}(\delta'') = x \;\&\; \iota_{\delta}(\delta''') = y \;\&\; (x \neq y)) \implies \\
&\qquad \neg(\mathsf{seq}(x) \wedge \mathsf{seq}(y)) \quad \in S_{\mathscr{R}})
\end{aligned}$$

II.3

Thus, by means of the clauses explicated above, the observer is enabled to reason about the observable behavior of an agent whose rules it knows (observability clauses $C_{\mathscr{R}}$, $L_{\mathscr{R}}$, and $P_{\mathscr{R}}$) as well as the mental state which it could have had ($M_{\mathscr{R}}$). Furthermore, by means of clauses stating properties of the agent's deliberation process ($S_{\mathscr{R}}$), the observer presumes that no two rules are applied simultaneously, and also expects to see only the actions of a single plan branch. This latter fact is realized by the right-hand part of the top-level meta-conjunction '&' in Definition II.16, expressing that for the plans of *any* two rules it holds that distinct dilutions of those plans (and hence also prefixes or substrings) cannot both be the sequence the observer expects to (eventually) see as the outcome of execution of those plans. This latter clause could have been expressed more succinctly by stating it more generally in reference to the predicate seq/1 and the natural numbers, but it is considered better practice to illustrate how this clause can be grounded in the set of rules that the observed agent is known to have at its disposition. The aforementioned technicalities have paved the way for formulating the background-part of an abductive theory about BDI-based agents' observed behavior, to complete the abductive theory in the next section the abducibles and observables are considered.

3.5 Abducibles and Observables

It was stated in Section 3.1 that an abductive theory is taken to consist of a background theory, along with pre-specified abducibles and observables. The formal foundation for the background theory of an observer reasoning about an observed agent on grounds of its rules was laid in Section 3.4; this section deals with the abducibles and observables.

3.5.1 Abducibles

Computational approaches to abduction typically employ a restricted set of *abducible* elements, which are elements that have the status of candidate abductive explanations. As noted by Flach & Kakas, this restriction is not imposed incidentally or solely for computational reasons, but has the deeper representational reason of reflecting the comprehensiveness of the accompanying background theory in regard to knowledge about some domain (2000, p. 19). It is assumed in this chapter that possible explanations for observed behavior are formulated in terms of the rule an agent has applied and the particular action sequence the observer expects, as follows.

Definition II.17 (abducibles $\mathscr{A}_{\mathscr{R}}$)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules. The set of abducibles $\mathscr{A}_{\mathscr{R}}$ is then as follows.*

$$\mathscr{A}_{\mathscr{R}} = \{\mathsf{rule}(n) \wedge \mathsf{seq}(i) \mid \exists (n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R} \exists \delta, \delta' \in \mathscr{L}_{\Delta} : (\delta \in OS(\pi)\ \&\ (\delta' \underline{\odot}\ \delta)\ \&\ \iota_{\delta}(\delta') = i)\}$$

As the above definition shows, the set of abducibles $\mathscr{A}_{\mathscr{R}}$ is grounded in a particular set of rules $\mathscr{R}$, employing the relation $\underline{\odot}$ to encompass $\underline{\blacktriangleright}$ and $\underline{\blacktriangledown}$, as in Definition II.16. Because elements of $\mathscr{A}_{\mathscr{R}}$ consist of a conjunction of a single instance of rule/1 and a single instance of seq/1, it follows that by choosing single abducibles as candidate explanations the constraining assumptions put forward in Definition II.16 are respected.

3.5.2 Observables

It was stated at the end of Section 3.1 that the primitive actions of an agent are considered to be observable *in principle*. That is to say; primitive actions are the aspect of an agent's behavior which can be perceived by the observer, whereas other aspects of its behavior cannot. Recall that the language $\mathscr{L}_1$ contains the predicate obs/2 pertaining to the observability of actions (cf. Definition II.8), noting that this predicate accepts elements from a subset, consisting of all single-action percepts, of the language $\mathscr{L}_{\Delta}$ that describes observable action sequences, so that it captures the intuitive interpretation of the term 'observable'. The predicate describes specifically which action is observable at which position, and as such gives more information than stating simply that primitive actions are observable; note, however, that because observation is incremental it is the case that actions which are in principle observable at some position have not necessarily been observed at the moment of explanation, so that — in contrast to typical approaches to abduction — it seems inappropriate to represent actual observations as facts for which the theory accounts (i.e. instances of the predicate obs/2).

In context of logical abduction the term 'observable' also has a technical meaning, referring to possible explananda. To represent actual observations (explananda) the predicate seen/2 is employed, which states that some action has actually been observed at some position. Given the assumption that the observer has no doubt regarding which particular actions it has observed, an observation takes the form of conjoined instances of the predicate seen/2. Because the assumption is maintained that it is not known by the observer whether it has failed to see particular actions of the agent, any actual observation is required to have consecutively numbered instances of seen/2. Also, it is assumed that the observer perceives only a single action 'simultaneously', so that no two different actions are observed at the same instant. The set of observables (in the technical sense) is therefore defined as follows.

Definition II.18 (observables $\Omega_{\mathscr{R}}$)**.** *Let $P_{\mathscr{R}}$ be the set of expressions concerning the partial observability of the actions of an agent with the set of* MYAPL *rules; i.e. the smallest closed*

under Definition II.13. The corresponding set of observables $\Omega_{\mathscr{R}} \subseteq \mathscr{L}_1$ *is then as follows.*

$$\forall(\phi \rightarrow (\mathsf{obs}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{obs}(\alpha_n, n)) \in P_{\mathscr{R}} :$$
$$\mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n) \in \Omega_{\mathscr{R}}$$

Note that because of Proposition II.1, which states that the dilution relation $\underline{\odot}$ that characterizes partial observation subsumes those of complete and late observation, it holds that observables in the set $\Omega_{\mathscr{R}}$, as defined in Definition II.18 in relation to $P_{\mathscr{R}}$, include those arising in conditions of complete and late observation (as seen also in Definitions II.16 and II.17). Furthermore note that, in principle, the language of $\mathscr{L}_1$ allows for expressing uncertainty in observation (i.e. disjoined instances of $\mathsf{seen}/2$ that have a distinct first but identical second argument) as well as the observation of concurrent actions (i.e. conjoined instances of $\mathsf{seen}/2$ that have a distinct first but identical second argument). Those topics, however, are considered to be outside the scope of this dissertation.

3.6 Explaining Observed Behavior

Given the material of preceding sections, the *abductive theory* of an observer reasoning about the behavior of an agent on grounds of this agent's set of rules can be defined using the elements put forward in preceding sections. It is noteworthy in this respect that the observables of Definition II.18 consist of conjoined instances of the predicate $\mathsf{seen}/2$, about which a theory regarding action observability, instantiated on grounds of Definitions II.9, II.12, or II.13, makes no claims! This is somewhat strange in regard to abductive explanation as defined in Section 1, where it is required for observations to be explained that they follow from the background theory, given the admission of some hypothesis; cf. Section 3.5.2. In order to remedy this, an 'observability operator' is defined which translates observables (i.e. elements of a set $\Omega_{\mathscr{R}}$ defined as in Definition II.18) to something which the background theory can account for.

Definition II.19 (observability operator)**.** *The observability operator* $\mathsf{o} : \mathscr{L}_1 \longrightarrow \mathscr{L}_1$ *is defined as follows.*

$$\mathsf{o}(\mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n)) = \mathsf{obs}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{obs}(\alpha_n, n)$$

As it turns out, this operator does nothing more than 'reverse' the result of Definition II.18, which extracted conjunctions of $\mathsf{seen}/2$ on grounds of conjunctions of $\mathsf{obs}/2$ implicated in $P_{\mathscr{R}}$! Thus, its existence may seem superfluous, and it could be suggested that, instead of employing the predicate $\mathsf{seen}/2$, observations should simply consist of instances of $\mathsf{obs}/2$. Such an approach, as stated earlier, is undesirable in our estimation, because the meaning of $\mathsf{obs}(\alpha, n)$ is defined as "action α is observable as the n'th action", whereas that of $\mathsf{seen}(\alpha, n)$ is defined as "action α is seen as the n'th action". The latter is clearly more appropriate for formalizing the actual observation of actions, so that our choice is to employ the observability operator, with the underlying rationale that any action which is actually seen must also be observable.[5]

[5] This issue reflects the difficulty of using classical logic to model dynamics; something which modal logic is better suited for, and one of the reasons why PDL is employed in later chapters.

Given the above, the abductive theory used by an observer in reasoning about the behavior of a MyAPL agent with known rules can be defined. It is hereby assumed that the observer, in fact, maintains *three distinct background theories*, each of which represents a different perceptory condition. Those background theories are defined as follows, where the subscripts C, L, P denote the perceptory conditions of complete, late, and partial observation, respectively, and a fixed underlying set of rules $\mathscr{R}$ is assumed so that its mention is omitted in regard to those theories in order to avoid clutter; note that the operator 'Th' is employed to warrant the term 'theories'.

Definition II.20 (background theories $\Theta_C, \Theta_L, \Theta_P$)**.** *Let $\mathscr{R}$ be a set of* MyAPL *rules, $C_{\mathscr{R}}$ the smallest set closed under Definition II.9, $L_{\mathscr{R}}$ the smallest set closed under Definition II.12, $P_{\mathscr{R}}$ the smallest set closed under Definition II.13, $M_{\mathscr{R}}$ the smallest set closed under Definition II.15, and $S_{\mathscr{R}}$ the smallest set closed under Definition II.16, all with respect to $\mathscr{R}$. The background theories Θ_C, Θ_L, and Θ_P are then as follows.*

$$\Theta_C = \mathsf{Th}(C_{\mathscr{R}} \cup M_{\mathscr{R}} \cup S_{\mathscr{R}})$$
$$\Theta_L = \mathsf{Th}(L_{\mathscr{R}} \cup M_{\mathscr{R}} \cup S_{\mathscr{R}})$$
$$\Theta_P = \mathsf{Th}(P_{\mathscr{R}} \cup M_{\mathscr{R}} \cup S_{\mathscr{R}})$$

Given the above, the observer's abductive theory is as follows.

Definition II.21 (abductive theory $\Lambda_{\mathscr{R}}$)**.** *Let $\mathscr{R}$ be a set of* MyAPL *rules from which the background theories Θ_C, Θ_L, and Θ_P are derived, as defined in Definition II.20, and let $\mathscr{A}_{\mathscr{R}}$ be the set of abducibles derived from $\mathscr{R}$ according to Definition II.17. The abductive theory $\Lambda_{\mathscr{R}}$ based on $\mathscr{R}$ is then as follows.*

$$\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$$

Given a sequence of actions the agent is observed to perform, the observer can attempt to explain this sequence on grounds of its abductive theory. To this extent the ternary explanatory relations $\mathrel{|\!\approx_C}$, $\mathrel{|\!\approx_L}$, and $\mathrel{|\!\approx_P}$ are defined below, in Definition II.22. It should be noted that the observables are not explicitly part of the abductive theory; instead it is simply assumed that any explanandum stems from a set of observables $\Omega_{\mathscr{R}}$, derived from the logical relations underlying the background theories as in Definition II.18.

Definition II.22 (abductive action explanation)**.** *Let $\mathscr{R}$ be a set of* MyAPL *rules, $\Lambda_{\mathscr{R}}$ an abductive theory as defined in Definition II.21, $\omega \in \Omega_{\mathscr{R}}$ an explanandum, and $\phi \in \mathscr{A}_{\mathscr{R}}$ an explanans. The explanatory relations $\mathrel{|\!\approx_C}$, $\mathrel{|\!\approx_L}$, and $\mathrel{|\!\approx_P}$ under complete, late, and partial observation, respectively, are then as follows.*

$$\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx_C} \phi \quad \Longleftrightarrow \quad \Theta_C \cup \{\phi\} \models \mathsf{o}(\omega)$$
$$\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx_L} \phi \quad \Longleftrightarrow \quad \Theta_L \cup \{\phi\} \models \mathsf{o}(\omega)$$
$$\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx_P} \phi \quad \Longleftrightarrow \quad \Theta_P \cup \{\phi\} \models \mathsf{o}(\omega)$$

Note that the symbols '$\mid\!\!\approx_C$', '$\mid\!\!\approx_L$', and '$\mid\!\!\approx_P$' are syntactically similar to '$\mid\!\!\approx$' as used for abductive explanation in Definition II.1. The following proposition shows that this similarity extends beyond the surface, in the sense that any explanation obtained on grounds of the relations defined in Definition II.22 is a classical abductive explanation.

Theorem II.1. *Let $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ be the abductive theory based on some set of* MYAPL *rules $\mathscr{R}$ and $\Omega_{\mathscr{R}}$ the corresponding observables. It is then the case that*

$$\forall X \in \{C, L, P\}\ \forall \phi \in \mathscr{A}_{\mathscr{R}}\ \forall \omega \in \Omega_{\mathscr{R}} :$$
$$\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx_X \phi \qquad \Longleftrightarrow \qquad \Theta_X, \mathsf{o}(\omega) \mid\!\!\approx \{\phi\}$$

Proof. Choose any $X \in \{C, L, P\}$. From the definitions of $\mid\!\!\approx_X$ it follows that $\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx_X \phi$ if and only if $\Theta_X \cup \{\phi\} \models \mathsf{o}(\omega)$. From the definition of $\mid\!\!\approx$ it follows that one requirement for $\Theta_X, \mathsf{o}(\omega) \mid\!\!\approx \{\phi\}$ is that $\Theta_X \cup \{\phi\} \models \mathsf{o}(\omega)$, which is thus already fulfilled on grounds of the definition of $\mid\!\!\approx_X$. Further requirements are that $\Theta_X \not\models \mathsf{o}(\omega)$, $\{\phi\} \not\models \mathsf{o}(\omega)$, and $\Theta_X \cup \{\phi\} \not\models \bot$. Thus, if the last three clauses of the definition of $\mid\!\!\approx$ hold for any $\phi \in \mathscr{A}_{\mathscr{R}}$ and $\omega \in \Omega_{\mathscr{R}}$ then the proof follows. To see that these clauses indeed hold, first take any $\omega \in \Omega_{\mathscr{R}}$, such that $\omega = \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n)$ and $\mathsf{o}(\omega)$ yields the corresponding conjunction of instances of obs/2. Observe that any theory Θ_X is based solely on implications of the form $\mathsf{rule}(n) \rightarrow \psi$ and $\mathsf{rule}(n) \wedge \mathsf{seq}(x) \rightarrow \psi$ for particular $\psi \in \mathscr{L}_1$, and facts of the form $\neg(\mathsf{rule}(m) \wedge \mathsf{rule}(n))$ and $\neg(\mathsf{seq}(x) \wedge \mathsf{seq}(y))$. Therefore, indeed, $\Theta_X \not\models \mathsf{o}(\omega)$. Take any $\phi \in \mathscr{A}_{\mathscr{R}}$, such that $\phi = \mathsf{rule}(n) \wedge \mathsf{seq}(i)$. Clearly, $\{\phi\} \not\models \mathsf{o}(\omega)$. Also, since ϕ contains only a single instance of both rule/1 and seq/1, there is no contradiction on grounds of facts of the form $\neg(\mathsf{rule}(m) \wedge \mathsf{rule}(n))$ and $\neg(\mathsf{seq}(x) \wedge \mathsf{seq}(y))$. Furthermore, for any $(\mathsf{rule}(n) \rightarrow \psi), (\mathsf{rule}(n) \wedge \mathsf{seq}(x) \rightarrow \psi) \in \Theta_X$ holds that ψ consists either of conjoined instances of obs/2 or descriptions of rule preconditions derived on grounds of Definition II.15 (i.e. conjunctions or disjunction of belief/1 and goal/1). In either case, $\Theta_X \cup \{\phi\} \not\models \bot$. □

3.6.1 Abductive Extensions

Earlier in this chapter the notion of 'extensions' was introduced informally, and it will be defined formally in the current section. As mentioned in Section 1.2, extensions can be seen as completions of an incomplete logical theory with defeasible information that is justified, in the sense that it can be inferred on grounds of some nonmonotonic relation. In our case that relation is abductive explanation, and an *abductive extension* is a completion of a background theory with some fact that can be abductively inferred as explanation for observed actions. Because there may exist multiple abductive extensions to a background theory given some observation and it is convenient to refer to those extensions as a whole, an operator is defined which, given some observation, generates all abductive extensions to a theory.

Definition II.23 (abductive extensions)**.** *Let $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ be the abductive theory based on some set of* MYAPL *rules $\mathscr{R}$, and $\omega \in \Omega_{\mathscr{R}}$ an observation. The functions $\Gamma^C_{\mathscr{R}}$,*

$\Gamma^L_{\mathscr{R}}$, *and* $\Gamma^P_{\mathscr{R}}$ *then denote the abductive extensions of their respective background theories on grounds of that observation, and are defined as follows.*

$$\Gamma^C_{\mathscr{R}}(\omega) = \{\mathsf{Th}(\Theta_C \cup \{\phi\}) \mid \Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx_C} \phi\}$$
$$\Gamma^L_{\mathscr{R}}(\omega) = \{\mathsf{Th}(\Theta_L \cup \{\phi\}) \mid \Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx_L} \phi\}$$
$$\Gamma^P_{\mathscr{R}}(\omega) = \{\mathsf{Th}(\Theta_P \cup \{\phi\}) \mid \Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx_P} \phi\}$$

It should be observed that the background theory itself is not an extension, although it is of course a subset of each extension. Thus, the operator $\Gamma_{\mathscr{R}}$ emphasizes the explanatory role of the abductive theory, which is non-existent if there exists no abductive extension based on some observation. The notion of extensions furthermore allows for characterizing properties of consequence relations. It was stated in Section 1.2 that a fact can be credulously inferred if it is entailed by some extension, and it can be skeptically inferred if it is entailed by all extensions. Formally, this is captured by two distinct consequence relations, each of which can in turn be defined per perceptory condition.

Definition II.24 (credulous and skeptical abduction)**.** *Let* $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ *be the abductive theory based on some set of* MyAPL *rules* $\mathscr{R}$*,* $\omega \in \Omega_{\mathscr{R}}$ *an observation, and* $\Gamma^C_{\mathscr{R}}$, $\Gamma^L_{\mathscr{R}}$, *and* $\Gamma^P_{\mathscr{R}}$ *the functions as defined in Definition II.23. Given* $X \in \{C, L, P\}$*, credulous abduction* $\mathrel{|\!\approx^{\mathsf{cr}}_X}$ *and skeptical abduction* $\mathrel{|\!\approx^{\mathsf{sk}}_X}$ *are defined as follows.*

$$\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{cr}}_X} \phi \quad \Longleftrightarrow \quad \exists \Phi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi \models \phi)$$
$$\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{sk}}_X} \phi \quad \Longleftrightarrow \quad \Gamma^X_{\mathscr{R}}(\omega) \neq \emptyset \text{ and } \forall \Psi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Psi \models \phi)$$

Note that it is explicitly required in the definition of skeptical abduction that the set of abductive extensions is not empty, because otherwise (by nature of the universal quantifier) any fact could be skeptically concluded in the absence of abductive explanations! From this it follows that any fact which can be skeptically concluded can also be credulously concluded, seen as follows.

Proposition II.5. *Let* $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ *be the abductive theory based on some set of* MyAPL *rules* $\mathscr{R}$*, and* $\omega \in \Omega_{\mathscr{R}}$ *an observation. Given* $X \in \{C, L, P\}$*,*

$$\forall \phi \in \mathscr{L}_1 : \quad \Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{sk}}_X} \phi \quad \Longrightarrow \quad \Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{cr}}_X} \phi$$

Proof. From Definition II.24 it follows that if $\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{sk}}_X} \phi$ then $\Gamma^X_{\mathscr{R}}(\omega) \neq \emptyset$ such that $\exists \Phi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi \models \phi)$, from which follows $\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{cr}}_X} \phi$. □

Also of note is that among the conclusions which can be skeptically (and thus also credulously) concluded are those which followed from the background theory alone.

Proposition II.6. *Let $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ be the abductive theory based on some set of* MYAPL *rules $\mathscr{R}$, and $\Omega_{\mathscr{R}}$ the corresponding observables.*

$$\forall X \in \{C, L, P\}\ \forall \omega \in \Omega_{\mathscr{R}}\ \forall \phi \in \mathscr{L}_1 :$$
$$(\Theta_X \models \phi \quad \& \quad \exists \Phi \in \Gamma^X_{\mathscr{R}}(\omega)) \implies (\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{sk}}_X \phi)$$

Proof. Consider any $\Gamma^X_{\mathscr{R}}(\omega)$ such that $\Gamma^X_{\mathscr{R}}(\omega) \neq \emptyset$. Furthermore, assume that $\Theta_X \models \phi$ and observe that from Definition II.23 follows $\forall \Phi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Theta_X \subset \Phi)$, such that $\forall \Phi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi \models \phi)$. This means that the requirements for $\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{sk}}_X \phi$ are fulfilled. □

Corollary II.1. *Let $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ be the abductive theory based on some set of* MYAPL *rules $\mathscr{R}$, and $\Omega_{\mathscr{R}}$ the corresponding observables.*

$$\forall X \in \{C, L, P\}\ \forall \omega \in \Omega_{\mathscr{R}}\ \forall \phi \in \mathscr{L}_1 :$$
$$(\Theta_X \models \phi \quad \& \quad \exists \Phi \in \Gamma^X_{\mathscr{R}}(\omega)) \implies (\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{cr}}_X \phi)$$

Proof. Straightforward, from Propositions II.5 and II.6. □

3.7 Example

In this section an elaborate example is presented in order to illustrate the formal notions put forward in the preceding sections. This example is set in an environment that could be encountered in typical a role-playing game (RPG), or in a life-simulation game such as THE SIMS; namely, in a room filled with objects, enabling a rich variety of behavior. Because the previous sections focus on observation of a single agent by an observer, the scenario which unfolds in the room is of a similar nature. It is assumed that the observer is a further unspecified entity that perceives the agent's actions and attempts to explain them, without interfering with its behavior.[6] The agent is stated to have the MYAPL rules shown in Listing II.2, which are known to the observer.

Note that for simplicity the rules in this example contain no conditional composition but only actions in sequence. The identifiers of the various action sequences are as follows, where for conciseness actions are referred to with the following variables.

$\alpha_1 = \mathsf{goto\_shelf}$ $\quad$ $\alpha_6 = \mathsf{goto\_table}$
$\alpha_2 = \mathsf{pickup\_paper}$ $\quad$ $\alpha_7 = \mathsf{squash\_bug}$
$\alpha_3 = \mathsf{goto\_chair}$ $\quad$ $\alpha_8 = \mathsf{pickup\_vase}$
$\alpha_4 = \mathsf{sit}$ $\quad$ $\alpha_9 = \mathsf{arrange\_flowers}$
$\alpha_5 = \mathsf{read}$

Because there are only nine actions in this example, a straightforward way to define the function ι_δ, which assigns identifiers to action sequences, is to take the number formed by

[6] The observer can be imagined to stand outside the room, looking in through a window.

putting the action variables' subscript numerals in sequence as identifier for that sequence. For example, '163' refers to 'goto_shelf; goto_table; goto_chair'. For further conciseness, the following variables are used in reference to atomic propositions.

$$\begin{aligned} p_1 &= \mathsf{paper\_read} & p_5 &= \mathsf{flowers\_arranged} \\ p_2 &= \mathsf{paper\_on\_shelf} & p_6 &= \mathsf{flowers\_on\_table} \\ p_3 &= \mathsf{bug\_squashed} & p_7 &= \mathsf{vase\_on\_shelf} \\ p_4 &= \mathsf{bug\_on\_table} \end{aligned}$$

The component Θ_C in the observer's abductive theory $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ is then as follows; i.e. the theory derived using the operator 'Th' on the set yielded by Definition II.9 with respect to the set of rules $\mathscr{R}$ in Listing II.2.

$$\begin{aligned} \Theta_C = \mathsf{Th}(\{&\mathsf{rule}(1) \rightarrow (\mathsf{goal}(p_1) \wedge \mathsf{belief}(p_2)), \\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(12345)) \rightarrow (\mathsf{obs}(\alpha_1, 1) \wedge \mathsf{obs}(\alpha_2, 2) \wedge \mathsf{obs}(\alpha_3, 3) \wedge \mathsf{obs}(\alpha_4, 4) \wedge \mathsf{obs}(\alpha_5, 5)), \\ &\mathsf{rule}(2) \rightarrow (\mathsf{goal}(p_3) \wedge (\mathsf{belief}(p_4) \wedge \mathsf{belief}(p_2))), \\ &(\mathsf{rule}(2) \wedge \mathsf{seq}(1267)) \rightarrow (\mathsf{obs}(\alpha_1, 1) \wedge \mathsf{obs}(\alpha_2, 2) \wedge \mathsf{obs}(\alpha_6, 3) \wedge \mathsf{obs}(\alpha_7, 4)), \\ &\mathsf{rule}(3) \rightarrow (\mathsf{goal}(p_5) \wedge (\mathsf{belief}(p_6) \wedge \mathsf{belief}(p_7))), \\ &(\mathsf{rule}(3) \wedge \mathsf{seq}(1869)) \rightarrow (\mathsf{obs}(\alpha_1, 1) \wedge \mathsf{obs}(\alpha_8, 2) \wedge \mathsf{obs}(\alpha_6, 3) \wedge \mathsf{obs}(\alpha_9, 4)), \\ &\neg(\mathsf{rule}(1) \wedge \mathsf{rule}(2)), \neg(\mathsf{rule}(1) \wedge \mathsf{rule}(3)), \neg(\mathsf{rule}(2) \wedge \mathsf{rule}(3)), \\ &\neg(\mathsf{seq}(12345) \wedge \mathsf{seq}(1267)), \neg(\mathsf{seq}(12345) \wedge \mathsf{seq}(1869)), \neg(\mathsf{seq}(1267) \wedge \mathsf{seq}(1869))\}) \end{aligned}$$

Assuming the systematic relation between the argument of the predicate seq/1 in the precondition of implications and instances of obs/2 in the postcondition to be clear, for conciseness the postcondition is replaced with '…' in the following explication of Θ_L and Θ_P. Furthermore, the clauses $\neg(\mathsf{seq}(x) \wedge \mathsf{seq}(y))$ resulting from Definition II.16 are not exhaustively mentioned, and $C_{\mathscr{R}}$ is taken to be the set of logical expressions given as argument to the operator Th in the above definition of Θ_C, so that it can be reused in defining the component Θ_L of the observer's abductive theory; note that $C_{\mathscr{R}}$ already comprises the clauses

```
1:  paper_read        <-  paper_on_shelf  |
    { goto_shelf; pickup_paper; goto_chair; sit; read }

2:  bug_squashed      <-  bug_on_table and paper_on_shelf     |
    { goto_shelf; pickup_paper; goto_table; squash_bug }

3:  flowers_arranged  <-  flowers_on_table and vase_on_shelf  |
    { goto_shelf; pickup_vase; goto_table; arrange_flowers }
```

Listing II.2

of $M_{\mathscr{R}}$ (cf. Definition II.15).

$$\begin{aligned}\Theta_L = \mathsf{Th}(C_{\mathscr{R}} \cup \{&(\mathsf{rule}(1) \wedge \mathsf{seq}(2345)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(345)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(45)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(5)) \rightarrow (\ldots),\\ &(\mathsf{rule}(2) \wedge \mathsf{seq}(267)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(67)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(7)) \rightarrow (\ldots),\\ &(\mathsf{rule}(3) \wedge \mathsf{seq}(869)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(69)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(9)) \rightarrow (\ldots),\\ &\neg(\mathsf{seq}(2345) \wedge \mathsf{seq}(345)), \ldots\})\end{aligned}$$

II.3

Taking $L_{\mathscr{R}}$ to be the set which in union with $C_{\mathscr{R}}$ forms the basis for the theory Θ_L above, the component Θ_P is defined as follows, where again the clauses resulting from Definition II.16 are not exhaustively mentioned.

$$\begin{aligned}\Theta_P = \mathsf{Th}(C_{\mathscr{R}} \cup L_{\mathscr{R}} \cup \{&(\mathsf{rule}(1) \wedge \mathsf{seq}(1)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(2)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(3)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(4)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(12)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(13)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(14)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(15)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(23)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(24)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(25)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(34)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(35)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(123)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(124)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(125)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(134)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(135)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(145)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(234)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(235)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(245)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(1234)) \rightarrow (\ldots),\\ &(\mathsf{rule}(1) \wedge \mathsf{seq}(1235)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(1245)) \rightarrow (\ldots), (\mathsf{rule}(1) \wedge \mathsf{seq}(1345)) \rightarrow (\ldots),\\ &(\mathsf{rule}(2) \wedge \mathsf{seq}(1)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(2)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(6)) \rightarrow (\ldots),\\ &(\mathsf{rule}(2) \wedge \mathsf{seq}(12)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(16)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(17)) \rightarrow (\ldots),\\ &(\mathsf{rule}(2) \wedge \mathsf{seq}(26)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(27)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(126)) \rightarrow (\ldots),\\ &(\mathsf{rule}(2) \wedge \mathsf{seq}(127)) \rightarrow (\ldots), (\mathsf{rule}(2) \wedge \mathsf{seq}(167)) \rightarrow (\ldots),\\ &(\mathsf{rule}(3) \wedge \mathsf{seq}(1)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(8)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(6)) \rightarrow (\ldots),\\ &(\mathsf{rule}(3) \wedge \mathsf{seq}(18)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(16)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(19)) \rightarrow (\ldots),\\ &(\mathsf{rule}(3) \wedge \mathsf{seq}(86)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(89)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(186)) \rightarrow (\ldots),\\ &(\mathsf{rule}(3) \wedge \mathsf{seq}(189)) \rightarrow (\ldots), (\mathsf{rule}(3) \wedge \mathsf{seq}(169)) \rightarrow (\ldots),\\ &\neg(\mathsf{seq}(1) \wedge \mathsf{seq}(2)), \ldots\})\end{aligned}$$

Finally, the set of abducibles $\mathscr{A}_{\mathscr{R}}$ derived from the set of MyAPL rules $\mathscr{R}$ listed in Listing II.2 can also be described as the set of all preconditions $\mathsf{rule}(n) \wedge \mathsf{seq}(i)$ in implications in $C_{\mathscr{R}} \cup L_{\mathscr{R}} \cup P_{\mathscr{R}}$, where $P_{\mathscr{R}}$ is the set which in union with $C_{\mathscr{R}}$ and $L_{\mathscr{R}}$ forms the basis for Θ_P, so that $\mathscr{A}_{\mathscr{R}}$ can be concisely summarized as follows.

$$\mathscr{A}_{\mathscr{R}} = \{\mathsf{rule}(1) \wedge \mathsf{seq}(12345), \ldots, \mathsf{rule}(3) \wedge \mathsf{seq}(169)\}$$

Assume that the virtual character with the aforementioned rules performs actions which are observed by an observer that employs the abductive theory defined above. Let the first

observed action be ‘goto_shelf’, and note that

$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\mathsf{goto\_shelf}, 1) \mid\!\approx^{\mathsf{cr}}_{C} \mathsf{rule}(1)$$
$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\mathsf{goto\_shelf}, 1) \mid\!\approx^{\mathsf{cr}}_{C} \mathsf{rule}(2)$$
$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\mathsf{goto\_shelf}, 1) \mid\!\approx^{\mathsf{cr}}_{C} \mathsf{rule}(3)$$

After all, for any abducible $\phi \in \{(\mathsf{rule}(1) \wedge (12345)), (\mathsf{rule}(2) \wedge (1267)), (\mathsf{rule}(3) \wedge (1869))\}$ holds $\Lambda_{\mathscr{R}}, \mathsf{seen}(\mathsf{goto\_shelf}, 1) \mid\!\approx \phi$, so that $\mathsf{Th}(\Theta_C \cup \{\phi\}) \in \Gamma^{C}_{\mathscr{R}}(\mathsf{seen}(\mathsf{goto\_shelf}, 1))$, which fulfills the requirement for credulous abductive explanation (cf. Definition II.24). Furthermore, note $\Lambda_{\mathscr{R}}, \mathsf{seen}(\mathsf{goto\_shelf}, 1) \mid\!\approx^{\mathsf{sk}}_{C} \mathsf{rule}(1) \vee \mathsf{rule}(2) \vee \mathsf{rule}(3)$ — i.e. the disjunction over three possible explanations can be skeptically inferred — and that explanation under late and partial observation yield identical results.

Now assume that the second observed action is ‘goto_table’ so that the corresponding observation is $\omega = \mathsf{seen}(\mathsf{goto\_shelf}, 1) \wedge \mathsf{seen}(\mathsf{goto\_table}, 2)$. Under the assumptions of complete or late observation, no abductive explanation can then be found for ω! However, under the condition of partial observation there exist four distinct and mutually exclusive explanations, which can be described as follows.

$\mathsf{rule}(2) \wedge \mathsf{seq}(16)$**:** The agent applied rule 2 and the observer expects to see the sequence ‘goto_shelf; goto_table’ with identifier ‘16’. In this case the observer has seen all the actions it expects to see.

$\mathsf{rule}(2) \wedge \mathsf{seq}(167)$**:** The agent applied rule 2 and the observer expects to see the sequence ‘goto_shelf; goto_table; squash_bug’ with identifier ‘167’. In this case the observer has seen two actions and expects to see a third.

$\mathsf{rule}(3) \wedge \mathsf{seq}(16)$**:** The agent applied rule 3 and the observer expects to see the sequence ‘goto_shelf; goto_table’ with identifier ‘16’. In this case the observer has seen all the actions it expects to see.

$\mathsf{rule}(3) \wedge \mathsf{seq}(169)$**:** The agent applied rule 3 and the observer expects to see the sequence ‘goto_shelf; goto_table; arrange_flowers’ with identifier ‘169’. In this case the observer has seen two actions and expects to see a third.

Apart from reasoning about behavior, the abductive approach sketched in this section also allows for defeasibly inferring the observed agent’s mental state at the time of rule application. Specifically, note that for any $X \in \{C, L, P\}$ the following holds.

$$\Theta_X \cup \{\mathsf{rule}(1)\} \models \mathsf{goal}(\mathsf{paper\_read}) \wedge \mathsf{belief}(\mathsf{paper\_on\_shelf})$$
$$\Theta_X \cup \{\mathsf{rule}(2)\} \models \mathsf{goal}(\mathsf{bug\_squashed}) \wedge (\mathsf{belief}(\mathsf{paper\_on\_shelf}) \wedge \mathsf{belief}(\mathsf{bug\_on\_table}))$$
$$\Theta_X \cup \{\mathsf{rule}(3)\} \models \mathsf{goal}(\mathsf{flowers\_arranged}) \wedge (\mathsf{belief}(\mathsf{flowers\_on\_table}) \wedge \mathsf{belief}(\mathsf{vase\_on\_shelf}))$$

Because the pure logical approach is somewhat unwieldy, the following chapter presents a representation of mental state inference that is more succinct.

4 A Functional Approach to Explanation of Behavior

Alternatively to modeling explanation of agents' behavior using an explicit abductive theory, one can employ a functional approach that is based on this theory but which abstracts from the underlying logical machinery. Such an approach has the benefit that it considers the subject from a different angle, and in doing so only takes particular aspects of the explanatory theory into account; also, it allows for study of certain properties in isolation. A downside is that this occurs at the expense of expressivity. This section presents such a functional approach to defeasibly inferring the mental state of MYAPL agents, given observed actions and knowledge of their rules, in context of the abductive theory presented in earlier sections. It may be noted that the distinction between 'logical' and 'functional' is not strict, technically speaking, because predicate logic involves the notion of functions. Nevertheless, given that in our case the functions are external to the logical formulation rather than being an inherent aspect of it, we consider it important to make the distinction.

II.4

4.1 Mental State Abduction

The rules of a MYAPL agent have the form '$n : \gamma <\!\!- \beta \mid \pi$', as defined in Definition II.3. By means of the structural relations on observables (Definition II.10) it is possible to map observed action sequences to goal/belief preconditions of rules, based on a relation with observable sequences that are extracted from corresponding plans. In earlier work (Sindlar et al., 2008) this approach was termed *mental state abduction*, based on the fact that if the agent's actual plan is taken to be selected on grounds of application of a rule from some set of known rules, then it is reasonable to (defeasibly) infer that the agent's mental state satisfied one of the goal/belief preconditions associated with plans that relate to the observed actions. More specifically, it was claimed that MYAPL rules of the form '$n : \gamma <\!\!- \beta \mid \pi$' can be treated as implications for the sake of abduction, with the following syllogism in mind.

Definition II.25 (syllogism of mental state abduction)**.** *Let $n : \gamma <\!\!- \beta \mid \pi$ be a* MYAPL *rule, and $\delta \in \mathscr{L}_\Delta$ some sequence of actions which an agent, known to have this rule at its disposition, is observed to perform.*

$$\frac{\textit{The action sequence } \delta \textit{ is observed, and is related to the plan } \pi. \qquad \textit{If the agent had goal } \gamma \textit{ and belief } \beta \textit{ then the agent selects the plan } \pi.}{\textit{The agent had goal } \gamma \textit{ and belief } \beta.}$$

This syllogism states that if some sequence of actions $\delta \in \mathscr{L}_\Delta$ is observed to be performed by the agent, and this sequence of actions is appropriately related to a plan π which occurs in one of the agent's rules of the form '$n : \gamma <\!\!- \beta \mid \pi$', then it can be *abduced* that the goal γ and belief β were satisfied by the agent's mental state at the time when it selected the plan π.[7] It is noteworthy that in this syllogism a clause occurs which constitutes the observed

[7] This syllogism illustrates the intuition underlying the functional approach, also putting our earlier work into perspective; a more elaborate formal interpretation is given in Section 4.2.

fact (the explanandum), as is common in abduction. Typically, however, this clause corresponds directly to the conclusion of the implication occurring in the syllogism, but here it does not. The reason for this discrepancy is that plans and their selection are taken not to be observable; only primitive actions are. Therefore the observation-clause consists of two conjuncts: the observed sequence of actions, and the fact that this sequence is related to some plan. The structural relations of Definition II.10 are possible relations to figure in a functional approach to mental state abduction, reflecting the perceptory conditions of complete, late, and partial observation, respectively. The corresponding explanatory functions then are as follows.

Definition II.26 (mental state abduction functions)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, let C, L, and P denote the perceptory conditions of complete, late, and partial observation, let τ_p and OS be the functions of Definitions II.5 and II.7, respectively, and let $\underline{\blacktriangleright}$, $\underline{\blacktriangledown}$, and $\underline{\odot}$ be the structural relations defined in Definition II.10.*

$$\mathsf{msa}_C(\delta,\mathscr{R}) = \{\,(\gamma,\beta) \mid \exists(\gamma <\!\!- \beta \mid \pi) \in \mathscr{R}\, \exists \delta' \in OS(\tau_p(\pi)) \colon (\delta \mathrel{\underline{\blacktriangleright}} \delta')\,\}$$
$$\mathsf{msa}_L(\delta,\mathscr{R}) = \{\,(\gamma,\beta) \mid \exists(\gamma <\!\!- \beta \mid \pi) \in \mathscr{R}\, \exists \delta' \in OS(\tau_p(\pi)) \colon (\delta \mathrel{\underline{\blacktriangledown}} \delta')\,\}$$
$$\mathsf{msa}_P(\delta,\mathscr{R}) = \{\,(\gamma,\beta) \mid \exists(\gamma <\!\!- \beta \mid \pi) \in \mathscr{R}\, \exists \delta' \in OS(\tau_p(\pi)) \colon (\delta \mathrel{\underline{\odot}} \delta')\,\}$$

Note that the above functions map to a representation which is not primarily logical: specifically, tuples of MYAPL rule preconditions. This is done by choice, because in the functional approach we wish to abstract from specific logical representations, leaving open the possibility of mapping those program elements to different representations. The functions defined in Definition II.26 capture the application of the mental state abduction syllogism under different perceptory conditions, mapping an observed sequence of actions to the set of those goal/belief pairs of which it can be defeasibly presumed that they were satisfied by the agent's mental state at the time of rule application. In the remainder of this dissertation the functional representation is often employed, because of its conciseness and convenient level of abstraction. The following brief example illustrates this aspect in regard to the example presented in Section 3.7.

$$\begin{aligned}(\delta = \mathsf{goto\_shelf}; \mathsf{goto\_table}) \implies\ & (\mathsf{msa}_C(\delta,\mathscr{R}) = \mathsf{msa}_L(\delta,\mathscr{R}) = \emptyset)\ \&\\ & \mathsf{msa}_P(\delta,\mathscr{R}) = \{(\mathsf{bug\_squashed}, \mathsf{bug\_on\_table\ and\ paper\_on\_shelf}),\\ & \qquad (\mathsf{flowers\_arranged}, \mathsf{flowers\_on\_table\ and\ vase\_on\_shelf})\}\end{aligned}$$

It should be evident that there is a correspondence between the logical and functional approaches in the sense that, given the above observed action sequence 'δ', explanations are only found under the assumption of partial observation, and that the goal/belief pairs in the output of the msa_P function have a logical counterpart (see the last paragraph of Section 3.7). In order to make this correspondence more precise, the next section focuses on clarifying the logical interpretation of those goal/belief pairs.

4.2 Interpretation of the Functional Approach

The functions of Definition II.26 have been introduced in earlier work (Sindlar et al., 2008) and are based (in spirit) on the abductive syllogism of Definition II.25. This syllogism, in its turn, is based on treating agent programming rules $n : \gamma <\!\!- \beta \mid \pi$ as implications of the kind "If the agent had goal γ and belief β then the agent selects the plan π". This treatment can be regarded as sufficiently accurate for the purpose of abduction; i.e. if one is only interested in defeasibly inferring the premise of the implication, based on knowledge regarding the conclusion. However, the implicative description is not correct from an 'operational' perspective, because the procedural interpretation of programming rules might be such that even when the premise of the implication is satisfied, the conclusion is not. Consider, for example, an agent with a goal base $\{p\}$ and a belief base $\{q, r\}$, and suppose it has the rules $\{1 : p <\!\!- q \mid \pi,\ 2 : p <\!\!- q \text{ and } r \mid \pi'\}$. If the rules of this agent are considered in relation to its belief and goal bases, then it follows that both rules are in principle applicable, seeing that their preconditions are satisfied. However, if the agent has a non-interleaving strategy (as we have assumed) then it would be reasonable for it to select at most one of those applicable plans, in which case the implicative description would strictly speaking not be correct. In relation to the example this could mean that it holds true that the agent selects π, whereas it does not select π', although the implicative description states that both plans are selected.

The fact that the implication in the syllogism of Definition II.25 does not correctly describe the behavior of the agent is not really problematic in relation to abductive explanation, because it then only matters which plans the agent *could* have selected; i.e., if the observed sequence of actions relates appropriately to a particular plan, then this could be the plan that the agent is executing and the preconditions of which can be abduced. But regardless of this claim, it can be deemed unsatisfactory for the functional approach to hinge on a syllogism that involves descriptions of behavior which are, strictly speaking, incorrect. In order to clarify those matters, the functional approach is put in relation to the abductive account of Section 3.

Theorem II.2 (skeptical interpretation of mental state abduction)**.** *Let* $\mathbf{R}$ *be the domain of* MYAPL *rules,* $\Lambda_{\mathscr{R}}$ *as defined in Definition II.21 with respect to some* $\mathscr{R} \subseteq \mathbf{R}$*, and* msa_X *the function defined in Definition II.26 with regard to* $X \in \{C, L, P\}$*. It then holds that*

$$\begin{aligned}
&\forall \mathscr{R} \subseteq \mathbf{R}\ \forall X \in \{C,L,P\}\ \forall \alpha_1;\ldots;\alpha_n \in \mathscr{L}_\Delta : \\
&\quad \mathsf{msa}_X(\alpha_1;\ldots;\alpha_n,\mathscr{R}) = \{(\gamma_1,\beta_1),\ldots,(\gamma_m,\beta_m)\} \quad \Longrightarrow \\
&\quad \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1,1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n,n) \mid\!\!\sim^{\mathsf{sk}}_X (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m))
\end{aligned}$$

Proof. Take an arbitrary perceptory condition $X \in \{C, L, P\}$, and arbitrary action sequence $\delta \in \mathscr{L}_\Delta$ such that $\delta = \alpha_1;\ldots;\alpha_i$. Let $\mathscr{R}$ be the set of rules given as argument to msa_X and underlying the abductive theory $\Lambda_{\mathscr{R}}$ as defined in Definition II.21, such that $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$. Consider Θ_X as determined by the choice of $X \in \{C, L, P\}$, and let R be the structural relation of Definition II.10 characterizing the perceptory condition X, and, in relation to Θ_X as defined in Definition II.21, specifically consider Definition II.9

if $R = \blacktriangleright$, Definition II.12 if $R = \blacktriangledown$, and Definition II.13 if $R = \odot$. Assume $\mathsf{msa}_X(\delta, \mathscr{R}) \neq \emptyset$ and pick any $(\gamma, \beta) \in \mathsf{msa}_X(\delta, \mathscr{R})$. Observe that because $(\gamma, \beta) \in \mathsf{msa}_X(\delta, \mathscr{R})$ it holds that $\exists (n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R} \exists \delta' \in OS(\pi) : (\delta R \delta')$. Let $\delta' = \alpha_1; \ldots; \alpha_j$, and note that from the definition of Θ_X it then follows that $((\mathsf{rule}(n) \wedge \mathsf{seq}(\iota_\delta(\delta'))) \rightarrow (\mathsf{obs}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{obs}(\alpha_j, j))) \in \Theta_X$. As Propositions II.2, II.3, and II.4 show, since $\alpha_1; \ldots; \alpha_i R \delta'$ there exists a (non-trivial) $\phi \in \mathscr{L}_1$ such that $\Theta_X \cup \{\phi\} \models \mathsf{obs}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{obs}(\alpha_i, i)$. Specifically, it holds that there exists an abducible $(\mathsf{rule}(n) \wedge \mathsf{seq}(y)) \in \mathscr{A}_{\mathscr{R}}$ for which, given $\omega = \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_i, i)$, it holds that $\Lambda_{\mathscr{R}}, \omega \not\approx_X (\mathsf{rule}(n) \wedge \mathsf{seq}(y))$. Let $\Phi = \mathsf{Th}(\Theta_X \cup \{\mathsf{rule}(n) \wedge \mathsf{seq}(y)\})$, noting that on grounds of the above it holds that $\Phi \in \Gamma^X_{\mathscr{R}}(\omega)$. Furthermore, because $(n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}$, as stated earlier, it follows from Definition II.15 that $\Phi \models \tau_g(\gamma) \wedge \tau_b(\beta)$. Since the choice of $(\gamma, \beta) \in \mathsf{msa}_X(\delta, \mathscr{R})$ was arbitrary, the above shows that for each $(\gamma', \beta') \in \mathsf{msa}_X(\delta, \mathscr{R})$ there exists a $\Phi' \in \Gamma^X_{\mathscr{R}}(\omega)$ such that $\Phi' \models \tau_g(\gamma') \wedge \tau_b(\beta')$. Thus, it follows by nature of disjunctive weakening that in all abductive extensions in $\Gamma^X_{\mathscr{R}}(\omega)$ the disjunction over translated elements of $\mathsf{msa}_X(\delta, \mathscr{R})$ is satisfied, so that on grounds of Definition II.24 the requirements for skeptical abduction are fulfilled. □

As Theorem II.2 shows, there is a correspondence between the output of the functions defined in Definition II.26 and facts which can be inferred from the abductive theory defined in Definition II.21, with respect to an underlying set of rules $\mathscr{R}$. Specifically, the disjunction over the conjoined translated goal/belief pairs in the output of an msa_X function, given a particular perceptory condition and input action sequence, can be skeptically inferred from the abductive theory. This claim is non-trivial, because as the proof shows there is in fact for each tuple in the output of an msa_X function an extension which satisfies the logical counterpart of that tuple, such that the disjunction over the counterparts of these tuples is satisfied in all extensions and can be skeptically inferred. This insight furthermore gives rise to Theorem II.3, which shows that for each such tuple holds that its logical counterpart can be credulously inferred.

Theorem II.3 (credulous interpretation of mental state abduction)**.** *Let* $\mathbf{R}$ *be the domain of* MYAPL *rules,* $\Lambda_{\mathscr{R}}$ *as defined in Definition II.21 with respect to some* $\mathscr{R} \subseteq \mathbf{R}$*, and* msa_X *as defined in Definition II.26 with regard to* $X \in \{C, L, P\}$*. It then holds that*

$$\begin{aligned} &\forall \mathscr{R} \subseteq \mathbf{R}\ \forall X \in \{C, L, P\}\ \forall \alpha_1; \ldots; \alpha_n \in \mathscr{L}_\Delta : \\ &\quad (\gamma, \beta) \in \mathsf{msa}_X(\alpha_1; \ldots; \alpha_n, \mathscr{R}) \quad \Longrightarrow \\ &\quad \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n) \not\approx^{\mathsf{cr}}_X \tau_g(\gamma) \wedge \tau_b(\beta) \end{aligned}$$

Proof. (This proof proceeds exactly like the proof of Theorem II.2, up to the last sentence, from where it is continued.) And if that is the case, then on grounds of Definition II.24 it follows that the requirements for credulous abduction are fulfilled. □

Both Theorems II.2 and II.3 establish a uni-directional correspondence, and one might wonder whether the converse of either of those claims also holds. As Propositions II.7 and II.8 show by means of counterexamples, this is not the case in general.

Proposition II.7. *Given the conditions of Theorem II.2, it holds that*

$$\begin{aligned}&\exists \mathscr{R} \subseteq \mathbf{R}\ \exists X \in \{C,L,P\}\ \exists \alpha_1;\ldots;\alpha_n \in \mathscr{L}_\Delta : \\ &\quad \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1,1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n,n) \mid\!\approx^{\mathsf{sk}}_X (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m)) \\ &\qquad \& \qquad \mathsf{msa}_X(\alpha_1;\ldots;\alpha_n,\mathscr{R}) \neq \{(\gamma_1,\beta_1),\ldots,(\gamma_m,\beta_m)\}\end{aligned}$$

II.4

Proof. Let $\mathscr{R} = \{p <\!\!- q \mid a, r <\!\!- s \mid b\}$, and verify that $\Lambda_{\mathscr{R}}, \mathsf{seen}(a,1) \mid\!\approx^{\mathsf{sk}}_C (\mathsf{goal}(p) \wedge \mathsf{belief}(q))$ and therefore also $\Lambda_{\mathscr{R}}, \mathsf{seen}(a,1) \mid\!\approx^{\mathsf{sk}}_C (\mathsf{goal}(p) \wedge \mathsf{belief}(q)) \vee (\mathsf{goal}(r) \wedge \mathsf{belief}(s))$. However, note that $\mathsf{msa}_C(a,\mathscr{R}) = \{(p,q)\}$ so that $(r,s) \notin \mathsf{msa}_C(a,\mathscr{R})$. □

Proposition II.8. *Given the conditions of Theorem II.3, it holds that*

$$\begin{aligned}&\exists \mathscr{R} \subseteq \mathbf{R}\ \exists X \in \{C,L,P\}\ \exists \alpha_1;\ldots;\alpha_n \in \mathscr{L}_\Delta : \\ &\quad \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1,1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n,n) \mid\!\approx^{\mathsf{cr}}_X \tau_g(\gamma) \wedge \tau_b(\beta) \\ &\qquad \& \qquad (\gamma,\beta) \notin \mathsf{msa}_X(\alpha_1;\ldots;\alpha_n,\mathscr{R})\end{aligned}$$

Proof. Let $\mathscr{R} = \{p <\!\!- q \mid a\}$, and consider $\tau_g(\gamma) = (\mathsf{goal}(p) \vee \mathsf{goal}(p'))$ and $\tau_b(\beta) = \mathsf{belief}(q)$, such that $\gamma = p$ or p' and $\beta = q$. Observe that, indeed, $\Lambda_{\mathscr{R}}, \mathsf{seen}(a,1) \mid\!\approx^{\mathsf{sk}}_C (\mathsf{goal}(p) \vee \mathsf{goal}(p')) \wedge \mathsf{belief}(q)$, but that $(p \text{ or } p', q) \notin \mathsf{msa}_C(a,\mathscr{R})$. □

Theorems II.2 and II.3 show that the functional approach captures an important aspect of the abductive theory; specifically, those theorems show that the logical counterpart of goal/belief pairs in the output of the mental state abduction functions can be deduced in the extensions of the abductive theory. However, the functional approach is also limited, in a sense, as reflected by Propositions II.7 and II.8 that illustrate the greater expressiveness of the pure logical approach, showing through in the fact that skeptical/credulous abduction allow for inference of facts that are not as such output by the functions. As stated before, though, it was our goal to gain conciseness and a convenient level of abstraction by using the functional approach, and the latter two propositions only show that there is (as usual) a price to pay. And, as the following propositions show, this price is not very high in regard to the limited expressiveness illustrated by Propositions II.7 and II.8, as slightly weaker properties still exists.

Proposition II.9. *Given the conditions of Theorem II.2, it holds that*

$$\begin{aligned}&\forall \mathscr{R} \subseteq \mathbf{R}\ \forall X \in \{C,L,P\}\ \forall \alpha_1;\ldots;\alpha_n \in \mathscr{L}_\Delta : \\ &\quad (\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1,1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n,n) \mid\!\approx^{\mathsf{sk}}_X (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m)) \\ &\qquad \& \qquad \mathsf{msa}_X(\alpha_1;\ldots;\alpha_n,\mathscr{R}) = \{(\gamma_1,\beta_1),\ldots,(\gamma_k,\beta_k)\}) \qquad \Longrightarrow \\ &\qquad\quad \{(\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_k) \wedge \tau_b(\beta_k))\} \models \\ &\qquad\qquad (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m))\end{aligned}$$

Proof. First of all, see that any instances of belief/1 and goal/1 that are found in an abductive extension on grounds of the relation $\mid\!\approx^{\mathsf{sk}}_X$, given some observed sequence actions, are there

because of a relation of the form $\mathsf{rule}(n) \rightarrow (\ldots)$; see Definition II.15. Thus, the skeptically inferred mental state description $(\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m))$ is either the disjunction over the mental state descriptions in each of the extensions, or a weakening thereof. As shown by Theorem II.2, it holds in the former case that each of those disjuncts has a counterpart in the output of the corresponding mental state abduction function, and this theorem applies. In the latter case, observe that this disjunctive weakening is entailed by the (translated) output of that function. □

Proposition II.10. *Given the conditions of Theorem II.3, it holds that*

$$\begin{aligned}
&\forall \mathscr{R} \subseteq \mathbf{R}\ \forall X \in \{C, L, P\}\ \forall \alpha_1; \ldots; \alpha_n \in \mathscr{L}_\Delta : \\
&\quad (\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n) \mid\!\!\approx^{\mathsf{cr}}_X (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m)) \\
&\qquad \Longrightarrow \quad \exists (\gamma, \beta) \in \mathsf{msa}_X(\alpha_1; \ldots; \alpha_n, \mathscr{R}) : \\
&\qquad\qquad \{(\tau_g(\gamma) \wedge \tau_b(\beta))\} \models (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m))
\end{aligned}$$

Proof. (Follows by analogy with the proof of Proposition II.9, noting that any credulously abduced mental state description either has a direct counterpart in the mental state abduction function that operates under the same perceptory condition, or can be inferred as a disjunctive weakening thereof.) □

4.3 Properties of Mental State Abduction

The purpose of this section is to explain and prove useful properties of the functions defined in Definition II.26, which can be given skeptical or credulous interpretations with respect to a logical abductive theory as explained in Section 4.2. Because of the uni-directional correspondence between the functional approach and the corresponding logical abductive theory, formal properties of the functional approach have their reflection in that theory as well.

Proposition II.11. *Let* $\mathbf{R}$ *be the domain of* MYAPL *rules,* $\mathsf{msa}_C, \mathsf{msa}_L, \mathsf{msa}_P$ *be the explanatory functions of Definition II.26, and* $\mathscr{L}_\Delta$ *the percept language. Then*

$$\forall \mathscr{R} \subseteq \mathbf{R}\ \forall \delta \in \mathscr{L}_\Delta : \quad \mathsf{msa}_C(\delta, \mathscr{R}) \subseteq \mathsf{msa}_L(\delta, \mathscr{R}) \subseteq \mathsf{msa}_P(\delta, \mathscr{R})$$

Proof. Because the explanatory functions differ only in their particular characterizing relation, it suffices to show that $(\blacktriangleright) \subseteq (\blacktriangledown) \subseteq (\underline{\odot})$, as proven in Proposition II.1. □

Proposition II.12. *Let* $\mathbf{R}$ *be the domain of* MYAPL *rules,* $\mathsf{msa}_C, \mathsf{msa}_L, \mathsf{msa}_P$ *be the explanatory functions of Definition II.26, and* $\mathscr{L}_\Delta$ *the percept language. Then*

$$\exists \mathscr{R} \subseteq \mathbf{R}\ \exists \delta \in \mathscr{L}_\Delta : \quad \mathsf{msa}_P(\delta, \mathscr{R}) \not\subseteq \mathsf{msa}_L(\delta, \mathscr{R}) \not\subseteq \mathsf{msa}_C(\delta, \mathscr{R})$$

Proof. Let $\mathscr{R} = \{p \mathrel{<\!\!-} q \mid a; b; c,\ r \mathrel{<\!\!-} q \mid b; a; c\}$ and consider the following counterexample. Given the percept $a; c \in \mathscr{L}_\Delta$, verify that $\mathsf{msa}_P(a; c, \mathscr{R}) = \{(p, q), (r, s)\}$ and $\mathsf{msa}_L(a; c, \mathscr{R}) = \{(r, s)\}$ and $\mathsf{msa}_C(a; c, \mathscr{R}) = \emptyset$. □

Informally put, Proposition II.11 shows that assumption of 'worse' perceptory conditions — e.g. partial observation instead of complete observation, as explained in Section 3.2 — cannot lead to a decrease in the number of observations. Intuitively this makes sense, because if in explaining the agent's behavior it is assumed that the agent *might* have done some actions other than those which had been already assumed — which is, essentially, what occurs if a worse perceptory condition is assumed to explain the same sequence of actions — one would expect not to lose the explanations which had already been inferred. Similarly, and also informally put, one would expect the set of explanations inferred under some particular perceptory condition not to increase with additional observations under the same perceptory condition. This is reflected in the following proposition.

Proposition II.13. *Let* **R** *be the domain of* MYAPL *rules,* $\mathsf{msa}_C, \mathsf{msa}_L, \mathsf{msa}_P$ *be the explanatory functions of Definition II.26,* $\mathscr{L}_\Delta$ *and* $R \in \{\underline{\blacktriangleright}, \underline{\blacktriangledown}, \underline{\odot}\}$ *the structural relation for* msa_X *(*$\underline{\blacktriangleright}$*,* $\underline{\blacktriangledown}$*, and* $\underline{\odot}$ *for* msa_C*,* msa_L*, and* msa_P*, respectively). It then holds that*

$$\begin{aligned}&\forall \mathscr{R} \subseteq \mathbf{R}\ \forall X \in \{C, L, P\}\ \forall \delta, \delta' \in \mathscr{L}_\Delta : \\ &\qquad (\delta, \delta') \in R \quad \Longrightarrow \quad \mathsf{msa}_X(\delta', \mathscr{R}) \subseteq \mathsf{msa}_X(\delta, \mathscr{R})\end{aligned}$$

Proof. For any set of rules $\mathscr{R}$ and any msa_X, the claim $\mathsf{msa}_X(\delta', \mathscr{R}) \subseteq \mathsf{msa}_X(\delta, \mathscr{R})$ holds if and only if $\forall(\gamma, \beta) \in \mathsf{msa}_X(\delta', \mathscr{R}) : (\gamma, \beta) \in \mathsf{msa}_X(\delta, \mathscr{R})$. From Definition II.26 follows for any (γ, β) that if $(\gamma, \beta) \in \mathsf{msa}_X(\delta', \mathscr{R})$ then $\exists(\gamma <\!\!- \beta \mid \pi) \in \mathscr{R}\, \exists \delta'' \in OS(\tau_p(\pi))$ such that $(\delta', \delta'') \in R$. By transitivity, from $(\delta, \delta') \in R$ and $(\delta', \delta'') \in R$ follows $(\delta, \delta'') \in R$, such that if $(\gamma, \beta) \in \mathsf{msa}_X(\delta', \mathscr{R})$ then it must be the case that $(\gamma, \beta) \in \mathsf{msa}_X(\delta, \mathscr{R})$. □

Corollary II.2. *Explanation is* monotonic *with respect to appended single actions, i.e.*

$$\forall \mathscr{R} \subseteq \mathbf{R}\ \forall X \in \{C, L, P\}\ \forall \delta, \alpha \in \mathscr{L}_\Delta : \quad \mathsf{msa}_X(\delta; \alpha, \mathscr{R}) \subseteq \mathsf{msa}_X(\delta, \mathscr{R})$$

Proof. For any $R \in \{\underline{\blacktriangleright}, \underline{\blacktriangledown}, \underline{\odot}\}$ and $\alpha, \delta \in \mathscr{L}_\Delta$ holds that $(\delta, \delta; \alpha) \in R$. From this fact and Proposition II.13 it follows that $\mathsf{msa}_X(\delta; \alpha, \mathscr{R}) \subseteq \mathsf{msa}_X(\delta, \mathscr{R})$. □

Proposition II.14. *Let* **R** *be the domain of* MYAPL *rules and* $R, R' \in \{\underline{\blacktriangleright}, \underline{\blacktriangledown}, \underline{\odot}\}$ *be the relations that characterize the function* msa_X *and* $\mathsf{msa}_{X'}$*; i.e.* $\underline{\blacktriangleright}$*,* $\underline{\blacktriangledown}$*, and* $\underline{\odot}$ *for* msa_C*,* msa_L*, and* msa_P*, respectively, where* R *belongs to* msa_X *and* R' *to* msa_X*. It then holds that*

$$\begin{aligned}&\forall \mathscr{R} \subseteq \mathbf{R}\ \forall R, R' \in \{\underline{\blacktriangleright}, \underline{\blacktriangledown}, \underline{\odot}\}\ \forall \delta, \delta' \in \mathscr{L}_\Delta : \\ &\qquad ((\delta, \delta') \in R \quad \& \quad R \subseteq R') \quad \Longrightarrow \quad \mathsf{msa}_X(\delta', \mathscr{R}) \subseteq \mathsf{msa}_{X'}(\delta, \mathscr{R})\end{aligned}$$

Proof. By nature of the subset relation follows that if $(\delta, \delta') \in R$ and $R \subseteq R'$ then $(\delta, \delta') \in R'$. Take any set of MYAPL rules $\mathscr{R}$ and note that Proposition II.13 shows that if $(\delta, \delta') \in R$ then $\mathsf{msa}_X(\delta', \mathscr{R}) \subseteq \mathsf{msa}_X(\delta, \mathscr{R})$. Note from the proof of Proposition II.11 that $R \subseteq R'$ is sufficient to conclude $\mathsf{msa}_X(\delta, \mathscr{R}) \subseteq \mathsf{msa}_{X'}(\delta, \mathscr{R})$, and by transitivity of $\subseteq$ it then follows from $\mathsf{msa}_X(\delta', \mathscr{R}) \subseteq \mathsf{msa}_X(\delta, \mathscr{R})$ and $\mathsf{msa}_X(\delta, \mathscr{R}) \subseteq \mathsf{msa}_{X'}(\delta, \mathscr{R})$ that $\mathsf{msa}_X(\delta', \mathscr{R}) \subseteq \mathsf{msa}_{X'}(\delta, \mathscr{R})$. □

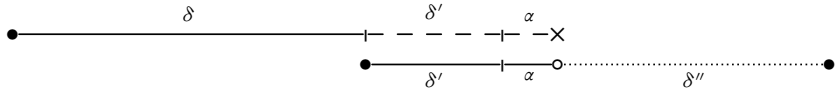


Figure II.2: Explainable sequences $\delta;\delta'$ and $\delta';\alpha$, and an unexplainable sequence $\delta;\delta';\alpha$.

Proposition II.13 shows that, indeed, the set of explanations inferred under some particular perceptory condition for a given observation does not increase with additional observations under the same perceptory condition. Corollary II.2 shows that this is specifically the case for any condition given a single 'novel' observed action, and Proposition II.14 generalizes this to different underlying relations. The latter result also conforms to intuition, because inference under a particular perceptory condition with a particular sequence of observed actions should not be expected to yield more explanations than inference with fewer observed actions under a worse assumed perceptory condition (i.e. a condition which allows for more missing observations because its characterizing relation is a superset of that of the better perceptory condition). The aforementioned theoretical results also have some practical implications. From Definition II.26 it should be evident that the presence of a particular goal/belief pair in the output of an msa_X function depends on the input sequence of actions being related to a plan that, in turn, is related to that goal/belief pair on grounds of a rule. Propositions II.11, II.13, and II.14, along with Corollary II.2, show, effectively, that with increasing observations one needs to reconsider only those rules which already yielded explanations; this can be pictured as the explanatory process 'zooming in' on the best explanation with novel observed actions. As the theoretical results show, this holds likewise for explanation under better perceptory conditions.

Another computational issue concerns the amount of backtracking the observer must perform in order to find explanations, in case some sequence of actions cannot be explained. After all, if explanation is to be a process which occurs in step with the agent performing its actions, then it can in principle occur that the observer is at some point unable to explain those actions because they do not match with the observable sequence of any plan (although it was assumed here that all observation are taken from $\Omega_{\mathscr{R}}$). This might happen because the agent either finished a plan or dropped a plan before its completion, if its commitment type allows this (cf. Section 2.1). It should be noted that if the observer 'backtracks' along the sequence that it is attempting to explain, in order to arrive at an explainable suffix, then it can occur in the case of complete observation that it must backtrack more than just a single action in order to find an explanation, because of possible overlap between sequences. This situation is visualized in Figure II.2, in which the sequence δ' is the suffix of the sequence $\delta;\delta'$ as well as the prefix of some other sequence $\delta';\alpha;\delta''$, and it holds that both can be explained by the msa_C function, whereas $\delta;\delta';\alpha$ cannot. If such a situation occurs and explanation fails, then the observer can backtrack along δ' until it finds a suffix that can be explained in coherence with the last observed action (e.g. α in Figure II.2, the initial action of δ''). It is possible to give a measure on how far the observer must backtrack, as shown

in Proposition II.16. First note, though, that the need for backtracking in principle only occurs if complete observation is presumed, because under the presumption of incomplete observation an explanation can always be found on grounds of the last observed action, as shown below.

Proposition II.15. *Let $\mathscr{R}$ be an agent's set of* MYAPL *rules and $\alpha \in \mathscr{L}_\Delta$ an action the agent is observed to perform, and which occurs in one of its plans. It then holds that*

$$(\mathsf{msa}_L(\alpha, \mathscr{R}) = \mathsf{msa}_P(\alpha, \mathscr{R})) \neq \emptyset$$

Proof. Given the fact that the agent rules are known, it must be the case that $\exists(n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}$ such that the action α was performed the agent on grounds of plan π. In that case, it holds that $\exists \delta \in OS(\pi) : (\alpha \underline{\blacktriangledown} \delta)$, noting that any single action which occurs in a sequence is per definition a substring of that sequence. □

Because of the fact that finding a measure for the required amount of backtracking is relevant mostly in the case of complete observation, it is presumed for the following proofs that, indeed, observation is complete. Furthermore, in order to provide the desired measures, the function $\mathsf{len} : \mathscr{L}_\Delta \longrightarrow \mathbb{N}_1$ is defined to map action sequences to their length (i.e. the number of actions in the sequence).

Proposition II.16. *Let* **R** *be the domain of* MYAPL *rules, and let $\delta \in \mathscr{L}_\Delta$ be an observed action sequence such that the first action of δ is also the initial action of the plan π the agent actually selected. It then holds that*

$$\begin{aligned} &\forall \mathscr{R} \subseteq \mathbf{R}\ \forall \alpha \in \mathscr{L}_\Delta : \\ &\quad (\mathsf{msa}_C(\delta, \mathscr{R}) \neq \emptyset)\ \&\ (\mathsf{msa}_C(\delta; \alpha, \mathscr{R}) = \emptyset) \implies \\ &\quad \exists \delta' \in \mathscr{L}_\Delta : \quad (\mathsf{msa}_C(\delta', \mathscr{R}) \neq \emptyset)\ \&\ (\mathsf{len}(\delta') \leq \mathsf{len}(\delta)) \end{aligned}$$

Proof. It can occur that, after performing the initial action of π, the agent drops this plan. In the worst case, with respect to the required amount of backtracking, if $\delta = \alpha'; \alpha''; \delta'$ then α'' could be the initial action of some plan π' which the agent selected after dropping π and it could hold that $\neg\exists(n : \gamma <\!\!- \beta \mid \pi'') \in \mathscr{R} \exists \delta'' \in OS(\pi'') : (\delta'; \alpha \underline{\blacktriangleright} \delta'')$. In that case, because observation is complete, it must be that $\exists(n : \gamma <\!\!- \beta \mid \pi'') \in \mathscr{R} \exists \delta'' \in OS(\pi'') : (\alpha''; \delta'; \alpha \underline{\blacktriangleright} \delta'')$ in which case $\mathsf{len}(\alpha''; \delta'; \alpha) = \mathsf{len}(\delta)$. If the situation is not worst-case, then $\exists(n : \gamma <\!\!- \beta \mid \pi'') \in \mathscr{R} \exists \delta'' \in OS(\pi'') \exists \delta''' \in \mathscr{L}_\Delta : (\delta''' \underline{\blacktriangleleft} \delta' \ \&\ \delta'''; \alpha \underline{\blacktriangleright} \delta'')$, in which case $\mathsf{len}(\delta''; \alpha) \leq \mathsf{len}(\delta)$. □

Corollary II.3. *Let $\mathscr{R}$ be such that for $(n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}$ holds that $\delta \in OS(\pi)$ is the longest observable sequence of any plan. Given complete observation, it then follows from Proposition II.16 that the maximum required amount of backtracking is* $\mathsf{len}(\delta)$.

5 Reflection

This chapter deals with explanation of behavior using logical abduction, an approach to defeasible reasoning discussed in Section 1. It assumes the perspective of an abstract observer, which perceives and explains the actions of a BDI-based software agent. Agents are assumed to be programmed in the language MYAPL, an agent programming language introduced in Section 2. This language is 'fictional' but shares elements with many existing agent programming languages, so that the approach taken in this work with respect to the MYAPL language can be related to those existing languages as well. In Section 3 the foundation for this approach is laid down in classical logic, focusing on abductive explanation of a BDI agent's observed behavior. Of particular note is the introduction of an abductive theory, which comprises three distinct classical theories for reasoning about the agent's behavior under different perceptory conditions that range from complete observation to partial observation. It is proven that use of this theory with specific classes of observations and explanations corresponds to classical abductive reasoning as defined at the beginning of this chapter, and an elaborate example illustrates that use. The classical logical approach has its benefits but also its limitations, one of which is that it is somewhat unwieldy. In order to remedy this aspect, a functional approach is defined in Section 4 which is shown to be closely related to the logical approach but abstracts from its specifics, allowing for concise expression of defeasible inference of an agent's mental state as explanation for its observed behavior.

This chapter pointed out another limitation of our formulation using the classical logical approach, which is that it is not particularly suited for reasoning about dynamics. Although approaches such as the situation calculus show that this kind of reasoning can be done in terms of classical logic, we wish to employ a formalism in which expressions about the dynamics of the system are not intertwined with expressions about the state of the system, and which has an inherent notion of 'possibility'. It was mentioned that modal logic is more suited for that task, which is so (amongst other reasons) because of the fact that modal semantics is essentially based on structures that enable the 'local' evaluation of expressions (Blackburn et al., 2001). This quality is put to use in the next chapter in order to provide an alternative interpretation of mental state abduction, one which does not only focus on the defeasible nature of explanations but also incorporates the dynamic nature of actions.

II.5

Chapter III

Dynamics in Ascription

This chapter focuses on modeling the dynamics of actions, in a context of explanation of observed behavior, using propositional dynamic logic. Focus is on identifying classes of models that represent interpretations of particular states of affairs in which mental states are attributed to an agent whose actions are observed. The previous chapter introduced a formal approach to reasoning about observed behavior using defeasible (abductive) reasoning in classical logic, that will serve as the basis for the approach presented here. As before, focus is on an observer that perceives and explains the actions of an agent. The chapter is structured as follows: first, in Section 1 the basic formalism is introduced and some of its properties are analyzed. In Section 2 this formalism is used to provide a dynamic counterpart for the mental state abduction functions introduced in the previous chapter. Section 3 elaborates on this account, showing the increased expressivity of this dynamic logical approach over the classical logical approach. Interspersed with brief and more elaborate examples that further develop the theme introduced earlier, this chapter concludes with Section 4.

1 A Propositional Dynamic Logic

Modal logics in general deal with relational structures, and provide the means to consider such structures from a 'local' perspective; see the book by Blackburn et al. (2001) for a thorough introduction to modal logic. Propositional dynamic logic (PDL) is such a modal logic, suitable for reasoning about dynamics. The modalities of PDL consist of dynamic expressions, and allow for statements such as "after program π it is necessarily the case that ϕ holds", or "after action α has been performed there always exists a state preceding α". The fact that complex expressions can occur 'inside' modalities makes PDL stand out from many other modal logics, whose languages typically comprise only a few basic modalities. In the following, the syntax and semantics are put forward of the propositional dynamic logic used in this chapter.

1.1 Syntax

Conveniently, the plan language $\mathscr{L}_{\Pi}$ defined in Definition II.4 can serve as the basis for our variant of PDL. This language is defined by mutual induction on $\mathscr{L}_{\Pi}$ and the propositional language $\mathscr{L}_0$ mentioned in Chapter II, which was left undefined there but which is defined at present.

Definition III.1 (propositional language $\mathscr{L}_0$)**.** *Let* $\top$ *denote the truth constant or 'verum' and* $\bot$ *the falsity constant or 'falsum', and let* Atom *be a set of propositional atoms such that* $\{\top, \bot\} \subseteq \mathsf{Atom}$*. The language* $\mathscr{L}_0$ *is then the smallest set satisfying the following clauses.*

- *If* $p \in \mathsf{Atom}$ *then* $p, {\sim} p \in \mathscr{L}_0$.
- *If* $\phi, \psi \in \mathscr{L}_0$ *then* $\phi \vee \psi, \phi \wedge \psi \in \mathscr{L}_0$.

It should be noted that the language $\mathscr{L}_0$ uses the symbol '$\sim$' for negation. The reason for this is to avoid confusion between the negation symbol '$\neg$' and $\mathscr{L}_0$-negation '$\sim$', the latter of which is shown in Section 1.2.1 to have a somewhat non-standard interpretation. Because of that fact, the logical connectives for disjunction and conjunction are both defined as primitives in $\mathscr{L}_0$. The observer language $\mathscr{L}_{\mathbf{M}}$ is then defined on top of $\mathscr{L}_\Pi$ and $\mathscr{L}_0$ by mutual induction, as follows, where it should be noted that the classical negation symbol '$\neg$' does occur in the language $\mathscr{L}_{\mathbf{M}}$.

Definition III.2 (observer language $\mathscr{L}_{\mathbf{M}}$)**.** *Let* $\mathscr{L}_0$ *be defined as in Definition III.1 and* $\mathscr{L}_\Pi$ *as in Definition II.4, based on the sets* Atom *of atomic propositions and* Act *of primitive actions. The observer language* $\mathscr{L}_{\mathbf{M}}$ *is then the smallest set satisfying the following clauses.*

III.1

- *If* $p \in \mathsf{Atom}$ *then* $p \in \mathscr{L}_{\mathbf{M}}$.
- *If* $\alpha \in \mathsf{Act}$ *then* $\mathbf{Obs}(\alpha) \in \mathscr{L}_{\mathbf{M}}$.
- *If* $\phi \in \mathscr{L}_0$ *then* $\mathbf{B}(\phi), \mathbf{G}(\phi) \in \mathscr{L}_{\mathbf{M}}$.
- *If* $\pi \in \mathscr{L}_\Pi$ *and* $\phi \in \mathscr{L}_{\mathbf{M}}$ *then* $[\pi](\phi), [\pi^-](\phi) \in \mathscr{L}_{\mathbf{M}}$.
- *If* $\phi, \psi \in \mathscr{L}_{\mathbf{M}}$ *then* $\neg\phi, \phi \vee \psi \in \mathscr{L}_{\mathbf{M}}$.

Non-primitive connectives of $\mathscr{L}_{\mathbf{M}}$ *are defined as follows, given* $\phi, \psi \in \mathscr{L}_{\mathbf{M}}$.

$$\begin{array}{lcl} \phi \wedge \psi & \triangleq & \neg(\neg\phi \vee \neg\psi) \\ \phi \rightarrow \psi & \triangleq & \neg\phi \vee \psi \\ \phi \leftrightarrow \psi & \triangleq & (\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi) \end{array}$$

Non-primitive modalities of $\mathscr{L}_{\mathbf{M}}$ *are defined as follows, given* $\phi \in \mathscr{L}_{\mathbf{M}}$ *and* $\pi \in \mathscr{L}_\Pi$.

$$\begin{array}{lcl} \langle\pi\rangle\phi & \triangleq & \neg[\pi]\neg\phi \\ \langle\pi^-\rangle\phi & \triangleq & \neg[\pi^-]\neg\phi \end{array}$$

The language $\mathscr{L}_{\mathbf{M}}$ is a variant of PDL with converse actions and poor tests, and serves to formalize reasoning about the actions and mental state of some agent by an observer. It is important to keep this perspective upon this framework in mind, and realize that the operators **B** and **G** qualify propositions *ascribed* to the agent by the observer as its beliefs and goals, respectively. This interpretation differs from that of BDI logics (Cohen & Levesque, 1990; Rao & Georgeff, 1991) or similar logics such as KARO (van der Hoek et al., 1998), which also employ doxastic and motivational operators but are intended to be used as external metalanguages for external reasoning about intelligent agents (Singh et al., 1999). In contrast, the intended reading of the language $\mathscr{L}_{\mathbf{M}}$ is as a language that assumes the perspective of the observer in reasoning about intelligent agents. It is noteworthy in this regard that the doxastic and motivational operators of $\mathscr{L}_{\mathbf{M}}$ are rather simplistic, and are best considered as notational devices that 'label' or 'mark' propositional expressions. This simplicity shows in the fact that the language does not allow nesting of doxastic/motivational operators, and

also shows in the semantics of those operators (of which more is said in the Section 1.2). It would appear that this simplicity is favorable in computational terms, because it ensures that the language $\mathscr{L}_{\mathbf{M}}$ is technically more akin to standard PDL than to the BDI or agent logics mentioned above. Last but not least, note that the operator **Obs** qualifies actions of the agent as having been observed and accepts only actions $\alpha \in \mathsf{Act}$.

It should be stressed at this point that the set Atom of atomic propositions is shared by $\mathscr{L}_{\mathbf{M}}$ and $\mathscr{L}_0$, and that the referents of atomic propositions in those languages are the same. The difference lies in the fact that the language $\mathscr{L}_{\mathbf{M}}$ expresses the perspective of the observer on states of affairs, i.e. its view on 'the world', so that, given $p \in \mathsf{Atom}$, $w \models p$ means that in the state w the observer takes p to be the case. Expressions of the language $\mathscr{L}_0$ only occur in $\mathscr{L}_{\mathbf{M}}$ as arguments to the **B** and **G** operators, and concern facts that are attributed to the agent, as its beliefs or goals, by the observer. Accordingly, $w \models \mathbf{B}(p)$ means that in state w the observer attributes the belief that p is the case to the agent. This reading of $\mathscr{L}_{\mathbf{M}}$ is perhaps straightforward, but it is our conviction that it deserves mentioning in order to avoid confusion. Last but not least, the modalities $[\pi]$ and $[\pi^-]$ allow for qualifying propositions that, according to the observer, *necessarily* hold after and before execution of some plan π by the agent, and their duals $\langle\pi\rangle$ and $\langle\pi^-\rangle$ do likewise in an existential sense, referring to things that are *possibly* the case.

1.2 Semantics

The observer language $\mathscr{L}_{\mathbf{M}}$ is a modal language, and as usual it has Kripke semantics (Harel et al., 2000). This language has underlying structures (frames) of the type $\mathfrak{F} = (\mathbb{W}, \{R_\alpha \mid \alpha \in \mathsf{Act}\})$, where $\mathbb{W}$ is a set of states, and $R_\alpha \subseteq \mathbb{W} \times \mathbb{W}$ are accessibility relations for primitive actions α. A model $\mathfrak{M}$ is a tuple $\mathfrak{M} = (\mathfrak{F}, \vartheta)$ of a frame and a valuation function $\vartheta : \mathscr{L}_{\mathbf{M}} \longrightarrow \wp(\mathbb{W})$, which for technical clarity and simplicity is in itself a tuple $\vartheta = (\vartheta_{\mathsf{B}}, \vartheta_{\mathsf{G}}, \vartheta_{\mathsf{I}}, \vartheta_{\mathsf{O}}, \varrho)$ of multiple functions, interpreted as follows.

Definition III.3 (semantics of $\mathscr{L}_{\mathbf{M}}$). *Let $\vartheta_{\mathsf{B}}, \vartheta_{\mathsf{G}} : \mathscr{L}_0 \longrightarrow \wp(\mathbb{W})$ be functions assigning sets of states to expressions of $\mathscr{L}_0$, $\vartheta_{\mathsf{I}} : \mathsf{Atom} \longrightarrow \wp(\mathbb{W})$ a function assigning sets of states to atoms, $\vartheta_{\mathsf{O}} : \mathsf{Act} \longrightarrow \wp(\mathbb{W})$ a function assigning sets of states to actions, and $\varrho : \mathscr{L}_\Pi \longrightarrow \wp(\mathbb{W} \times \mathbb{W})$ a function assigning sets of pairs of states to dynamic expressions. The valuation function ϑ is defined on top of those functions, as follows, given $p \in \mathsf{Atom}$, $\phi, \phi' \in \mathscr{L}_{\mathbf{M}}$, $\psi, \psi' \in \mathscr{L}_0$, $\varkappa \in \{\mathbf{B}(\psi), \neg\mathbf{B}(\psi), \mathbf{G}(\psi), \neg\mathbf{G}(\psi)\}$, $\alpha \in \mathsf{Act}$, and $\pi, \pi' \in \mathscr{L}_\Pi$. Note that $\vartheta_{\mathsf{B}}(\top) = \vartheta_{\mathsf{B}}(\sim\bot) = \vartheta_{\mathsf{G}}(\top) = \vartheta_{\mathsf{G}}(\sim\bot) = \vartheta_{\mathsf{I}}(\top) = \mathbb{W}$, and $\vartheta_{\mathsf{B}}(\bot) = \vartheta_{\mathsf{B}}(\sim\top) = \vartheta_{\mathsf{G}}(\bot) = \vartheta_{\mathsf{G}}(\sim\top) = \emptyset$.*

$$
\begin{aligned}
\vartheta(p) &= \vartheta_{\mathsf{I}}(p) \\
\vartheta(\neg\phi) &= \mathbb{W} - \vartheta(\phi) \\
\vartheta(\phi \vee \phi') &= \vartheta(\phi) \cup \vartheta(\phi') \\
\vartheta(\mathbf{B}(\psi)) &= \vartheta_{\mathsf{B}}(\psi) \\
\vartheta_{\mathsf{B}}(\psi \vee \psi') &= \vartheta_{\mathsf{B}}(\psi) \cup \vartheta_{\mathsf{B}}(\psi') \\
\vartheta_{\mathsf{B}}(\psi \wedge \psi') &= \vartheta_{\mathsf{B}}(\psi) \cap \vartheta_{\mathsf{B}}(\psi')
\end{aligned}
$$

$$
\begin{aligned}
\vartheta(\mathbf{G}(\psi)) &= \vartheta_{\mathsf{G}}(\psi)\\
\vartheta_{\mathsf{G}}(\psi \vee \psi') &= \vartheta_{\mathsf{G}}(\psi) \cup \vartheta_{\mathsf{G}}(\psi')\\
\vartheta_{\mathsf{G}}(\psi \wedge \psi') &= \vartheta_{\mathsf{G}}(\psi) \cap \vartheta_{\mathsf{G}}(\psi')\\
\vartheta(\mathbf{Obs}(\alpha)) &= \vartheta_{\mathsf{O}}(\alpha)\\
\vartheta([\pi](\phi)) &= \{w \mid \forall (w,w') \in \varrho(\pi) : w' \in \vartheta(\phi)\}\\
\vartheta([\pi^-](\phi)) &= \{w \mid \forall (w,w') \in \varrho(\pi)^{-1} : w' \in \vartheta(\phi)\}\\
\varrho(\alpha) &= R_\alpha\\
\varrho(x?) &= \{(w,w) \mid w \in \vartheta(x)\}\\
\varrho(\pi;\pi') &= \varrho(\pi) \circ \varrho(\pi')\\
\varrho(\pi + \pi') &= \varrho(\pi) \cup \varrho(\pi')\\
\varrho(\pi^*) &= \bigcup_{n \in \mathbb{N}_0} \varrho(\pi^n), \textit{ where } \pi^0 = \mathbf{B}(\top)?
\end{aligned}
$$

III.1

The operator $^{-1}$ denotes the inverse of a relation, such that for any relation R, it is the case that $R^{-1} = \{(y,x) \mid (x,y) \in R\}$. In the definition of ϱ, the operator $\circ$ denotes relational composition, such that for any two relations R and R', their composition $R \circ R' = \{(x,z) \mid (x,y) \in R, (y,z) \in R'\}$. As usual $\cup$ denotes set union and π^* denotes iteration of π, the semantics of which is the reflexive transitive closure of $\varrho(\pi)$ (Harel et al., 2000). Zero-length iteration of π is defined as the '$\mathbf{B}(\top)?$' action, i.e. a test action on whether the agent is attributed belief in truth, which in this particular incarnation of PDL can be seen as the counterpart of a 'skip' action. Given a model $\mathfrak{M} = (\mathfrak{F}, \vartheta)$, the fact that a particular state w in the model satisfies a formula $\phi \in \mathscr{L}_{\mathbf{M}}$ is defined as

$$\mathfrak{M}, w \models \phi \iff w \in \vartheta(\phi)$$

$\mathfrak{M} \models \phi$ denotes that any state in $\mathfrak{M}$ satisfies ϕ, and $\models \phi$ denotes that ϕ is valid in all of the models considered. If $\mathscr{M}$ is a class of models then $\mathscr{M} \models \phi$ is equivalent to $\forall \mathfrak{M} \in \mathscr{M} : (\mathfrak{M} \models \phi)$. As a shorthand $\mathfrak{M}, \{w_1, \ldots, w_n\} \models \phi$ can be used instead of $\mathfrak{M}, w_1 \models \phi$ & … & $\mathfrak{M}, w_n \models \phi$. Also, if $w \models \phi$ is used then it is assumed that the model $\mathfrak{M}$ is evident from context. For notational convenience, given that $\mathfrak{M} = (\mathfrak{F}, \vartheta)$ and $\mathfrak{F} = (\mathrm{W}, \{R_\alpha \mid \alpha \in \mathsf{Act}\})$, it holds that '$\mathrm{W}_{\mathfrak{M}}$' denotes W.

1.2.1 Semantics of Ascription

Ascription[1] of beliefs and goals by means of the operators **B** and **G** and the language $\mathscr{L}_0$, has semantics that deviate somewhat from those of classical (propositional) logic, in the sense that its interpretation makes it possible to express a kind of 'four-valuedness' (Belnap, 1977)

[1] 'Ascription' is used here as synonymous to 'attribution', and if either seems at any point favored over the other in a particular context, then this is solely for the sake of variation.

in the language, specific to ascription of mental states, to be interpreted as follows. First of all, ascription allows for a sort of *paraconsistency*, meaning that opposite literals can be simultaneously true without giving rise to the classical kind of inconsistency that allows for 'ex falso quodlibet' (Stanford Encyclopedia of Philosophy, 2009). Also, the semantics of $\mathscr{L}_0$ allow for *indeterminacy*, meaning that it is possible that neither some atomic proposition nor its literal opposite is taken to be true. The rationale behind this approach is that the proposed formalism is intended for characterization of an observer that reasons about the behavior of an observed agent. In that light, it is not hard to imagine situations in which the observer has no grounds to ascribe to the agent belief in either some atomic proposition or its negation. Or, on grounds of observed behavior, to be justified in attributing belief to the agent in some atom and likewise its negation, but being unable to determine conclusively which should be ruled out. Similarly, this holds for attribution of goals.

In regard to the informal interpretation of the operators **B** and **G** qualifying ascribed beliefs and goals, respectively, Section 1.1 already mentioned that their intended use differs from that of similar operators in BDI or agent logics. Such logics typically use accessibility relations between states (or 'worlds') for interpretation of those modalities, in contrast to our approach where evaluation of those operators occurs in relation to a single state. This fact makes it interesting to compare those approaches, especially in cases where syntax is identical but the difference lies in semantics. The following proposition expresses such a difference, where for the sake of illustration a necessity-type doxastic modality $\mathbf{B}^{\Box}$ is assumed that accepts the propositional logical subset of $\mathscr{L}_{\mathbf{M}}$ (i.e. atoms composed with $\neg$ and $\vee$), which is interpreted by the relation $B^{\Box}$ with **KD45** properties, as typical for doxastic operators (Blackburn et al., 2001; Meyer & van der Hoek, 1995). Since $\mathbf{B}^{\Box}$ is a necessity-type operator, its semantics would be defined as for $[\pi]$ (Definition III.3) if it were actually part of the language, i.e. $\vartheta(\mathbf{B}^{\Box}(\phi)) = \{w \mid \forall(w,w') \in B^{\Box} : (w' \in \vartheta(\phi)\})$. Note that in the proof below '||' denotes disjunction in the metalanguage.

Proposition III.1. *Let $\phi, \phi' \in \mathscr{L}_{\mathbf{M}}$ be expressions in the propositional logical subset of $\mathscr{L}_{\mathbf{M}}$, and let $\psi, \psi' \in \mathscr{L}_0$. It holds that $\not\models \mathbf{B}^{\Box}(\phi \vee \phi') \rightarrow (\mathbf{B}^{\Box}(\phi) \vee \mathbf{B}^{\Box}(\phi'))$. In contrast, for the doxastic operator in our framework it holds that $\models \mathbf{B}(\psi \vee \psi') \rightarrow (\mathbf{B}(\psi) \vee \mathbf{B}(\psi'))$.*

Proof. The first claim of our proposition can be proven with a counterexample. Take an arbitrary model $\mathfrak{M}$ for $\mathscr{L}_{\mathbf{M}}$, and let $w, w', w'' \in \mathrm{W}_{\mathfrak{M}}$ be such that $(w,w'),(w,w'') \in B^{\Box}$ and $\forall w''' \in \mathrm{W}_{\mathfrak{M}} : (w,w''') \in B^{\Box} \Rightarrow (w''' = w' \;||\; w''' = w'')$. Furthermore, let $w' \in \vartheta(p)$ and $w'' \in \vartheta(q)$, but $w' \notin \vartheta(q)$ and $w'' \notin \vartheta(p)$. It is then the case that $w \models \mathbf{B}^{\Box}(p \vee q)$ but $w \not\models \mathbf{B}^{\Box}(p)$ and $w \not\models \mathbf{B}^{\Box}(q)$, such that $w \notin \vartheta(\mathbf{B}^{\Box}(p)) \cup \vartheta(\mathbf{B}^{\Box}(q))$ and the first claim is proven. The second claim is proven by considering the valuation function ϑ_{B} which interprets arguments of the doxastic operator **B**. Syntactically this operator only accepts elements of $\mathscr{L}_0$ as its argument, so choose arbitrary $\psi, \psi' \in \mathscr{L}_0$ and consider any $w \in \mathrm{W}_{\mathfrak{M}}$ for which holds that $w \models \mathbf{B}(\psi \vee \psi')$. The semantics dictate that then $w \in \vartheta(\mathbf{B}(\psi \vee \psi'))$, i.e. $w \in \vartheta_{\mathsf{B}}(\psi \vee \psi')$, i.e. $w \in (\vartheta_{\mathsf{B}}(\psi) \cup \vartheta_{\mathsf{B}}(\psi'))$, i.e. $w \in (\vartheta(\mathbf{B}(\psi)) \cup \vartheta(\mathbf{B}(\psi')))$, i.e. $w \models \mathbf{B}(\psi) \vee \mathbf{B}(\psi')$. Since $\mathfrak{M}$ is arbitrary and so is $w \in \mathrm{W}_{\mathfrak{M}}$, the claim is proven. □

Proposition III.2. *It holds that $\models \mathbf{G}(\psi \vee \psi') \rightarrow (\mathbf{G}(\psi) \vee \mathbf{G}(\psi'))$.*

Proof. Along the lines of Proposition III.1. □

Proposition III.1 shows that our state-based semantics for the belief ascription operator is somewhat coarse in comparison to the accessibility relations typically used in epistemic (doxastic) logic; Proposition III.2 shows that likewise this is the case for the goal ascription operator. In frameworks that interpret goals or beliefs by means of accessibility relations, truth is evaluated in relation to 'possible worlds'. For instance, stating that "an agent believes that it rains or snows" means that in every world that the agent considers possible on grounds of its beliefs, it either rains or snows. This does *not* mean (in the case of a modal belief operator) that in every possible world it rains, or in every possible world it snows. The state-based operators that we employ, however, dictate that if an agent, in some state, is ascribed the belief that it rains or snows, then this flatly means that in that particular state the agent is ascribed the belief that either it rains or snows; or both. In this sense our semantics is more simplistic than the standard modal semantics, which is acceptable to us given the fact that the intended application of our framework is in realizing artificially intelligent systems. Such can be done either by reasoning directly with the logic, or by using the formalism as basis for software implementation; in either case, the state-based interpretation of ascription operators is useful because its simplicity involves restriction of the ascription operators (which cannot be nested), and translates more straightforwardly to software implementation than semantics in terms of possible worlds (cf. Chapter V, Section 3.2). Notwithstanding this fact, as the following propositions show, in the case of conjunction our language is similarly expressive to that of languages with typical belief/goal modalities.

Proposition III.3. *It holds that* $\models \mathbf{B}^{\Box}(\phi \wedge \phi') \rightarrow (\mathbf{B}^{\Box}(\phi) \wedge \mathbf{B}^{\Box}(\phi'))$. *Likewise, for the doxastic operator in our framework it holds that* $\models \mathbf{B}(\psi \wedge \psi') \rightarrow (\mathbf{B}(\psi) \wedge \mathbf{B}(\psi'))$.

Proof. Take an arbitrary model $\mathfrak{M}$ for $\mathscr{L}_{\mathbf{M}}$, and any $w \in \mathrm{W}_{\mathfrak{M}}$ such that $w \models \mathbf{B}^{\Box}(\phi \wedge \phi')$, i.e. $w \in \vartheta(\mathbf{B}^{\Box}(\phi \wedge \phi'))$, i.e. $w \in \{w'' \mid \forall (w'', w') \in B^{\Box} : w' \in \vartheta(\phi \wedge \phi')\}$. This means that $\forall (w, w') \in B^{\Box} : w' \in \vartheta(\phi \wedge \phi')$, i.e. $\forall (w, w') \in B^{\Box} : w' \in \vartheta(\phi) \cap \vartheta(\phi')$. If that is the case, then $\forall (w, w') \in B^{\Box} : w' \in \vartheta(\phi)$ and $\forall (w, w') \in B^{\Box} : w' \in \vartheta(\phi')$, i.e. $w \models \mathbf{B}^{\Box}(\phi)$ and $w \models \mathbf{B}^{\Box}(\phi')$, so that $w \models \mathbf{B}^{\Box}(\phi) \wedge \mathbf{B}^{\Box}(\phi')$. Since $\mathfrak{M}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ are arbitrary, the first claim is proven.

For the second claim, also take an arbitrary $w \in \mathrm{W}_{\mathfrak{M}}$. Let $w \models \mathbf{B}(\psi \wedge \psi')$, and note that then $w \in \vartheta_{\mathrm{B}}(\psi \wedge \psi')$, i.e. $w \in (\vartheta_{\mathrm{B}}(\psi) \cap \vartheta_{\mathrm{B}}(\psi'))$, i.e. $w \in (\vartheta(\mathbf{B}(\psi)) \cap \vartheta(\mathbf{B}(\psi')))$, i.e. $w \models \mathbf{B}(\psi) \wedge \mathbf{B}(\psi')$. Again $w \in \mathrm{W}_{\mathfrak{M}}$ is arbitrary, proving the second claim. □

Proposition III.4. *It holds that* $\models \mathbf{G}(\psi \wedge \psi') \rightarrow (\mathbf{G}(\psi) \wedge \mathbf{G}(\psi'))$.

Proof. Along the lines of Proposition III.3. □

Furthermore, the ascription operators of $\mathscr{L}_{\mathbf{M}}$ are quite expressive in another sense, as is shown by the following propositions which illustrate that the concepts of indeterminacy and paraconsistency can be expressed in terms of $\mathscr{L}_0$.

Proposition III.5. *It holds that* $\models \mathbf{B}^{\Box}(p \vee \neg p)$. *In contrast, for the doxastic operator of our framework holds* $\not\models \mathbf{B}(p \vee \sim p)$.

Proof. Take an arbitrary model $\mathfrak{M}$ for $\mathscr{L}_{\mathbf{M}}$ and any $w \in \mathrm{W}_{\mathfrak{M}}$, and consider that for $w \models \mathbf{B}^{\Box}(p \vee \neg p)$ to be the case it must hold that $w \in \vartheta(\mathbf{B}^{\Box}(p \vee \neg p))$, i.e. $w \in \{w' \mid \forall(w', w'') \in B^{\Box} : (w'' \in \vartheta(p \vee \neg p)\}$. Note that $w'' \in \vartheta(p \vee \neg p)$ if and only if $w'' \in (\vartheta(p) \cup \vartheta(\neg p))$, i.e. $w'' \in (\vartheta(p) \cup (\mathrm{W}_{\mathfrak{M}} - \vartheta(p)))$, i.e. $w'' \in \mathrm{W}_{\mathfrak{M}}$, such that $w \in \{w' \mid \forall(w', w'') \in B^{\Box} : (w'' \in \mathrm{W}_{\mathfrak{M}}\})$ must hold for the claim to be true. This is obviously the case, and since $\mathfrak{M}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ are arbitrary the claim for $\mathbf{B}^{\Box}$ is proven. To prove the second claim, consider any $w \in \mathrm{W}_{\mathfrak{M}}$ such that $w \notin \vartheta_{\mathsf{B}}(p)$ and $w \notin \vartheta_{\mathsf{B}}(\sim p)$, the existence of which is made possible by the fact that the function ϑ_{B} is defined for elements of $\{p, \sim p \mid p \in \mathsf{Atom}\}$ individually. It is then the case that $w \notin (\vartheta_{\mathsf{B}}(p) \cup \vartheta_{\mathsf{B}}(\sim p))$, such that $w \notin (\vartheta_{\mathsf{B}}(p \vee \sim p))$, such that $w \notin \vartheta(\mathbf{B}(p \vee \sim p))$ and therefore $w \not\models \mathbf{B}(p \vee \sim p)$, proving the second claim. □

Proposition III.6. *It holds that* $\not\models \mathbf{G}(p \vee \sim p)$.

Proof. Along the lines of Proposition III.5. □

Proposition III.7. *It holds that* $\models \neg\mathbf{B}^{\Box}(p \wedge \neg p)$. *In contrast, for the doxastic operator of our framework holds* $\not\models \neg\mathbf{B}(p \wedge \sim p)$.

Proof. Take an arbitrary model $\mathfrak{M}$ for $\mathscr{L}_{\mathbf{M}}$ and any $w \in \mathrm{W}_{\mathfrak{M}}$, and consider that for $w \models \neg\mathbf{B}^{\Box}(p \wedge \neg p)$ to be the case it must hold that $w \in \vartheta(\neg\mathbf{B}^{\Box}(p \wedge \neg p))$, i.e. $w \in (\mathrm{W}_{\mathfrak{M}} - \vartheta(\mathbf{B}^{\Box}(p \wedge \neg p)))$, i.e. $w \in (\mathrm{W}_{\mathfrak{M}} - \{w' \mid \forall(w', w'') \in B^{\Box} : w'' \in \vartheta(p \wedge \neg p)\}$. Note that $w'' \in \vartheta(p \wedge \neg p)$ if and only if $w'' \in \vartheta(p) \cap \vartheta(\neg p)$, i.e. $w'' \in \vartheta(p) \cap (\mathrm{W}_{\mathfrak{M}} - \vartheta(p))$, i.e. $w'' \in \emptyset$, such that $w \in \{w' \mid \forall(w', w'') \in B^{\Box} : w'' \in \emptyset\}$ must hold for the claim to be true. This is obviously never the case, and since $\mathfrak{M}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ are arbitrary the first claim is proven. To prove the second claim, consider any $w \in \mathrm{W}_{\mathfrak{M}}$ such that $w \in \vartheta_{\mathsf{B}}(p)$ and $w \in \vartheta_{\mathsf{B}}(\sim p)$, the existence of which is made possible by the fact that the function ϑ_{B} is defined for elements of $\{p, \sim p \mid p \in \mathsf{Atom}\}$ individually. It is then the case that $w \in (\vartheta_{\mathsf{B}}(p) \cap \vartheta_{\mathsf{B}}(\sim p))$, such that $w \in (\vartheta_{\mathsf{B}}(p \wedge \sim p))$, such that $w \in \vartheta(\mathbf{B}(p \wedge \sim p))$, such that $w \notin (\mathrm{W}_{\mathfrak{M}} - \vartheta(\mathbf{B}(p \wedge \sim p)))$, such that $w \notin \vartheta(\neg\mathbf{B}(p \wedge \sim p))$, and therefore $w \not\models \mathbf{B}(p \wedge \sim p)$, proving the second claim. □

Proposition III.8. *It holds that* $\not\models \neg\mathbf{G}(p \wedge \sim p)$.

Proof. Along the lines of Proposition III.7. □

Propositions III.5–III.8 show that validities which typically hold in the case of the modal doxastic operator do not hold in the case of the ascription operators with our special semantics. In the modal case an agent is assumed to believe either any atom or its literal opposite because of the classical semantics with which atoms are interpreted. Since our objective is to formalize ascription of mentalistic facts *on grounds of observation*, it is noteworthy that with indeterminacy on the level of the operators **B** and **G** it is indeed the case that ascription does not hold by definition. As Proposition III.5 shows, it can be the case

that the agent is not ascribed belief in some fact p, nor in $\sim p$. This allows for formalizing observation-based ascription, in the sense that evidence must exist — in the form of observed behavior — that warrants belief in p or $\sim p$ to be ascribed to the agent. The above illustrates this deviation of our ascription semantics from standard modal/two-valued semantics; note that this deviation is restricted to expressions from $\mathscr{L}_0$, and that the negation symbol $\neg$ in the language $\mathscr{L}$ behaves normally, as shown by the following propositions and their corollaries.

Proposition III.9. $\models \phi \vee \neg\phi$.

Proof. Take any model $\mathfrak{M}$, any $w \in \mathrm{W}_{\mathfrak{M}}$, and any $\phi \in \mathscr{L}_{\mathbf{M}}$, and see that if $w \models \phi \vee \neg\phi$ is to be true, then it must hold that $w \in (\vartheta(\phi) \cup \vartheta(\neg\phi))$, i.e. $w \in (\vartheta(\phi) \cup (\mathrm{W}_{\mathfrak{M}} - \vartheta(\phi)))$, i.e. $w \in \mathrm{W}_{\mathfrak{M}}$. This is obviously the case, and since $\mathfrak{M}$, $w \in \mathrm{W}_{\mathfrak{M}}$, and $\phi \in \mathscr{L}_{\mathbf{M}}$ are arbitrary, the claim is proven. □

III.1

Corollary III.1. $\models \mathbf{B}(\psi) \vee \neg\mathbf{B}(\psi)$ *and* $\models \mathbf{G}(\psi) \vee \neg\mathbf{G}(\psi)$.

Proposition III.10. $\not\models \phi \wedge \neg\phi$.

Proof. Take any model $\mathfrak{M}$, any $w \in \mathrm{W}_{\mathfrak{M}}$, and any $\phi \in \mathscr{L}_{\mathbf{M}}$, and see that if $w \models \phi \wedge \neg\phi$ is to be true, then it must hold that $w \in (\vartheta(\phi) \cap \vartheta(\neg\phi))$, i.e. $w \in (\vartheta(\phi) \cap (\mathrm{W}_{\mathfrak{M}} - \vartheta(\phi)))$, i.e. $w \in \emptyset$. This is never the case, and since $\mathfrak{M}$, $w \in \mathrm{W}_{\mathfrak{M}}$, and $\phi \in \mathscr{L}_{\mathbf{M}}$ are arbitrary, the claim is proven. □

Corollary III.2. $\not\models \mathbf{B}(\psi) \wedge \neg\mathbf{B}(\psi)$ *and* $\not\models \mathbf{G}(\psi) \wedge \neg\mathbf{G}(\psi)$.

To wrap up this section, the following proposition complements Propositions III.1–III.4.

Proposition III.11. *The following holds for any* $\psi, \psi' \in \mathscr{L}_0$.

$$\models \mathbf{B}(\psi \vee \psi') \leftrightarrow (\mathbf{B}(\psi) \vee \mathbf{B}(\psi'))$$
$$\models \mathbf{B}(\psi \wedge \psi') \leftrightarrow (\mathbf{B}(\psi) \wedge \mathbf{B}(\psi'))$$
$$\models \mathbf{G}(\psi \vee \psi') \leftrightarrow (\mathbf{G}(\psi) \vee \mathbf{G}(\psi'))$$
$$\models \mathbf{G}(\psi \wedge \psi') \leftrightarrow (\mathbf{G}(\psi) \wedge \mathbf{G}(\psi'))$$

Proof. Propositions III.1–III.4 already showed that the above statements hold for the $\rightarrow$ direction of $\leftrightarrow$, so it remains to show that they hold in the $\leftarrow$ direction as well. This can be shown straightforwardly by considering the proofs of the aforementioned propositions, which show that, given an operator $\mathbf{Q} \in \{\mathbf{B}, \mathbf{G}\}$ and $\vartheta_{\mathrm{Q}} \in \{\vartheta_{\mathrm{B}}, \vartheta_{\mathrm{G}}\}$ as its corresponding valuation function, it holds for any model $\mathfrak{M}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ that from $w \models \mathbf{Q}(\psi) \vee \mathbf{Q}(\psi')$ follows $w \in (\vartheta(\mathbf{Q}(\psi)) \cup \vartheta(\mathbf{Q}(\psi')))$, i.e. $w \in (\vartheta_{\mathrm{Q}}(\psi) \cup \vartheta_{\mathrm{Q}}(\psi'))$, i.e. $w \in \vartheta_{\mathrm{Q}}(\psi \vee \psi')$, i.e. $w \in \vartheta(\mathbf{Q}(\psi \vee \psi'))$, i.e. $w \models \mathbf{Q}(\psi \vee \psi')$. Likewise, from $w \models \mathbf{Q}(\psi) \wedge \mathbf{Q}(\psi')$ follows $w \in (\vartheta(\mathbf{Q}(\psi)) \cap \vartheta(\mathbf{Q}(\psi')))$, i.e. $w \in (\vartheta_{\mathrm{Q}}(\psi) \cap \vartheta_{\mathrm{Q}}(\psi'))$, i.e. $w \in \vartheta_{\mathrm{Q}}(\psi \wedge \psi')$, i.e. $w \in \vartheta(\mathbf{Q}(\psi \wedge \psi'))$, i.e. $w \models \mathbf{Q}(\psi \wedge \psi')$. □

It was stated in Section 1.1 that the formalism used here is somewhat simplistic in comparison to BDI logics (Cohen & Levesque, 1990; Rao & Georgeff, 1991) or agent logics such as KARO (van der Hoek et al., 1998), and in the current section this claim was given formal support. This simplicity lies in the fact that the doxastic and motivational operators of $\mathscr{L}_{\mathbf{M}}$ do not have modal semantics, and in this sense our approach is similar to recent approaches to reasoning about the behavior of agents (Alechina et al., 2007, 2010). A notable difference between such approaches and ours is the fact that these other approaches focus on proving properties of an agent's behavior given particular mental states, whereas in our case focus is on attribution of mental states given particular behavior. In this sense our semantics of ascription is reasonable, justified by the rationale given in Section 1.1 and at the beginning of the current section. The next chapter provides an additional argument in favor of our choice to consider doxastic and motivational operators as 'markings' of propositional expressions, as opposed to providing them with modal semantics. It should furthermore be noted that ascription could be forced to allow only indeterminacy but not paraconsistency (e.g. by constraining models to not allow both $\mathbf{B}(p)$ and $\mathbf{B}(\sim p)$ to be true simultaneously, and likewise for $\mathbf{G}$), or the other way around, or to allow for neither.

1.3 Definitions

In Section 1.1 the syntactic primitives of the language $\mathscr{L}_{\mathbf{M}}$ were put forward, and their semantics was given in Section 1.2. In the current section those primitives are used to define several syntactic shorthands, which allow for expressing certain notions more succinctly than can be done using only the primitives themselves.

1.3.1 Ascription

In Section 1.2 the semantics of $\mathscr{L}_{\mathbf{M}}$ were defined, with Section 1.2.1 focusing specifically on the semantics of ascription by means of $\mathscr{L}_0$, the language that is accepted by the operators $\mathbf{B}$ and $\mathbf{G}$ for ascription of beliefs and goals, respectively. It is common in logics with many-valued semantics that apart from the classical truth values 'true' and 'false' others are defined, typically the values 'unknown' and/or 'both' (Stanford Encyclopedia of Philosophy, 2009). The framework employed in this chapter does not have those values as semantic primitives, but does allow for their definition in terms of constructs which are primitive. Because interest here is in the ascription of beliefs and goals to an observed agent, the definitions given below focus on the *conclusiveness of ascription* in relation to literals instead of their 'truth'. For notational convenience, a superscript 'inversion' operator is defined that accepts literals in $\mathsf{Lit} = \{p, \sim p \mid p \in \mathsf{Atom}\}$, such that for any $p, \sim p \in \mathsf{Lit}$ holds $\overline{p} = \sim p$ and $\overline{\sim p} = p$.

Definition III.4 (conclusiveness of ascription)**.** *Let* $\mathsf{Lit} = \{p, \sim p \mid p \in \mathsf{Atom}\}$ *be the literals*

of $\mathscr{L}_0$, and let $\psi \in \mathsf{Lit}$ be any literal.

$$\begin{aligned}
\mathbf{BConc}(\psi) &\triangleq \mathbf{B}(\psi) \wedge \neg\mathbf{B}(\overline{\psi}) \\
\mathbf{BInc}(\psi) &\triangleq \mathbf{B}(\psi) \wedge \mathbf{B}(\overline{\psi}) \\
\mathbf{BUnk}(\psi) &\triangleq \neg\mathbf{B}(\psi) \wedge \neg\mathbf{B}(\overline{\psi}) \\
\mathbf{GConc}(\psi) &\triangleq \mathbf{G}(\psi) \wedge \neg\mathbf{G}(\overline{\psi}) \\
\mathbf{GInc}(\psi) &\triangleq \mathbf{G}(\psi) \wedge \mathbf{G}(\overline{\psi}) \\
\mathbf{GUnk}(\psi) &\triangleq \neg\mathbf{G}(\psi) \wedge \neg\mathbf{G}(\overline{\psi})
\end{aligned}$$

The interpretation of the above expressions is quite straightforward: $\mathbf{BConc}(\psi)$ denotes that the observer ascribes the literal ψ *conclusively* as belief to the observed agent, $\mathbf{BInc}(\psi)$ that it does so *inconclusively*, and $\mathbf{BUnk}(\psi)$ states that ascription is *unknown* which means the observer ascribes ψ nor its complementary literal to the agent. Because of the state-based interpretation of the operators $\mathbf{B}$ and $\mathbf{G}$, the above definitions can be straightforwardly extended to also allow for conjoint or disjoint formulae: $\mathbf{Q}(\psi \wedge \psi') \triangleq \mathbf{Q}(\psi) \wedge \mathbf{Q}(\psi')$ and $\mathbf{Q}(\psi \vee \psi') \triangleq \mathbf{Q}(\psi) \vee \mathbf{Q}(\psi')$, where $\mathbf{Q} \in \{\mathbf{BConc}, \mathbf{BInc}, \mathbf{BUnk}, \mathbf{GConc}, \mathbf{GInc}, \mathbf{GUnk}\}$. Note that some of the expressions which can thus be formed 'make no sense'; that is to say, e.g. $\mathbf{BConc}(p \wedge {\sim}p)$ cannot be true, because it is equivalent to $\vartheta((\mathbf{B}(p) \wedge \neg\mathbf{B}({\sim}p)) \wedge (\mathbf{B}({\sim}p) \wedge \neg\mathbf{B}(p))) = \emptyset$. In other words, it is never the case that the observer conclusively attributes both p and ${\sim}p$ as beliefs (or goals, for that matter) to an agent.

1.3.2 Action and Observation

This section focuses on shorthand expressions relating to actions and their observation. PDL-based approaches typically employ propositions of the type $\mathbf{Done}(\alpha) \in \mathsf{Atom}$ in relation to actions $\alpha \in \mathsf{Act}$, to state that a particular action in a particular state is considered 'done'. Because our logical framework is used for modeling an observer that reasons about the behavior of an agent and the actions in Act are those of the agent, the interpretation of expressions describing particular actions as 'done' is intrinsically related to those actions' underlying accessibility relations (i.e. the valuation of atoms $\mathbf{Done}(\alpha)$ by ϑ_{I} is directly tied to R_α). Specifically, the accessibility relation for some action relates two states if, and only if, in the resulting state this action is considered done. This leads to the following formal definition.

Definition III.5 (actions considered 'done')**.** *Let* Act *be a set of primitive actions such that R_α is the accessibility relation of action $\alpha \in \mathsf{Act}$, and let $\mathfrak{M}$ be any model for $\mathscr{L}_{\mathbf{M}}$. The following then holds with regard to the expression* $\mathbf{Done}(\alpha)$*.*

$$\forall \alpha \in \mathsf{Act} \forall w \in \mathbb{W}_{\mathfrak{M}} : \quad (\exists w' \in \mathbb{W}_{\mathfrak{M}} : (w', w) \in R_\alpha) \iff (w \in \vartheta(\mathbf{Done}(\alpha)))$$

In order to be able to talk easily about sequences of actions in our logical language, '**Done**' is defined to extend to elements of the language $\mathscr{L}_\Delta$ put forward in Definition II.6. As stated

previously $\mathscr{L}_{\Delta} \subseteq \mathscr{L}_{\Pi}$ so that this is feasible. It should once more be noted that models of the type defined in Section 1.2 represent the observer's 'view' on some state of affairs, and $\alpha \in \mathsf{Act}$ pertains to actions *of the observed agent*.

$$\mathbf{Done}(\alpha_1;\ldots;\alpha_n) \triangleq \mathbf{Done}(\alpha_n) \wedge \langle \alpha_n{}^{-}\rangle(\mathbf{Done}(\alpha_{n-1}) \wedge \langle \alpha_{n-1}{}^{-}\rangle(\mathbf{Done}(\alpha_{n-2}) \wedge \ldots \wedge \langle \alpha_2{}^{-}\rangle(\mathbf{Done}(\alpha_1))\ldots))$$

The above definition has an existential connotation, in the sense that if a sequence of actions is considered 'done' in some state then there must exist some notion of a past state to which this sequence leads (as shown by Proposition III.14). In addition to allowing us to express that an action has been done, the logical framework presented in Sections 1.1 and 1.2 offers the operator '**Obs**' which accepts actions and expresses the fact that some action has been observed, and which can be extended to sequences in a similar fashion as done with '**Done**'. In line with Chapter II it is hereby assumed that the observer in any state has no doubt about which actions it has observed. There, this assumption translated into the fact that observation was expressed with an instance of the predicate $\mathsf{seen}/2$ that has a unique 'position' argument. In the current chapter the focus is on dynamics, and the definition of observation of sequences of actions take this focus into account along the lines of the definition of '**Done**', as follows.

$$\mathbf{Obs}(\alpha_1;\ldots;\alpha_n) \triangleq \mathbf{Obs}(\alpha_n) \wedge \langle \alpha_n{}^{-}\rangle(\mathbf{Obs}(\alpha_{n-1}) \wedge \langle \alpha_{n-1}{}^{-}\rangle(\mathbf{Obs}(\alpha_{n-2}) \wedge \ldots \wedge \langle \alpha_2{}^{-}\rangle(\mathbf{Obs}(\alpha_1))\ldots))$$

It is noteworthy that the operator **Obs** is interpreted by means of the valuation function ϑ_{O}, which allows us to 'decouple' it from the action accessibility relations in order to express actions which the observer considers to have been done by the agent but which it has not observed. One could say that the observer has 'missed' such an action, and this is defined as follows.

$$\mathbf{Missed}(\alpha) \triangleq \mathbf{Done}(\alpha) \wedge \neg\mathbf{Obs}(\alpha)$$

Again, this expression can be extended to sequences of actions, denoting that the observer has missed a sequence of consecutive actions $\alpha_1;\ldots;\alpha_n$.

$$\mathbf{Missed}(\alpha_1;\ldots;\alpha_n) \triangleq \mathbf{Missed}(\alpha_n) \wedge \langle \alpha_n{}^{-}\rangle(\mathbf{Missed}(\alpha_{n-1}) \wedge \langle \alpha_{n-1}{}^{-}\rangle(\mathbf{Missed}(\alpha_{n-2}) \wedge \ldots \wedge \langle \alpha_2{}^{-}\rangle(\mathbf{Missed}(\alpha_1))\ldots))$$

1.4 Basic Properties

The interpretation of $\mathbf{Done}(\alpha)$ in Definition III.5 leads to the following properties.

Proposition III.12. *Let $\alpha \in \mathsf{Act}$ be a primitive action. It then holds that*

$$\models [\alpha]\mathbf{Done}(\alpha)$$

Proof. Take any model $\mathfrak{M}$ for $\mathscr{L}_{\mathbf{M}}$, any $w \in \mathrm{W}_{\mathfrak{M}}$ and $\alpha \in \mathsf{Act}$, and note from the semantics that $w \models [\alpha]\mathbf{Done}(\alpha)$ is the case if $w \in \{w'' \mid \forall(w'', w') \in \varrho(\alpha) : w' \in \vartheta(\mathbf{Done}(\alpha))\}$. Observe that $\varrho(\alpha) = R_\alpha$, and that from Definition III.5 follows that for any $(w'', w') \in R_\alpha$ holds $w' \in \vartheta(\mathbf{Done}(\alpha))$. Thus, indeed $w \in \{w'' \mid \forall(w'', w') \in \varrho(\alpha) : w' \in \vartheta(\mathbf{Done}(\alpha))\}$ and, since α, $\mathfrak{M}$, and w are arbitrary, the proposition is proven. □

Proposition III.13. *Let $\alpha \in \mathsf{Act}$ be a primitive action. It then holds that*

$$\models \mathbf{Done}(\alpha) \leftrightarrow \langle \alpha^- \rangle(\top)$$

Proof. Take any action $\alpha \in \mathsf{Act}$, model $\mathfrak{M}$ and state $w \in \mathrm{W}_{\mathfrak{M}}$, and let $w \models \mathbf{Done}(\alpha)$. Semantics dictate that then $w \in \vartheta(\mathbf{Done}(\alpha))$, and from Definition III.5 then follows $\exists w' : (w', w) \in R_\alpha$, i.e. $\exists w' : (w, w') \in R_\alpha^{-1}$. Take any $(w, w') \in R_\alpha^{-1}$ and observe that $w \models \langle \alpha^- \rangle(\top)$ is equivalent to $w \models \neg[\alpha^-]\neg(\top)$, which is true if $w \in \vartheta(\neg[\alpha^-]\neg(\top))$, i.e. if $w \in (\mathrm{W}_{\mathfrak{M}} - \vartheta([\alpha^-]\neg(\top)))$, i.e. if $w \in (\mathrm{W} - \vartheta([\alpha^-](\bot)))$, i.e. if $w \in (\mathrm{W}_{\mathfrak{M}} - \{w'' \mid \forall(w'', w''') \in R_\alpha^{-1} : w''' \in \vartheta(\bot)\})$, i.e. if $w \in (\mathrm{W}_{\mathfrak{M}} - \{w'' \mid \forall(w'', w''') \in R_\alpha^{-1} : (w''' \in \emptyset)\})$. Let $Z = \{w'' \mid \forall(w'', w''') \in R_\alpha^{-1} : (w''' \in \emptyset)\}$, observing that it is known that $(w, w') \in R_\alpha^{-1}$ but that $w' \notin \emptyset$. Thus, $w \notin Z$, so that it must be the case that $w \in \mathrm{W}_{\mathfrak{M}} - Z$, and therefore $w \models \langle \alpha^- \rangle(\top)$. Note that α, $\mathfrak{M}$, and w are arbitrary, so that the proposition is proven for '$\rightarrow$'. Assuming $w \models \langle \alpha^- \rangle(\top)$ and retracing the steps of the above proof yields the result for '$\leftarrow$'. □

This property translates to the extension of '**Done**' to sequences of actions, as shown in the following proposition.

Proposition III.14. *Let $\alpha_1, \ldots, \alpha_n \in \mathsf{Act}$ be primitive actions. It then holds that*

$$\models \mathbf{Done}(\alpha_1; \ldots; \alpha_n) \leftrightarrow \langle \alpha_1; \ldots; \alpha_n{}^- \rangle(\top)$$

Proof. Considering the definition of '**Done**' with respect to sequences, as presented in Section 1.3, it follows on grounds of Proposition III.13 that $\models \mathbf{Done}(\alpha_1; \ldots; \alpha_n) \leftrightarrow \langle \alpha_n{}^- \rangle(\top \wedge \langle \alpha_{n-1}{}^- \rangle(\top \wedge \ldots \wedge \langle \alpha_1{}^- \rangle(\top) \ldots))$, i.e. $\models \mathbf{Done}(\alpha_1; \ldots; \alpha_n) \leftrightarrow \langle \alpha_1; \ldots; \alpha_n{}^- \rangle(\top)$. □

In Section 3.2 of Chapter II it was stated in regard to the approach using logical abduction that the observer, *if* it perceives the actions of an agent, it perceives them sequentially and non-concurrently. It is therefore reasonable to require in this dynamic framework as well that only a single action can lead to a single state, such that in any such state only a single action is considered 'done'; as opposed to multiple actions being considered 'done' simultaneously. This assumption is reflected as follows.

Definition III.6 (strict sequentiality of actions)**.** *It holds for any model $\mathfrak{M}$ for the language $\mathscr{L}_{\mathbf{M}}$ and any $w \in \mathrm{W}_{\mathfrak{M}}$ that*

$$\forall \alpha, \alpha' \in \mathsf{Act} : \qquad \mathfrak{M}, w \models \mathbf{Done}(\alpha) \wedge \mathbf{Done}(\alpha') \qquad \Longrightarrow \qquad (\alpha = \alpha')$$

The semantics of $\mathscr{L}_{\mathbf{M}}$, as given in Section 1.2, define that observation of actions is formalized through the valuation function ϑ_{O} which interprets expressions of the form $\mathbf{Obs}(\alpha)$ for actions $\alpha \in \mathsf{Act}$. This function assigns sets of states to actions, and technically it is decoupled from atoms of the form $\mathbf{Done}(\alpha)$. Nevertheless, it seems reasonable to suppose a relation to exist between the observation and 'doneness' of actions; specifically, it does not seem reasonable to maintain the possibility that an action is taken to be observed but not done, which is reflected in the following definition.

Definition III.7 (observed actions are considered 'done')**.** *Let* Act *be a set of primitive actions and* $\mathfrak{M}$ *any model for the language* $\mathscr{L}_{\mathbf{M}}$*. It then holds that*

$$\forall \alpha \in \mathsf{Act} \forall w \in \mathrm{W}_{\mathfrak{M}} : \qquad w \in \vartheta_{\mathsf{O}}(\alpha) \implies w \in \vartheta_{\mathsf{I}}(\mathbf{Done}(\alpha))$$

Proposition III.15. *Let* $\alpha \in \mathsf{Act}$ *be a primitive action. It then holds that*

$$\models \mathbf{Obs}(\alpha) \rightarrow \mathbf{Done}(\alpha)$$

Proof. Take any model $\mathfrak{M}$ and state $w \in \mathrm{W}_{\mathfrak{M}}$. Assume that $w \models \mathbf{Obs}(\alpha)$, i.e. $w \in \vartheta(\mathbf{Obs}(\alpha))$, i.e. $w \in \vartheta_{\mathsf{O}}(\alpha)$. From Definition III.7 then follows that for $\mathbf{Done}(\alpha) \in \mathsf{Atom}$ holds that $w \in \vartheta_{\mathsf{I}}(\mathbf{Done}(\alpha))$, i.e. $w \in \vartheta(\mathbf{Done}(\alpha))$, i.e. $w \models \mathbf{Done}(\alpha)$. Observe that $\mathfrak{M}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ are arbitrary, so that the claim is proven. □

Furthermore, the following then holds.

Proposition III.16. *Let* $\alpha_1, \ldots, \alpha_n \in \mathsf{Act}$ *be primitive actions. It then holds that*

$$\models \mathbf{Obs}(\alpha_1; \ldots; \alpha_n) \rightarrow \langle \alpha_1; \ldots; \alpha_n{}^{-} \rangle (\top)$$

Proof. This follows straightforwardly from Propositions III.15 and III.14. □

2 Dynamics in Mental State Abduction

In Chapter II, mental state abduction was presented as an approach to explanation of BDI-based agents' observed behavior on grounds of observed actions and knowledge of their behavioral rules. A family of functions was defined (cf. Definition II.26) that group together rule preconditions which can be defeasibly inferred as a result of explaining the observed behavior of an agent under different perceptory conditions. Those functions were given both credulous and skeptical interpretations in terms of a ground fragment of predicate logic, focusing on entailment of the inferred rule preconditions in abductive extensions. Because classical logic — such as that used to formalize the abductive approach in the previous chapter — is not very well suited for formalizing dynamics, the current chapter employs the language $\mathscr{L}_{\mathbf{M}}$ to give various explanatory interpretations of agents' observed behavior, focusing on dynamics. Such an approach has the benefit that reasoning with actions or more complex dynamic expressions can be concisely expressed by means of modalities. Also, the fact that a special operator is used to specify action observation allows for elegant expression of reasoning about dynamics under conditions of incomplete observation.

2.1 Rules and Plans Revisited

The interpretations presented in this chapter are the dynamic counterparts of the interpretations of the msa_X functions given in Theorems II.2 and II.3. Arriving at those interpretations is somewhat easier than in the previous chapter, because the formal machinery employed in that chapter is compatible with the dynamic logic employed here. Before proceeding, however, a missing piece of that machinery must be put into place — something which was omitted in Section 3.3 of Chapter II because the language $\mathscr{L}_0$ was not yet defined: the function for translation of MYAPL ⟨*testaction*⟩-elements (cf. Definition II.3). To this extent the function τ_q is here given, which translates queries in the MYAPL language to $\mathscr{L}_0$, as follows.

Definition III.8 (MYAPL query translation function τ_q)**.** *Let ϕ, ϕ' be any* MYAPL ⟨*query*⟩ *or* ⟨*goalquery*⟩ *element, and* Atom *a set of atoms from which this element is composed. The function τ_q then maps such query elements to $\mathscr{L}_0$, as follows.*

$$
\begin{aligned}
\tau_q(p) &= p \quad \textit{if} \quad p \in \mathsf{Atom} \\
\tau_q(-p) &= {\sim}p \quad \textit{if} \quad p \in \mathsf{Atom} \\
\tau_q(\phi \text{ or } \phi') &= \tau_q(\phi) \vee \tau_q(\phi') \\
\tau_q(\phi \text{ and } \phi') &= \tau_q(\phi) \wedge \tau_q(\phi')
\end{aligned}
$$

The plan translation function τ_p of Definition II.5 is then finalized, and repeated here once more in full because it occurs frequently in this chapter; albeit 'under the hood'.

$$
\begin{aligned}
\tau_p(\alpha) &= \alpha \quad \textit{if} \quad \alpha \in \mathsf{Act} \\
\tau_p(\mathsf{B}(\phi)) &= \mathbf{B}(\tau_q(\phi))? \\
\tau_p(\mathsf{G}(\phi)) &= \mathbf{G}(\tau_q(\phi))? \\
\tau_p(\pi;\pi') &= \tau_p(\pi);\tau_p(\pi') \\
\tau_p(\text{if } \phi \text{ then } \{\pi\} \text{ else } \{\pi'\}) &= (\tau_p(\phi);\tau_p(\pi)) + (\neg\tau_p(\phi);\tau_p(\pi')) \\
\tau_p(\text{while } \phi \text{ do } \{\pi\}) &= (\tau_p(\phi);\tau_p(\pi))^*;\neg\tau_p(\phi)
\end{aligned}
$$

Note that the fact that the scope of negation '$\sim$' in $\mathscr{L}_0$ only ranges over atoms is not problematic in translation using τ_q, because the scope of MYAPL negation '–' also ranges over atoms. This occurs without lack of generality, as the translation function τ_q can be applied to (propositional) languages with negation that scopes over compound expressions by bringing those expressions into a conjunctive or disjunctive normal form. The functions which translate MYAPL programming rules into their formal counterpart are now all defined. Because it is easier to operate directly on the logical expressions and it is somewhat tedious and superfluous to mention the appropriate translation functions every time, it is assumed throughout this chapter that programming rules are specified in a formal language. Specifically, it is the case that if $\mathscr{R}$ is said to be a set of MYAPL rules, this generally means

(unless clearly otherwise) that $\mathscr{R} = \{(n : \tau_q(\gamma) <\!\!- \tau_q(\beta) \mid \tau_p(\pi)) \mid (n : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R}'\}$, where $\mathscr{R}'$ is the corresponding set of rules in MYAPL syntax. In fact, in this chapter we take the liberty to let plans be given by any $\pi \in \mathscr{L}_\Pi$. Naturally, the functions msa_X defined in Definition II.26 are assumed to operate accordingly on the translation of π.

The nonmonotonic interpretations of the mental state abduction functions given in Chapter II (cf. Theorems II.2 and II.3) show that the goal/belief pairs in the set that is the output of those functions can be given a defeasible (skeptical or credulous) interpretation. It is noteworthy that this is taken to be the observer's interpretation, such that a more imaginative reading is that the observer defeasibly *ascribes* (or *attributes*) the aforementioned goals and beliefs to the agent. However, if one focuses on the dynamics of the matter, then this interpretation is found lacking. The reason for this is that the goal/belief pairs attributed by the observer — of which there may exist multiple in the output of the msa_X functions, and which may be conflicting — should not be regarded as the agent's current mental state, but as the mental state which the observer presumes it to have possibly had when it selected the rule on grounds of which that pair is inferred. More specifically: the inferred goal/belief pairs should be regarded as having possibly been entailed by the agent's mental state in the state preceding some sequence of actions to which observed actions are related. The simplest case in this regard is that of complete observation, because there the inferred mental state should be ascribed to the agent in a state preceding exactly that particular sequence which the observer assumes it has completely observed.

2.2 Ascription in the Case of Complete Observation

The dynamic interpretations of observed behavior which are presented in this chapter, are classes of models that are based on the mental state abduction functions of Chapter II. In modal logics, such as PDL, models are relational structures (frames) of interconnected states, along with an assignment of atomic propositions to states (valuation) which is *contingent*. Models are the principal semantic unit of such logics, and expressions from the logical language are evaluated with regard to models. This evaluation can occur in different ways: formulae can be evaluated locally (with respect to specific states), globally across a model (with respect to all states in the model), and globally across models (with respect to all states of all models). This gives rise to notions of modal 'truth' on different levels; yet, in the words of Blackburn et al. (2001): "statements only deserve the description 'logical' if they are *invariant* under changes of contingent information" (see that work for a thorough exposition on the notion of truth in modal logic). In this chapter our focus is on providing interpretations which are logically *valid*, in the sense that they are invariant under changes of contingent information in models. Accordingly, we are looking for classes of models in which certain things are invariably true, irrespective of any contingent truths; of course those models are such that they in any case adhere to the basic properties stated in Section 1.

In line with the fact that in Chapter II the logical theories are based on some set of MYAPL rules the observer knows the agent to have, the classes of models defined in this chapter are also based on a set of MYAPL rules. Before proceeding to realize our intended interpretations, it seems pertinent to clarify the conceptual rationale behind our approach. Most of all, it should be stressed that the notion of action explanation has both a dynamic

and a static connotation. On the one hand, actions convey dynamics, as their occurrence can result in change of state. On the other hand, the observation of (sequences of) actions can be referred to statically, as it is the case that a particular (sequence of) action(s) is — or is not — observed *at a particular moment*. PDL models allow for capturing both those qualities, in the sense that they can be used to represent the observer's perspective as a 'snapshot' of the way things evolved in relation to actions of the agent, and individual states can be regarded as possible 'moments'. The following then gives the class of models that represents the basic dynamic interpretation of the agent's behavior under the assumption of complete observation.

Definition III.9 (basic dynamic interpretation — complete observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\mathscr{L}_\Delta$ the percept language of Definition II.6, msa_C the mental state abduction function defined Definition II.26, and $\mathscr{M}$ the class of all models for $\mathscr{L}_{\mathbf{M}}$. $\mathscr{D}_C^{\mathscr{R}}$ is then the class of models reflecting the basic dynamic interpretation of mental state abduction with respect to $\mathscr{R}$, under the assumption of complete observation, as follows.*

III.2

$$\mathscr{D}_C^{\mathscr{R}} = \{\mathfrak{M} \in \mathscr{M} \mid \exists \delta \in \mathscr{L}_\Delta \exists (\gamma, \beta) \in \mathsf{msa}_C(\delta, \mathscr{R}) \exists w \in \mathrm{W}_{\mathfrak{M}} : \mathfrak{M}, w \models \mathbf{Obs}(\delta) \wedge \langle \delta^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\}$$

Each model in the class $\mathscr{D}_C^{\mathscr{R}}$ thus represents an interpretation of (at least) some observed sequence of actions and some corresponding initial mental state (i.e. goal/belief pair that is the precondition of some rule) that is attributed to the agent on grounds of that particular observed sequence of actions. It is noteworthy that it is required of any model $\mathfrak{M}$ in this class that it actually interprets some observed sequence δ (i.e. $\exists w \in \mathrm{W}_{\mathfrak{M}} : (w \models \mathbf{Obs}(\delta))$), for which the output of the mental state abduction is non-empty (i.e. $\exists (\gamma, \beta) \in \mathsf{msa}_C(\delta, \mathscr{R})$), in relation to some goal/belief pair γ, β (i.e. $w \models \langle \delta^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$). This is considered a notion of *relevance* of the selected models.

The class $\mathscr{D}_C^{\mathscr{R}}$ in itself is not of primary interest to us; its contribution lies mostly in grouping together models that are in different ways relevant to a particular set of rules. In doing so, it encompasses subclasses which are of specific interest, such as the class $\mathscr{D}_C^{\mathsf{sk}}$ that represents the dynamic counterpart of the skeptical interpretation of mental state abduction under complete observation. It should be noted that in the remainder of this chapter the set of $\mathscr{R}$ underlying subclasses of the basic class $\mathscr{D}_C^{\mathscr{R}}$ is omitted from notation in '$\mathscr{D}_C^{\mathsf{sk}}$' for the sake of conciseness.

Definition III.10 (skeptical dynamic interpretation — complete observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\mathscr{L}_\Delta$ the percept language of Definition II.6, msa_C the mental state abduction function defined in Definition II.26, and $\mathscr{D}_C^{\mathscr{R}}$ the class of models defined in Definition III.9 with respect to $\mathscr{R}$. The class of models $\mathscr{D}_C^{\mathsf{sk}}$ reflecting the skeptical dynamic interpretation of mental state abduction is then as follows.*

$$\mathscr{D}_C^{\mathsf{sk}} = \{\mathfrak{M} \in \mathscr{D}_C^{\mathscr{R}} \mid \forall \delta \in \mathscr{L}_\Delta : \quad \mathsf{msa}_C(\delta, \mathscr{R}) = \{(\gamma_1, \beta_1), \ldots, (\gamma_n, \beta_n)\} \neq \emptyset \implies \mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow [\delta^-]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_n) \wedge \mathbf{B}(\beta_n)))\}$$

It follows straightforwardly from the definition of $\mathscr{D}_C^{\mathsf{sk}}$ that $\mathscr{D}_C^{\mathsf{sk}} \subseteq \mathscr{D}_C^{\mathscr{R}}$. Furthermore, it holds that $\mathscr{D}_C^{\mathsf{sk}}$ is non-trivial, in the sense that its 'existence' is directly tied to $\mathscr{D}_C^{\mathscr{R}}$, as shown in the following proposition.

Proposition III.17. *Let $\mathscr{R}$ be a set of* MyAPL *rules, and $\mathscr{D}_C^{\mathscr{R}}$ and $\mathscr{D}_C^{\mathsf{sk}}$ as defined in Definitions III.9 and III.10 with respect to those rules, respectively. It then holds that*

$$(\mathscr{D}_C^{\mathscr{R}} \neq \emptyset) \qquad \Longleftrightarrow \qquad (\mathscr{D}_C^{\mathsf{sk}} \neq \emptyset)$$

Proof. ($\Rightarrow$) From Definition III.9 it follows that for some $\mathfrak{M} \in \mathscr{D}_C^{\mathscr{R}}$ to exist, it must be the case that $\mathsf{msa}_C(\delta, \mathscr{R}) \neq \emptyset$ for some $\delta \in \mathscr{L}_\Delta$. Take any single action $\alpha \in \mathscr{L}_\Delta$ such that $\mathsf{msa}_C(\alpha, \mathscr{R}) \neq \emptyset$, and choose $\gamma, \beta \in \mathscr{L}_0$ such that $(\gamma, \beta) \in \mathsf{msa}_C(\alpha, \mathscr{R})$. Let $\mathfrak{M}$ be the following model, such that $w' \models \mathbf{Obs}(\alpha)$ and $w \models \mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)$.

$$w \xrightarrow[\alpha]{} w'$$

Consider the criteria of Definition III.9, and observe that $\mathfrak{M} \in \mathscr{D}_C^{\mathscr{R}}$. Furthermore, see that $\forall\{(\gamma_1, \beta_1), \ldots, (\gamma_m, \beta_m)\} \subseteq \mathsf{msa}_C(\alpha, \mathscr{R}) : (\mathfrak{M} \models \mathbf{Obs}(\alpha) \rightarrow [\alpha^-]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_m) \wedge \mathbf{B}(\beta_m))))$ by disjunctive weakening, so that $\mathfrak{M}' \in \mathscr{D}_C^{\mathsf{sk}}$ and $\mathscr{D}_C^{\mathsf{sk}} \neq \emptyset$.

($\Leftarrow$) Straightforward, as $\mathscr{D}_C^{\mathsf{sk}} \subseteq \mathscr{D}_C$. □

By contraposition, it follows that $(\mathscr{D}_C^{\mathscr{R}} = \emptyset) \Longleftrightarrow (\mathscr{D}_C^{\mathsf{sk}} = \emptyset)$. Note that the model which is used in the proof of Proposition III.17 is not coincidental or dependent on the specific set of rules; a model of this type can always be found if $\mathsf{msa}_C(\delta, \mathscr{R}) \neq \emptyset$ for some $\delta \in \mathscr{L}_\Delta$, and the above proof will be referred to in subsequent proofs of the same sort, illustrating that the characterizing implication is not only satisfied by default.

It was stated that the class of models $\mathscr{D}_C^{\mathsf{sk}}$ defined in Definition III.10 is the dynamic counterpart of the skeptical interpretation of the mental state abduction functions, as given in Theorem II.2. Specifically, it holds that the disjunction over the abduced mental states which in Theorem II.2 was shown to be satisfied by all extensions, in regard to some particular observed sequence, is satisfied by all states preceding that sequence in the dynamic models. This can be seen formally as follows, where for technical simplicity it is assumed that the functions τ_g and τ_b accept elements from $\mathscr{L}_0$.

Theorem III.1. *Let $\mathscr{R}$ be a set of* MyAPL *rules, $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ the abductive theory based on those rules as defined in Definition II.21, and $\mathscr{D}_C^{\mathsf{sk}}$ the class of models defined in Definition III.10 with respect to those rules. It then holds that*

$$\begin{aligned}
&\forall \alpha_1; \ldots; \alpha_n \in \mathscr{L}_\Delta \forall \gamma_1, \ldots, \gamma_m, \beta_1, \ldots, \beta_m \in \mathscr{L}_0 : \\
&\quad \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n) \mathrel{|\!\approx\!\!\!\approx}_C^{\mathsf{sk}} (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m)) \\
&\qquad \Longrightarrow \\
&\quad \mathscr{D}_C^{\mathsf{sk}} \models \mathbf{Obs}(\alpha_1; \ldots; \alpha_n) \rightarrow [\alpha_1; \ldots; \alpha_n{}^-]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_m) \wedge \mathbf{B}(\beta_m)))
\end{aligned}$$

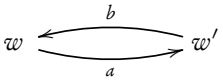


Figure III.1

Proof. Take any $\delta \in \mathscr{L}_\Delta$ and let $\delta = \alpha_1;\ldots;\alpha_n$, noting that $\Lambda, \mathsf{seen}(\alpha_1,1)\wedge\ldots\wedge\mathsf{seen}(\alpha_n,n) \not\approx^{\mathsf{sk}}_C (\tau_g(\gamma_1)\wedge\tau_b(\beta_1))\vee\ldots\vee(\tau_g(\gamma_m)\wedge\tau_b(\beta_m))$ entails that for some $k \leq m$ holds $\mathsf{msa}_C(\delta,\mathscr{R}) = \{(\gamma_1,\beta_1),\ldots,(\gamma_k,\beta_k)\}$ and $\{(\tau_g(\gamma_1)\wedge\tau_b(\beta_1))\vee\ldots\vee(\tau_g(\gamma_k)\wedge\tau_b(\beta_k))\} \models (\tau_g(\gamma_1)\wedge\tau_b(\beta_1))\vee\ldots\vee(\tau_g(\gamma_m)\wedge\tau_b(\beta_m))$, as shown in Proposition II.9. Take any $\mathfrak{M} \in \mathscr{D}^{\mathsf{sk}}_C$, such that from the definition of $\mathscr{D}^{\mathsf{sk}}_C$ follows $\mathfrak{M} \models \mathbf{Obs}(\alpha_1;\ldots;\alpha_n) \rightarrow [\alpha_1;\ldots;{\alpha_n}^-]((\mathbf{G}(\gamma_1)\wedge\mathbf{B}(\beta_1))\vee\ldots\vee(\mathbf{G}(\gamma_k)\wedge\mathbf{B}(\beta_k)))$, so that $\mathfrak{M} \models \mathbf{Obs}(\alpha_1;\ldots;\alpha_n) \rightarrow [\alpha_1;\ldots;{\alpha_n}^-]((\mathbf{G}(\gamma_1)\wedge\mathbf{B}(\beta_1))\vee\ldots\vee(\mathbf{G}(\gamma_m)\wedge\mathbf{B}(\beta_m)))$ follows by disjunctive weakening. □

III.2

Thus, it is shown that $\mathscr{D}^{\mathscr{R}}_C$ contains a subclass $\mathscr{D}^{\mathsf{sk}}_C$ that forms the dynamic counterpart of the skeptical interpretation of mental state abduction, given $\mathscr{R}$. Note that the relation proven in Theorem III.1 does not hold in the other direction (i.e. '$\not\Leftrightarrow$' is the case because of '$\not\Leftarrow$'), due to the fact that models which capture observation may force cases that are not handled by the abductive theory. This is easily seen in Figure III.1, given that $w \models \mathbf{Obs}(b)\wedge\mathbf{G}(p)\wedge\mathbf{B}(q)$ and $w' \models \mathbf{Obs}(a)\wedge\mathbf{G}(r)\wedge\mathbf{B}(s)$, so that this is a model agreeing with skeptical interpretation of the set of rules $\mathscr{R} = \{p <\!\!- q \mid a, r <\!\!- s \mid b\}$. Observe, however, that $w \models \mathbf{Obs}(a;b)\wedge[a;b^-](\mathbf{G}(r)\wedge\mathbf{B}(s))$ holds, but that $\Lambda_{\mathscr{R}}, \mathsf{seen}(a,1)\wedge\mathsf{seen}(b,2) \not\approx^{\mathsf{sk}}_C \tau_g(r)\wedge\tau_b(s)$. Thus, it is possible that models in $\mathscr{D}^{\mathsf{sk}}_C$ provide interpretations for action sequences that are *not* generated by some set of MYAPL rules, in addition to providing interpretations for sequences that are. This, however, is not problematic because we are looking for dynamic models for the abductive theory and not the other way around. Furthermore, observe that a comparable case is found in Chapter II in relation to the mental state abduction functions and the theory, where the syntactic restriction on elements of msa_X is the limiting factor (cf. Propositions II.7 and II.8). Similarly to the above, the dynamic counterpart of the credulous interpretation of mental state abduction exists, seen as follows.

Definition III.11 (credulous dynamic interpretation — complete observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\mathscr{L}_\Delta$ the percept language of Definition II.6, msa_C the mental state abduction function defined in Definition II.26, and $\mathscr{D}^{\mathscr{R}}_C$ the class of models defined in Definition III.9. The class of models $\mathscr{D}^{\mathsf{cr}[\![\gamma,\beta]\!]}_C$ reflecting the credulous dynamic interpretation of mental state abduction for particular $\gamma,\beta \in \mathscr{L}_0$ is then as follows.*

$$\mathscr{D}_C^{\mathsf{cr}[\![\gamma,\beta]\!]} = \{\mathfrak{M} \in \mathscr{D}_C^{\mathscr{R}} \mid \exists \delta \in \mathscr{L}_\Delta \exists w \in \mathrm{W}_{\mathfrak{M}} : ((\gamma,\beta) \in \mathsf{msa}_C(\delta,\mathscr{R}) \quad \& \quad \mathfrak{M}, w \models \mathbf{Obs}(\delta)) \;\&\; \forall \delta \in \mathscr{L}_\Delta : \quad ((\gamma,\beta) \in \mathsf{msa}_C(\delta,\mathscr{R}) \implies \mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow [\delta^-](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)))\}$$

The classes of models $\mathscr{D}_C^{\mathsf{cr}[\![\cdot]\!]}$ are each defined above relative to a particular mental state in the output of the mental state abduction function, of which they provide a credulous interpretation; note that $\mathscr{D}_C^{\mathsf{cr}[\![\phi,\psi]\!]} = \emptyset$ for $\phi, \psi \in \mathscr{L}_0$ for which holds that for no $\delta \in \mathscr{L}_\Delta$ it is the case that $(\phi,\psi) \in \mathsf{msa}_C(\delta,\mathscr{R})$, because in the definition of $\mathscr{D}_C^{\mathsf{cr}[\![\cdot]\!]}$ the relevance of models has been explicitly demanded with regard to the interpreted mental state. As with $\mathscr{D}_C^{\mathsf{sk}}$, it holds that those classes are non-trivial.

Proposition III.18. *Let $\mathscr{R}$ be a set of* MyAPL *rules, and $\mathscr{D}_C^{\mathscr{R}}$ and $\mathscr{D}_C^{\mathsf{cr}[\![\cdot]\!]}$ as defined in Definitions III.9 and III.11, respectively. It then holds that*

$$(\mathscr{D}_C^{\mathscr{R}} \neq \emptyset) \qquad \Longleftrightarrow \qquad \exists \gamma, \beta \in \mathscr{L}_0 : \quad (\mathscr{D}_C^{\mathsf{cr}[\![\gamma,\beta]\!]} \neq \emptyset)$$

Proof. ($\Rightarrow$) (*Consider the proof of Proposition III.17, and see that for the model $\mathfrak{M}$ described there holds $\mathfrak{M} \in \mathscr{D}_C^{\mathsf{cr}[\![\gamma,\beta]\!]}$, for suitable choice of $\gamma, \beta \in \mathscr{L}_0$ and $\alpha \in \mathscr{L}_\Delta$.*)

($\Leftarrow$) Straightforward, as $\mathscr{D}_C^{\mathsf{cr}[\![\gamma,\beta]\!]} \subseteq \mathscr{D}_C^{\mathscr{R}}$ for any $\gamma, \beta \in \mathscr{L}_0$. □

Similarly to how Proposition III.17 shows the existence of a non-trivial class of models $\mathscr{D}_C^{\mathsf{sk}}$ that is the dynamic counterpart of skeptical mental state abduction under complete observation, Proposition III.18 shows the existence of $\mathscr{D}_C^{\mathsf{cr}[\![\cdot]\!]}$, the dynamic reflection of credulous mental state abduction under complete observation. This can be illustrated formally as follows, along the lines of Theorem III.1.

Theorem III.2. *Let $\mathscr{R}$ be a set of* MyAPL *rules, $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ the abductive theory based on those rules as defined in Definition II.21, and $\mathscr{D}_C^{\mathsf{cr}[\![\cdot]\!]}$ the class of models defined in Definition III.11 with respect to those rules. It then holds that*

$$\forall \alpha_1;\ldots;\alpha_n \in \mathscr{L}_\Delta \forall \gamma, \beta \in \mathscr{L}_0 : \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1,1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n,n) \mid\!\approx_C^{\mathsf{cr}} (\tau_g(\gamma) \wedge \tau_b(\beta)) \implies \exists \gamma', \beta' \in \mathscr{L}_0 : \quad \mathscr{D}_C^{\mathsf{cr}[\![\gamma',\beta']\!]} \models \mathbf{Obs}(\alpha_1;\ldots;\alpha_n) \rightarrow [\alpha_1;\ldots;\alpha_n{}^-](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$$

Proof. Take any $\delta \in \mathscr{L}_\Delta$ and let $\delta = \alpha_1;\ldots;\alpha_n$, and observe that it holds that $\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1,1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n,n) \mid\!\approx_C^{\mathsf{cr}} (\tau_g(\gamma) \wedge \tau_b(\beta))$ implies $\exists (\gamma',\beta') \in \mathsf{msa}_C(\delta,\mathscr{R}) : \{\tau_g(\gamma') \wedge \tau_b(\beta')\} \models$

$\tau_g(\gamma) \wedge \tau_b(\beta)$, as shown in Proposition II.10. It should be observed that $\mathscr{D}_C^{\mathsf{cr}[\![\gamma',\beta']\!]} \models \mathbf{Obs}(\alpha_1;\ldots;\alpha_n) \rightarrow [\alpha_1;\ldots;\alpha_n{}^-](\mathbf{G}(\gamma') \wedge \mathbf{B}(\beta'))$, from which it then follows that $\mathscr{D}_C^{\mathsf{cr}[\![\gamma',\beta']\!]} \models \mathbf{Obs}(\alpha_1;\ldots;\alpha_n) \rightarrow [\alpha_1;\ldots;\alpha_n{}^-](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$. □

Thus, as Theorem III.2 shows, for every mental state that can be credulously inferred — in the sense of Chapter II — as a defeasible explanation for the agent's observed behavior, there exists a class of models that provides a dynamic interpretation of this explanation; meaning that this particular mental state is attributed to the agent in any state that precedes a state in which some sequence of actions, on grounds of which this mental state can be credulously inferred, is considered observed.

Interestingly, Theorems III.1 and III.2 rely on similar properties of the classes $\mathscr{D}_C^{\mathsf{sk}}$ and $\mathscr{D}_C^{\mathsf{cr}[\![\cdot]\!]}$ with respect to $\mathscr{D}_C^{\mathscr{R}}$, as the relations $\mid\!\approx^{\mathsf{sk}}$ and $\mid\!\approx^{\mathsf{cr}}$, defined in Definition II.24, did with respect to the abductive extensions. Specifically, for $\mid\!\approx^{\mathsf{sk}}$ it was required that the skeptically inferred fact is true in all abductive extensions, and furthermore that the set of extensions is non-empty. Compare this to Theorem III.1 which shows a correspondence between $\mid\!\approx^{\mathsf{sk}}$ and the class $\mathscr{D}_C^{\mathsf{sk}}$ pertaining to all abduced mental states. Furthermore, observe that this correspondence is non-trivial in the sense that $\mathscr{D}_C^{\mathsf{sk}}$ contains models interpreting the observed sequence of actions. Likewise, observing that for $\mid\!\approx^{\mathsf{cr}}$ it was required that the credulously inferred fact is true in some abductive extension, Theorem III.2 shows correspondence with some subclass of $\mathscr{D}_C^{\mathscr{R}}$ that pertains to a particular abduced mental state.

2.3 Ascription in Cases of Incomplete Observation

The assumption of incomplete observation, in the guise of the perceptory conditions of late and partial observation described in Section 3.2 of Chapter II, occurs when the observer admits the possibility that it has failed to see some of the agent's actions. Those perceptory conditions are implemented by the functions msa_L and msa_P defined in Definition II.26, and it would seem that also here dynamic interpretations can be posited, such as the one presented in Definition III.9, and that properties of particular subclasses can be shown as in Theorems III.1 and III.2. Before attempting this, however, it is important to appropriately define the different notions of incomplete observation. Based on the structural relations of Definition II.10 that characterize late/partial observation in the explanatory functions of Definition II.26, it can be stated that sequences which are incompletely (i.e. late or partially) observed but *not* completely observed, are sequences to which an actually observed sequence is related by means of the substring or dilution relation, but not the prefix relation. Such incompletely observed sequences may contain future actions which the agent is yet to perform; which, as a matter of fact, also holds with regard to complete observation: after all, a completely observed sequence is the prefix of some observable sequence, which may be also followed by actions which are yet to occur. The difference between complete and incomplete (i.e. late or partial) observation resides in the fact that possibly actions have been missed *before* the last observed action. Based on this insight, a satisfactory definition of incomplete observation can be given in our logical language $\mathscr{L}_{\mathbf{M}}$.

2.3.1 Late Observation

In defining incomplete observation, the fact is kept in mind that the dynamic interpretations are to focus on what could plausibly have been the case before the sequence of actions, to which the actually observed sequence is related, occurred. It is thereby required that if some sequence of actions is said to be incompletely (i.e. late or partially) observed then the last action(s) of this sequence must have actually been observed. This leads us to the following definition of late observation.

$$\begin{aligned}&\textbf{Late-Obs}(\delta,\delta) &&\triangleq \textbf{Obs}(\delta)\\ &\textbf{Late-Obs}(\delta,\delta';\delta) &&\triangleq \textbf{Obs}(\delta)\wedge\langle\delta^{-}\rangle(\textbf{Missed}(\delta'))\end{aligned}$$

Informally, this definition states that a sequence (the second argument of **Late-Obs**) is late observed if it corresponds to some actually observed sequence (the first argument of **Late-Obs**), or if this actually observed sequence is the suffix of the late observed sequence. It can also be shown formally that this definition captures the idea of late observation, as follows, where it should be noted that for any structural relation R, $\delta R\delta'$ states that δ is R-related to δ' (e.g. $\delta \mathrel{\underline{\blacktriangleleft}} \delta'$ states that δ is a suffix of δ').

Proposition III.19. *Let $\mathfrak{M}$ be any model for $\mathscr{L}_{\mathbf{M}}$, w any state in $\mathbb{W}_{\mathfrak{M}}$, and $\underline{\blacktriangleleft}$ the suffix relation defined in Definition II.10. It then holds that*

$$\begin{aligned}&\forall\delta,\delta'\in\mathscr{L}_{\Delta}:\\ &\quad\mathfrak{M},w\models\textbf{Late-Obs}(\delta,\delta') \quad\Longrightarrow\quad (\delta \mathrel{\underline{\blacktriangleleft}} \delta')\end{aligned}$$

Proof. As the definition of **Late-Obs** states, $\mathfrak{M},w\models\textbf{Late-Obs}(\delta,\delta')$ can only be true if $\delta=\delta'$, or if $\delta'=\delta'';\delta$ for some non-empty $\delta''\in\mathscr{L}_{\Delta}$. In either case $\delta \mathrel{\underline{\blacktriangleleft}} \delta'$. □

Corollary III.3. *Let $\mathfrak{M}$ be any model for $\mathscr{L}_{\mathbf{M}}$, w any state in $\mathbb{W}_{\mathfrak{M}}$, and $\underline{\blacktriangledown}$ the substring relation defined in Definition II.10. It then holds that*

$$\begin{aligned}&\forall\delta,\delta'\in\mathscr{L}_{\Delta}:\\ &\quad\mathfrak{M},w\models\textbf{Late-Obs}(\delta,\delta') \quad\Longrightarrow\quad (\delta \mathrel{\underline{\blacktriangledown}} \delta')\end{aligned}$$

Proof. Follows immediately from Proposition III.19, since $(\underline{\blacktriangleleft})\subseteq(\underline{\blacktriangledown})$. □

Because the logical formulation of late observation requires matching some (actually observed) sequence to another (incompletely observed) sequence, the following function is defined that allows for concisely expressing that formulation.

Definition III.12. *Given a goal/belief precondition (γ,β), a set of* MyAPL *rules $\mathscr{R}$, the function OS defined in Definition II.7, and the prefix relation $\underline{\blacktriangleright}$ defined in Definition II.10, the* opfx *function maps these arguments to prefixes of observable sequences extracted from plans related to the given precondition on grounds of the rules in $\mathscr{R}$.*

$$\mathsf{opfx}(\gamma,\beta,\mathscr{R})=\{\delta\in\mathscr{L}_{\Delta}\mid\exists(n:\gamma<\!-\beta\mid\pi)\in\mathscr{R}\exists\delta'\in OS(\pi):(\delta \mathrel{\underline{\blacktriangleright}} \delta')\}$$

Given the definition of late observation in $\mathscr{L}_{\mathbf{M}}$, the following dynamic interpretation can be given, which reflects the fact that the sequence of actually observed actions allows attributing some initial mental state to the agent on grounds of a late observed prefix of an observable sequence of a corresponding plan.

Definition III.13 (dynamic interpretation — late observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, and $\mathscr{M}$ the class of all models for $\mathscr{L}_{\mathbf{M}}$. $\mathscr{D}_L^{\mathscr{R}}$ is then the class of models reflecting the basic dynamic interpretation of mental state abduction with respect to $\mathscr{R}$, under the assumption of late observation, as follows.*

$$\mathscr{D}_L^{\mathscr{R}} = \{\mathfrak{M} \in \mathscr{M} \mid \exists \delta \in \mathscr{L}_{\Delta} \exists (\gamma, \beta) \in \mathsf{msa}_L(\delta, \mathscr{R}) \exists \delta' \in \mathsf{opfx}(\gamma, \beta, \mathscr{R}) \exists w \in \mathrm{W}_{\mathfrak{M}} : \\ \mathfrak{M}, w \models \textbf{Late-Obs}(\delta, \delta') \wedge \langle \delta'^{-} \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\}$$

III.2

Note that Definition III.13 employs the observable prefix δ' of some observable sequence on grounds of which the pair (γ, β) is inferred. This δ' is the sequence which is considered 'late observed' in regard to the actually observed sequence δ. The rationale for this approach is that, as stated earlier, our focus is on moments at which sequences of actions have been observed, and accordingly actions which have possibly been missed before observing the sequence δ are taken into account in attributing the observed agent's past mental state under the assumption of incomplete observation. By means of the interpretation given in Definition III.13, models in $\mathscr{D}_L^{\mathscr{R}}$ automatically take into account actions which possibly have been missed because of the fact that ascription of the initial mental state occurs in a state preceding the late observed prefix δ'. Similar to how in Section 2.2 it was shown that subclasses exist of $\mathscr{D}_C^{\mathscr{R}}$ that provide dynamic interpretations of the classical approach of the previous chapter, this can be shown for $\mathscr{D}_L^{\mathscr{R}}$. First, though, partial observation is formalized in $\mathscr{L}_{\mathbf{M}}$, after which the interpretations under both perceptory conditions are put forward.

2.3.2 Partial Observation

The definition of partial observation is somewhat tricky, because the underlying structural relation is one which allows gaps in matching sequences to occur in random places, as long as the actions of the observed sequence occur in the same order in the matching sequence. Fortunately, relying on the definition of late observation can aid in defining partial observation, which is not coincidental given the fact that the dilution relation $\odot$ that characterizes msa_P is also defined in terms of the substring relation $\blacktriangledown$ that characterizes msa_L (cf. Definition II.10). Using this approach partial observation can be defined as a finite disjunction of the various partitionings of the partially observed sequence in terms of late observed subsequences.

$$\textbf{Partial-Obs}(\alpha,\delta) \triangleq \textbf{Late-Obs}(\alpha,\delta)$$
$$\textbf{Partial-Obs}(\alpha_1;\ldots;\alpha_n,\delta) \triangleq$$
$$\bigvee\{\textbf{Late-Obs}(\alpha_n,\delta_n)\wedge\langle\delta_n{}^{-}\rangle(\textbf{Late-Obs}(\alpha_{n-1},\delta_{n-1})\wedge\ldots\wedge$$
$$(\langle\delta_2{}^{-}\rangle(\textbf{Late-Obs}(\alpha_1,\delta_1)))\ldots) \mid \exists\delta_1,\ldots,\delta_n\in\mathscr{L}_{\Delta}:(\delta=\delta_1;\ldots;\delta_n)\}$$

Thus, the above definition of partial observation (informally) states that the sequence δ is partially observed if it can be segmented into subsequences $\delta_1, \ldots, \delta_n$ which are all late observed, in such a way that their last actions (i.e. $\alpha_1, \ldots, \alpha_n$) are actually observed, and the sequence of all actually observed actions (i.e. the last actions of the late observed subsequences) makes up the total observed sequence $\alpha_1;\ldots;\alpha_n$ in relation to which the sequence δ is considered partially observed. That this definition captures the notion of partial observation can formally be shown as follows.

Proposition III.20. *Let $\mathfrak{M}$ be any model for $\mathscr{L}_{\mathbf{M}}$, w any state $w\in\mathrm{W}_{\mathfrak{M}}$, and $\underline{\odot}$ the dilution relation of Definition II.10. It then holds that*

$$\forall\alpha_1;\ldots;\alpha_n,\delta\in\mathscr{L}_{\Delta}:$$
$$\mathfrak{M},w\models\textbf{Partial-Obs}(\alpha_1;\ldots;\alpha_n,\delta) \quad\Longrightarrow\quad (\alpha_1;\ldots;\alpha_n\mathrel{\underline{\odot}}\delta)$$

Proof. If $\alpha_1;\ldots;\alpha_n$ is a single action, say $\alpha\in\mathscr{L}_{\Delta}$, then $\mathfrak{M},w\models\textbf{Partial-Obs}(\alpha_1;\ldots;\alpha_n,\delta)$ is defined as $\mathfrak{M},w\models\textbf{Late-Obs}(\alpha,\delta)$. Proposition III.19 shows that then $\alpha\blacktriangleleft\delta$, such that indeed $\alpha\mathrel{\underline{\odot}}\delta$. Otherwise, observe that if $\mathfrak{M},w\models\textbf{Partial-Obs}(\alpha_1;\ldots;\alpha_n,\delta)$ is to be true then δ must comprise at least n actions such that it can be represented as $\delta_1;\ldots;\delta_n$. Following the definition of **Partial-Obs** it can then be observed that $\mathfrak{M},w\models\textbf{Late-Obs}(\alpha_n,\delta_n)\wedge\langle\delta_n{}^{-}\rangle(\textbf{Late-Obs}(\alpha_{n-1},\delta_{n-1})\wedge\ldots\wedge(\langle\delta_2{}^{-}\rangle(\textbf{Late-Obs}(\alpha_1,\delta_1)))\ldots)$ follows from the fact that $\mathfrak{M},w\models\textbf{Partial-Obs}(\alpha_1;\ldots;\alpha_n,\delta)$, for some choice of $\delta=\delta_1;\ldots;\delta_n$, meaning that some disjunct in the definition of **Partial-Obs** is satisfied. Since $\mathfrak{M},w\models\textbf{Late-Obs}(\alpha_n,\delta_n)$ entails $\alpha_n\mathrel{\underline{\blacktriangleleft}}\delta_n$, see that thus follows $\alpha_1\mathrel{\underline{\blacktriangleleft}}\delta_1$ and …and $\alpha_n\mathrel{\underline{\blacktriangleleft}}\delta_n$, i.e. $\alpha_1\mathrel{\underline{\odot}}\delta_1$ and …and $\alpha_n\mathrel{\underline{\odot}}\delta_n$. Observe that for any $\delta,\delta',\delta'',\delta'''\in\mathscr{L}_{\Delta}$ holds that if $\delta\mathrel{\underline{\odot}}\delta'$ and $\delta''\mathrel{\underline{\odot}}\delta'''$ then $\delta;\delta''\mathrel{\underline{\odot}}\delta';\delta'''$, so that follows $\alpha_1;\ldots;\alpha_n\mathrel{\underline{\odot}}\delta_1;\ldots;\delta_n$. □

The dynamic interpretation of mental state abduction assuming partial observation is then as follows; similar to the case of late observation in Definition III.13.

Definition III.14 (dynamic interpretation — partial observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, and $\mathscr{M}$ the class of all models for $\mathscr{L}_{\mathbf{M}}$. $\mathscr{D}_P^{\mathscr{R}}$ is then the class of models reflecting the basic dynamic interpretation of mental state abduction with respect to $\mathscr{R}$, under the assumption of late observation, as follows.*

$$\mathscr{D}_P^{\mathscr{R}}=\{\mathfrak{M}\in\mathscr{M}\mid\exists\delta\in\mathscr{L}_{\Delta}\exists(\gamma,\beta)\in\mathsf{msa}_P(\delta,\mathscr{R})\exists\delta'\in\mathsf{opfx}(\gamma,\beta,\mathscr{R})\exists w\in\mathrm{W}_{\mathfrak{M}}:$$
$$\mathfrak{M},w\models\textbf{Partial-Obs}(\delta,\delta')\wedge\langle\delta'{}^{-}\rangle(\mathbf{G}(\gamma)\wedge\mathbf{B}(\beta))\}$$

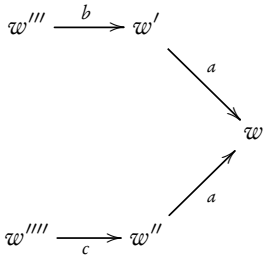


Figure III.2

The formal machinery is now in place which is used in the next section to identify particular subclasses of $\mathscr{D}_L^{\mathscr{R}}$ and $\mathscr{D}_P^{\mathscr{R}}$ that interpret an agent's observed behavior on grounds of its rules, under the assumption of incomplete observation, and to show how those classes correspond to the abductive interpretations given in Chapter II.

2.3.3 Skeptical and Credulous Interpretations

Before formally showing the relation between the dynamic interpretations of mental state abduction under late or partial observation and their counterparts in terms of the abductive theory defined in Chapter II, it is important to understand the formalization of the concept of incomplete observation in the language $\mathscr{L}_{\mathbf{M}}$. Specifically, it is important to note that in a single state it can be the case that multiple distinct sequences are considered to have been incompletely observed with respect to the same actually observed sequence. To see this, consider the model depicted in Figure III.2, assuming that $w \models \mathbf{Obs}(a)$, but $w' \models \mathbf{Missed}(b)$ and $w'' \models \mathbf{Missed}(c)$, so that $w \models \mathbf{Late\text{-}Obs}(a, b; a) \wedge \mathbf{Late\text{-}Obs}(a, c; a)$. Given this insight, consider the following definition of a subclass of $\mathscr{D}_L^{\mathscr{R}}$ that provides a dynamic interpretation of mental state abduction under late observation, as Definition III.10 did for $\mathscr{D}_C^{\mathscr{R}}$.

Definition III.15 (skeptical dynamic interpretation — late observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, msa_L the mental state abduction function defined in Definition II.26, and $\mathscr{D}_L^{\mathscr{R}}$ the class of models defined in Definition III.13. The class of models $\mathscr{D}_L^{\mathsf{sk}}$ reflecting the skeptical dynamic interpretation of mental state abduction is then as follows.*

$$\begin{aligned}\mathscr{D}_L^{\mathsf{sk}} = \{\mathfrak{M} \in \mathscr{D}_L^{\mathscr{R}} \mid \forall \delta \in \mathscr{L}_\Delta : \quad & \mathsf{msa}_L(\delta, \mathscr{R}) = \{(\gamma_1, \beta_1), \ldots, (\gamma_n, \beta_n)\} \neq \emptyset \implies \\ & \forall (\gamma, \beta) \in \mathsf{msa}_L(\delta, \mathscr{R}) \forall \delta' \in \mathsf{opfx}(\gamma, \beta, \mathscr{R}) : \\ & \mathfrak{M} \models \mathbf{Late\text{-}Obs}(\delta, \delta') \rightarrow [\delta'^{-}]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_n) \wedge \mathbf{B}(\beta_n)))\}\end{aligned}$$

As in the case of complete observation, it can be shown that this class is non-trivial.

Proposition III.21. *Let $\mathscr{R}$ be a set of* MYAPL *rules, and $\mathscr{D}_L^{\mathscr{R}}$ and $\mathscr{D}_L^{\mathsf{sk}}$ as defined in Definitions III.13 and III.15, respectively. It then holds that*

$$(\mathscr{D}_L^{\mathscr{R}} \neq \emptyset) \qquad \Longleftrightarrow \qquad (\mathscr{D}_L^{\mathsf{sk}} \neq \emptyset)$$

Proof. ($\Rightarrow$) Following the principle of the proof for Proposition III.17, observing that for the model $\mathfrak{M}$ depicted there holds $\mathfrak{M} \in \mathscr{D}_L^{\mathscr{R}}$ and $\mathfrak{M} \in \mathscr{D}_L^{\mathsf{sk}}$, given suitable choice of $\gamma, \beta \in \mathscr{L}_0$ and $\alpha \in \mathscr{L}_\Delta$, because $\models \mathbf{Obs}(\delta) \leftrightarrow \mathbf{Late\text{-}Obs}(\delta, \delta)$.

($\Leftarrow$) Straightforward, as $\mathscr{D}_L^{\mathsf{sk}} \subseteq \mathscr{D}_L^{\mathscr{R}}$. □

Similarly, the above can be shown to hold for partial observation.

Definition III.16 (skeptical dynamic interpretation — partial observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, msa_P the mental state abduction function defined in Definition II.26, and $\mathscr{D}_P^{\mathscr{R}}$ the class of models defined in Definition III.14. The class of models $\mathscr{D}_P^{\mathsf{sk}}$ reflecting the skeptical dynamic interpretation of mental state abduction is then as follows.*

$$\begin{aligned}\mathscr{D}_P^{\mathsf{sk}} = \{\mathfrak{M} \in \mathscr{D}_P^{\mathscr{R}} \mid \forall \delta \in \mathscr{L}_\Delta : \quad & \mathsf{msa}_P(\delta, \mathscr{R}) = \{(\gamma_1, \beta_1), \ldots, (\gamma_n, \beta_n)\} \quad \Longrightarrow \\ & \forall (\gamma, \beta) \in \mathsf{msa}_P(\delta, \mathscr{R}) \forall \delta' \in \mathsf{opfx}(\gamma, \beta, \mathscr{R}) : \\ & \mathfrak{M} \models \mathbf{Partial\text{-}Obs}(\delta, \delta') \rightarrow [\delta'^{-}]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_n) \wedge \mathbf{B}(\beta_n)))\}\end{aligned}$$

Proposition III.22. *Let $\mathscr{R}$ be a set of* MYAPL *rules, and $\mathscr{D}_P^{\mathscr{R}}$ and $\mathscr{D}_P^{\mathsf{sk}}$ as defined in Definitions III.14 and III.16 with respect to those rules, respectively. It then holds that*

$$(\mathscr{D}_P^{\mathscr{R}} \neq \emptyset) \qquad \Longleftrightarrow \qquad (\mathscr{D}_P^{\mathsf{sk}} \neq \emptyset)$$

Proof. ($\Rightarrow$) Following the principle of the proof for Proposition III.17, observing that for the model $\mathfrak{M}$ depicted there holds $\mathfrak{M} \in \mathscr{D}_P^{\mathscr{R}}$ and $\mathfrak{M} \in \mathscr{D}_P^{\mathsf{sk}}$, given suitable choice of $\gamma, \beta \in \mathscr{L}_0$ and $\alpha \in \mathscr{L}_\Delta$, because $\models \mathbf{Obs}(\delta) \leftrightarrow \mathbf{Partial\text{-}Obs}(\delta, \delta)$.

($\Leftarrow$) Straightforward, as $\mathscr{D}_P^{\mathsf{sk}} \subseteq \mathscr{D}_P^{\mathscr{R}}$. □

It can furthermore be shown that the dynamic interpretations of incomplete observation correspond to the abductive interpretations, as follows.

Theorem III.3. *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ the abductive theory based on those rules as defined in Definition II.21, and $\mathscr{D}_L^{\mathsf{sk}}$ the class of models identified in Definition III.15 with respect to those rules. It then holds that*

$$\begin{aligned}\forall \delta \in \mathscr{L}_\Delta \forall \gamma_1, \ldots, \gamma_m, \beta_1, \ldots, \beta_m \in \mathscr{L}_0 : \quad & (\delta = \alpha_1; \ldots; \alpha_n) \qquad \& \\ & \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n) \mathrel{|\!\approx}_L^{\mathsf{sk}} (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m)) \\ & \Longrightarrow \\ & \forall (\gamma, \beta) \in \mathsf{msa}_L(\delta, \mathscr{R}) \forall \delta' \in \mathsf{opfx}(\gamma, \beta, \mathscr{R}) : \\ & \mathscr{D}_L^{\mathsf{sk}} \models \mathbf{Late\text{-}Obs}(\delta, \delta') \rightarrow [\delta'^{-}]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_m) \wedge \mathbf{B}(\beta_m)))\end{aligned}$$

Proof. Take any $\delta \in \mathscr{L}_\Delta$ and let $\delta = \alpha_1;\ldots;\alpha_n$, noting that from $\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1,1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n,n) \mathrel{|\!\approx^{\mathsf{sk}}_L} (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m))$ follows that $\mathsf{msa}_L(\delta,\mathscr{R}) = \{(\gamma_1,\beta_1),\ldots,(\gamma_k,\beta_k)\}$, so that $\{(\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_k) \wedge \tau_b(\beta_k))\} \models (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m))$, as shown in Proposition II.9. It must then be the case that $\forall(\gamma,\beta) \in \mathsf{msa}_L(\delta,\mathscr{R}) \exists \delta' \in \mathsf{opfx}(\gamma,\beta,\mathscr{R}) : (\delta \mathbin{\underline{\blacktriangledown}} \delta')$. Take any $\mathfrak{M} \in \mathscr{D}^{\mathsf{sk}}_L$, such that from the definition of $\mathscr{D}^{\mathsf{sk}}_L$ follows $\mathfrak{M} \models \textbf{Late-Obs}(\delta,\delta') \rightarrow [\delta'^{-}]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_k) \wedge \mathbf{B}(\beta_k)))$, and by disjunctive weakening $\mathfrak{M} \models \textbf{Late-Obs}(\delta,\delta') \rightarrow [\delta'^{-}]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_m) \wedge \mathbf{B}(\beta_m)))$. □

Theorem III.4. *Let $\mathscr{R}$ be a set of* MyAPL *rules, $\Lambda_{\mathscr{R}} = (\Theta_C,\Theta_L,\Theta_P,\mathscr{A}_{\mathscr{R}})$ the abductive theory based on those rules as defined in Definition II.21, and $\mathscr{D}^{\mathsf{sk}}_P$ the class of models identified in Definition III.16 with respect to those rules. It then holds that*

III.2

$$\begin{aligned}
&\forall \delta \in \mathscr{L}_\Delta \forall \gamma_1,\ldots,\gamma_m,\beta_1,\ldots,\beta_m \in \mathscr{L}_0 : \qquad (\delta = \alpha_1;\ldots;\alpha_n) \qquad \&\\
&\quad \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1,1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n,n) \mathrel{|\!\approx^{\mathsf{sk}}_P} (\tau_g(\gamma_1) \wedge \tau_b(\beta_1)) \vee \ldots \vee (\tau_g(\gamma_m) \wedge \tau_b(\beta_m))\\
&\qquad \Longrightarrow\\
&\quad \forall(\gamma,\beta) \in \mathsf{msa}_P(\delta,\mathscr{R}) \forall \delta' \in \mathsf{opfx}(\gamma,\beta,\mathscr{R}) :\\
&\qquad \mathscr{D}^{\mathsf{sk}}_P \models \textbf{Partial-Obs}(\delta,\delta') \rightarrow [\delta'^{-}]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma_m) \wedge \mathbf{B}(\beta_m)))
\end{aligned}$$

Proof. The proof is identical to that of Theorem III.3 with replacement of '$\mathscr{D}^{\mathsf{sk}}_L$' by '$\mathscr{D}^{\mathsf{sk}}_P$', '**Late-Obs**' by '**Partial-Obs**', 'msa_L' by 'msa_P', '$\underline{\blacktriangledown}$' by '$\underline{\odot}$', and '$|\!\approx^{\mathsf{sk}}_L$' by '$|\!\approx^{\mathsf{sk}}_P$'. □

Similarly to how $\mathscr{D}^{\mathsf{cr}[\![\cdot]\!]}_C \subseteq \mathscr{D}^{\mathscr{R}}_C$ was identified alongside $\mathscr{D}^{\mathsf{sk}}_C \subseteq \mathscr{D}^{\mathscr{R}}_C$ to provide a credulous dynamic interpretation under complete observation of particular abduced goal/belief pairs, it is done here for late and partial observation.

Definition III.17 (credulous dynamic interpretation — late observation)**.** *Let $\mathscr{R}$ be a set of* MyAPL *rules, msa_L the mental state abduction function defined in Definition II.26, and $\mathscr{D}^{\mathscr{R}}_L$ the class of models defined in Definition III.13. The class of models $\mathscr{D}^{\mathsf{cr}[\![\gamma,\beta]\!]}_L$ reflecting the credulous dynamic interpretation of mental state abduction under late observation, for particular $\gamma,\beta \in \mathscr{L}_0$, is then as follows.*

$$\begin{aligned}
\mathscr{D}^{\mathsf{cr}[\![\gamma,\beta]\!]}_L = \{\mathfrak{M} \in \mathscr{D}^{\mathscr{R}}_L \mid\ & \exists \delta \in \mathscr{L}_\Delta \exists w \in \mathrm{W}_{\mathfrak{M}} \exists \delta' \in \mathsf{opfx}(\gamma,\beta,\mathscr{R}) :\\
& ((\gamma,\beta) \in \mathsf{msa}_L(\delta,\mathscr{R}) \quad \& \quad \mathfrak{M},w \models \textbf{Late-Obs}(\delta,\delta'))\\
& \qquad \&\\
& \forall \delta \in \mathscr{L}_\Delta \forall \delta' \in \mathsf{opfx}(\gamma,\beta,\mathscr{R}) : \quad (\gamma,\beta) \in \mathsf{msa}_L(\delta,\mathscr{R}) \quad \Longrightarrow\\
& \qquad \mathfrak{M} \models \textbf{Late-Obs}(\delta,\delta') \rightarrow [\delta'^{-}](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\}
\end{aligned}$$

Definition III.18 (credulous dynamic interpretation — partial observation)**.** *Let $\mathscr{R}$ be a set of* MyAPL *rules, msa_P the mental state abduction function defined in Definition II.26, and*

$\mathscr{D}_P^{\mathscr{R}}$ the class of models defined in Definition III.14. The class of models $\mathscr{D}_P^{\mathsf{cr}[\![\gamma,\beta]\!]}$ reflecting the credulous dynamic interpretation of mental state abduction under partial observation, for particular $\gamma, \beta \in \mathscr{L}_0$, is then as follows.

$$\begin{aligned}\mathscr{D}_P^{\mathsf{cr}[\![\gamma,\beta]\!]} = \{\mathfrak{M} \in \mathscr{D}_P^{\mathscr{R}} \mid \exists \delta \in \mathscr{L}_\Delta \exists w \in \mathbb{W}_{\mathfrak{M}} \exists \delta' \in \mathsf{opfx}(\gamma,\beta,\mathscr{R}) :\\ ((\gamma,\beta) \in \mathsf{msa}_P(\delta,\mathscr{R}) \quad \& \quad \mathfrak{M}, w \models \textbf{Partial-Obs}(\delta,\delta'))\\ \&\\ \forall \delta \in \mathscr{L}_\Delta \forall \delta' \in \mathsf{opfx}(\gamma,\beta,\mathscr{R}) : \quad (\gamma,\beta) \in \mathsf{msa}_P(\delta,\mathscr{R}) \implies\\ \mathfrak{M} \models \textbf{Partial-Obs}(\delta,\delta') \rightarrow [\delta'^{-}](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\}\end{aligned}$$

As in the case of complete observation, the notion of relevance is employed to ensure that $\mathscr{D}_X^{\mathsf{cr}[\![\phi,\psi]\!]} = \emptyset$ if $\neg\exists \delta \in \mathscr{L}_\Delta : ((\phi,\psi) \in \mathsf{msa}_X(\delta,\mathscr{R}))$, for $X \in \{L,P\}$. Also, the non-triviality of $\mathscr{D}_L^{\mathsf{cr}[\![\cdot]\!]}$ and $\mathscr{D}_P^{\mathsf{cr}[\![\cdot]\!]}$ can be shown, as follows.

Proposition III.23. *Let $\mathscr{R}$ be a set of* MYAPL *rules, and $\mathscr{D}_L^{\mathscr{R}}$ and $\mathscr{D}_L^{\mathsf{cr}[\![\cdot]\!]}$ as defined in Definitions III.13 and III.17, respectively. It then holds that*

$$(\mathscr{D}_L^{\mathscr{R}} \neq \emptyset) \qquad \iff \qquad \exists \gamma, \beta \in \mathscr{L}_0 : \quad (\mathscr{D}_L^{\mathsf{cr}[\![\gamma,\beta]\!]} \neq \emptyset)$$

Proof. ($\Rightarrow$) (*Following the principle of the proof for Proposition III.18, observing that for the model $\mathfrak{M}$ depicted there holds $\mathfrak{M} \in \mathscr{D}_L^{\mathsf{cr}[\![\gamma,\beta]\!]}$, given suitable choice of $\gamma, \beta \in \mathscr{L}_0$ and $\alpha \in \mathscr{L}_\Delta$.*)

($\Leftarrow$) Straightforward, as $\mathscr{D}_L^{\mathsf{cr}[\![\gamma,\beta]\!]} \subseteq \mathscr{D}_L^{\mathscr{R}}$ for any $\gamma, \beta \in \mathscr{L}_0$. □

Proposition III.24. *Let $\mathscr{R}$ be a set of* MYAPL *rules, and $\mathscr{D}_P^{\mathscr{R}}$ and $\mathscr{D}_P^{\mathsf{cr}[\![\cdot]\!]}$ as defined in Definitions III.14 and III.18, respectively. It then holds that*

$$(\mathscr{D}_P^{\mathscr{R}} \neq \emptyset) \qquad \iff \qquad \exists \gamma, \beta \in \mathscr{L}_0 : \quad (\mathscr{D}_P^{\mathsf{cr}[\![\gamma,\beta]\!]} \neq \emptyset)$$

Proof. ($\Rightarrow$) (*Following the principle of the proof for Proposition III.18, observing that for the model $\mathfrak{M}$ depicted there holds $\mathfrak{M} \in \mathscr{D}_P^{\mathsf{cr}[\![\gamma,\beta]\!]}$, given suitable choice of $\gamma, \beta \in \mathscr{L}_0$ and $\alpha \in \mathscr{L}_\Delta$.*)

($\Leftarrow$) Straightforward, as $\mathscr{D}_P^{\mathsf{cr}[\![\gamma,\beta]\!]} \subseteq \mathscr{D}_P^{\mathscr{R}}$ for any $\gamma, \beta \in \mathscr{L}_0$. □

As for the case of complete observation, it can be shown for incomplete observation that mental states that can be abduced credulously as a result of explaining observed behavior, on grounds of the approach formalized in Chapter II, are reflected in dynamic models which are subclasses of $\mathscr{D}_L^{\mathsf{cr}[\![\cdot]\!]}$ and $\mathscr{D}_P^{\mathsf{cr}[\![\cdot]\!]}$.

Theorem III.5. *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ the abductive theory based on those rules as defined in Definition II.21, and $\mathscr{D}_L^{\mathsf{cr}[\![\cdot]\!]}$ the class of models defined*

in Definition III.17 with respect to those rules. It then holds that

$$\begin{aligned}\forall \delta \in \mathscr{L}_\Delta \forall \gamma, \beta \in \mathscr{L}_0 : \quad & (\delta = \alpha_1; \ldots; \alpha_n) \quad \& \\ & \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n) \mathrel{|\!\approx^{\mathsf{cr}}_L} (\tau_g(\gamma) \wedge \tau_b(\beta)) \\ & \quad \Longrightarrow \\ & \exists \delta' \in \mathscr{L}_\Delta \exists \gamma', \beta' \in \mathscr{L}_0 : \quad \mathscr{D}_L^{\mathsf{cr}[\![\gamma', \beta']\!]} \models \textbf{Late-Obs}(\delta, \delta') \rightarrow [\delta'^{-}](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\end{aligned}$$

Proof. Take any $\delta \in \mathscr{L}_\Delta$ such that $\delta = \alpha_1; \ldots; \alpha_n \in \mathscr{L}_\Delta$, and see that $\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n) \mathrel{|\!\approx^{\mathsf{cr}}_L} (\tau_g(\gamma) \wedge \tau_b(\beta))$ implies $\exists (\gamma', \beta') \in \mathsf{msa}_L(\delta, \mathscr{R}) : \{\tau_g(\gamma') \wedge \tau_b(\beta')\} \models \tau_g(\gamma) \wedge \tau_b(\beta)$, as shown in Proposition II.10. It then holds that $\exists \mathscr{D}_L^{\mathsf{cr}[\![\gamma', \beta']\!]} \subseteq \mathscr{D}_L^{\mathscr{R}}$ and $\mathscr{D}_L^{\mathsf{cr}[\![\gamma', \beta']\!]} \models \textbf{Late-Obs}(\delta, \delta') \rightarrow [\delta'^{-}](\mathbf{G}(\gamma') \wedge \mathbf{B}(\beta'))$ for any $\delta' \in \mathsf{opfx}(\gamma', \beta', \mathscr{R})$, so that $\mathscr{D}_L^{\mathsf{cr}[\![\gamma', \beta']\!]} \models \textbf{Late-Obs}(\delta, \delta') \rightarrow [\delta'^{-}](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$ follows. □

III.2

Theorem III.6. *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\Lambda_{\mathscr{R}} = (\Theta_C, \Theta_L, \Theta_P, \mathscr{A}_{\mathscr{R}})$ the abductive theory based on those rules as defined in Definition II.21, and $\mathscr{D}_P^{\mathsf{cr}}$ the class of models defined in Definition III.18 with respect to those rules. It then holds that*

$$\begin{aligned}\forall \delta \in \mathscr{L}_\Delta \forall \gamma, \beta \in \mathscr{L}_0 : \quad & (\delta = \alpha_1; \ldots; \alpha_n) \quad \& \\ & \Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n) \mathrel{|\!\approx^{\mathsf{cr}}_P} (\tau_g(\gamma) \wedge \tau_b(\beta)) \\ & \quad \Longrightarrow \\ & \exists \delta' \in \mathscr{L}_\Delta \exists \gamma', \beta' \in \mathscr{L}_0 : \quad \mathscr{D}_P^{\mathsf{cr}[\![\gamma', \beta']\!]} \models \textbf{Partial-Obs}(\delta, \delta') \rightarrow [\delta'^{-}](\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\end{aligned}$$

Proof. The proof is identical to that of Proposition III.5 with replacement of '$\mathscr{D}_L^{\mathscr{R}}$' by '$\mathscr{D}_P^{\mathscr{R}}$'; '$\mathscr{D}_L^{\mathsf{cr}[\![\cdot]\!]}$' by '$\mathscr{D}_P^{\mathsf{cr}[\![\cdot]\!]}$'; '**Late-Obs**' by '**Partial-Obs**'; and 'msa_L' by 'msa_P'. □

2.4 Reflection

In this section the material presented in Sections 2.2 and 2.3 is reflected upon, before it is illustrated with an example.

2.4.1 'Skepticism' and 'Credulity'

In Chapter II, explanation of observed behavior was formalized in terms of nonmonotonic inference with respect to a classical logical theory about an agent's observable behavior. Skeptical and credulous inference in this case have a natural meaning, in the sense that they reflect the conviction with which the reasoner states something to be the case, given that inference is known to be defeasible. In contrast, the classes of models presented in this chapter as dynamic counterparts of skeptical and credulous interpretations of the mental state abduction function are, in fact, models of what monotonically can be stated (in modal terms) to necessarily/possibly have been the case in states preceding the actions of the agent. Thus, the labels 'skeptical' and 'credulous' are justified in reference to those

classes only in light of a relation with the aforementioned respective nonmonotonic inferential modi. Note that changing perspective in order to consider nonmonotonic inference as a monotonic inference is as such not suspicious, though; it is recognized (Flach & Kakas, 2000) that abduction (being nonmonotonic) can be regarded as deduction in the abductive extension (which is monotonic), as done in Chapter II (cf. Definition II.24). Also, providing monotonic encodings, modal or otherwise, of nonmonotonic problems is precedented in literature (Zhang & Foo, 2005; Meyer & van der Hoek, 1991).

2.4.2 Different Levels of Gullibility

As just stated, the models discussed so far represent interpretations of a monotonic theory pertaining to states of affairs preceding sequences of actions performed by some agent. Classes of models are defined, based on functions that have been given a nonmonotonic interpretation in Section 4.2 of the previous chapter, in light of certain assumptions pertaining to ascription and observation. It is noteworthy in this regard that the skeptical interpretations defined in Definition III.10, III.15, and III.16 can be said to represent the 'ultimate' skeptical observer. That is to say, based on those models the observer maintains the assumption that *any* observed action could have been the first action of the agent's plan, and does so in regard to plans that accompany any rule it knows the agent to have. To illustrate this with a simple example: if the agent's set of rules is $\{1 : p <\!\!- q \mid a, 2 : r <\!\!- s \mid a, 3 : t <\!\!- u \mid a; a\}$ then in any state satisfying $\mathbf{Obs}(a;a)$ the skeptical observer maintains the possibility that the agent applied any of those three rules, as $\models \mathbf{Obs}(a;a) \rightarrow \mathbf{Obs}(a)$. In cases of incomplete observation the skeptical observer furthermore maintains the possibility that the agent could have performed any of the action sequences to which observed actions are appropriately (i.e. late or partially) related, and in this sense can be called 'ultimately skeptical'.

Likewise, the credulous interpretations defined in Definitions III.11, III.17, and III.18 represent a 'quite credulous' observer, which has jumped to the conclusion that, given some observed sequence of actions, the agent had a particular mental state, and, in cases of incomplete observation, could have performed any sequence of actions to which the observed sequence is appropriately related. It could be argued that an even more credulous observer (i.e. an 'ultimately credulous' observer) ascribes a particular mental state to the agent based on observation of *a particular* (incompletely) observed sequence of actions instead of *any* (incompletely) observed sequence of actions. In the treatment given in this chapter, though, the notions of 'skepticism' and 'credulity' are restricted to selection of the mental state attributed to the agent, not the presumption of incompletely observed sequences. Nevertheless, this shows that the constraints in defining dynamic interpretations can be varied to reflect observers that differ in their willingness to make assumptions, i.e. their 'gullibility'.

2.4.3 Related Interpretations, and a Final Remark

In line with intuition it can be shown that the classes of models interpreting different perceptory conditions are related to each other similarly to how the structural relations $\blacktriangleright$, $\underline{\blacktriangledown}$, and $\underline{\odot}$ are (cf. Proposition II.1), and likewise the functions msa_C, msa_L, and msa_P (cf. Proposition II.11).

Theorem III.7. *Let $\mathscr{R}$ be a set of* MYAPL *rules, with respect to which $\mathscr{D}_C^{\mathscr{R}}$ is the class of models defined by Definition III.9, $\mathscr{D}_L^{\mathscr{R}}$ the class of models defined by Definition III.13, and $\mathscr{D}_P^{\mathscr{R}}$ the class of models defined by Definition III.14. It then holds that*

$$\mathscr{D}_C^{\mathscr{R}} \subseteq \mathscr{D}_L^{\mathscr{R}} \subseteq \mathscr{D}_P^{\mathscr{R}}$$

Proof. Consider Definitions III.9, III.13, and III.14, and note that they differ in the fact that the preconditions of the implications which they respectively enforce are **Obs**(δ), **Late-Obs**(δ, δ'), and **Partial-Obs**(δ, δ'), for particular $\delta, \delta' \in \mathscr{L}_\Delta$. Because $\models$ **Obs**$(\delta) \rightarrow$ **Late-Obs**(δ, δ) it straightforwardly follows that $\mathscr{D}_C^{\mathscr{R}} \subseteq \mathscr{D}_L^{\mathscr{R}}$, since any model satisfying the conditions of Definitions III.9 satisfies those of Definition III.13. Likewise, from $\models$ **Late-Obs**$(\delta, \delta') \rightarrow$ **Partial-Obs**(δ, δ') follows $\mathscr{D}_L^{\mathscr{R}} \subseteq \mathscr{D}_P^{\mathscr{R}}$, because any model which satisfies the conditions of Definitions III.13 also satisfies those of Definition III.14. □

Thus, it is shown that the dynamic interpretations for different conditions are related to each other similarly to the way the abductive and functional approaches for those conditions are. It is furthermore noteworthy that the dynamic interpretations define ascription relative to the 'moment' at which action sequences are observed, and do not consider actions which may occur at future moments (with respect to the moment of observation). This is especially noteworthy in the case of partial observation, where the observer considers it possible that it has failed to perceive different actions that the agent presumedly performed in the past. It could then be argued that the observer should consider the possibility that it also failed to perceive actions the agent performed *after* the last action the observer did actually perceive. This argument makes sense, and accordingly our approach does not intend to rule out the modeling of presumed future actions; it just focuses on interpretation of agents' intentions, based on their observed actions, by ascription of mental states. From this perspective, it seems natural to take the moment of having observed the last action as reference point.

A final remark concerns the fact that, thus far, in ascription only the information about an observed agent's mental state has been used which derives from the goal/belief preconditions of the agent's rules. However, the postconditions of rules — i.e. the agent's plans — can also be used a source of information about agents' mental states. This matter is dealt with in the next section, but first the matter that has been discussed so far in the current section is illustrated with an example.

2.5 Example

In this example, the dynamic interpretations presented in Section 2 are illustrated by means of the rules utilized earlier in the example in Section 3.7 of Chapter II, which are repeated here in Listing III.1. The earlier discussion of this example focused on initial observation

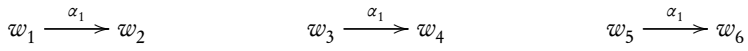


Figure III.3

of the action goto_shelf, warranting the following explanation.

$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\mathsf{goto\_shelf}, 1) \mid\!\approx_C \mathsf{rule}(1)$$
$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\mathsf{goto\_shelf}, 1) \mid\!\approx_C \mathsf{rule}(2)$$
$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\mathsf{goto\_shelf}, 1) \mid\!\approx_C \mathsf{rule}(3)$$

Accordingly, the following then holds, where for the sake of brevity again $\alpha_1 = \mathsf{goto\_shelf}$, $p_1 = \mathsf{paper\_read}$, $p_2 = \mathsf{paper\_on\_shelf}$, $p_3 = \mathsf{bug\_squashed}$, $p_4 = \mathsf{bug\_on\_table}$, $p_5 = \mathsf{flowers\_arranged}$, $p_6 = \mathsf{flowers\_on\_table}$, and $p_7 = \mathsf{vase\_on\_shelf}$.

$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \mid\!\approx^{\mathsf{cr}}_C \mathsf{goal}(p_1) \wedge \mathsf{belief}(p_2)$$
$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \mid\!\approx^{\mathsf{cr}}_C \mathsf{goal}(p_3) \wedge (\mathsf{belief}(p_4) \wedge \mathsf{belief}(p_2))$$
$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \mid\!\approx^{\mathsf{cr}}_C \mathsf{goal}(p_5) \wedge (\mathsf{belief}(p_6) \wedge \mathsf{belief}(p_7))$$

Figure III.3 then shows three models; model $\mathfrak{M}$ such that $\mathrm{W}_{\mathfrak{M}} = \{w_1, w_2\}$, model $\mathfrak{M}'$ such that $\mathrm{W}_{\mathfrak{M}'} = \{w_3, w_4\}$, and model $\mathfrak{M}''$ such that $\mathrm{W}_{\mathfrak{M}''} = \{w_5, w_6\}$. Assume that $\{w_2, w_4, w_6\} \models \mathbf{Obs}(\alpha_1)$, that $w_1 \models \mathbf{G}(p_1) \wedge \mathbf{B}(p_2)$, that $w_3 \models \mathbf{G}(p_3) \wedge \mathbf{B}(p_4 \wedge p_2)$, and that $w_1 \models \mathbf{G}(p_5) \wedge \mathbf{B}(p_6 \wedge p_7)$. Observe that it is then the case that $\mathfrak{M} \in \mathscr{D}_C^{\mathsf{cr}[\![p_1, p_2]\!]}$, $\mathfrak{M}' \in \mathscr{D}_C^{\mathsf{cr}[\![p_3, p_4 \wedge p_2]\!]}$, $\mathfrak{M}'' \in \mathscr{D}_C^{\mathsf{cr}[\![p_5, p_6 \wedge p_7]\!]}$, and $\{\mathfrak{M}, \mathfrak{M}', \mathfrak{M}''\} \subseteq \mathscr{D}_C^{\mathsf{sk}}$.

The example in Section 3.7 of Chapter II takes $\alpha_6 = \mathsf{goto\_table}$ as the second observed action, and here we do the same. As was noted in that previous chapter, no explanation can be found under the assumption of complete and late observation, yet under partial

```
1:  paper_read          <-  paper_on_shelf   |
    { goto_shelf; pickup_paper; goto_chair; sit; read }

2:  bug_squashed        <-  bug_on_table and paper_on_shelf      |
    { goto_shelf; pickup_paper; goto_table; squash_bug }

3:  flowers_arranged    <-  flowers_on_table and vase_on_shelf   |
    { goto_shelf; pickup_vase; goto_table; arrange_flowers }
```

Listing III.1

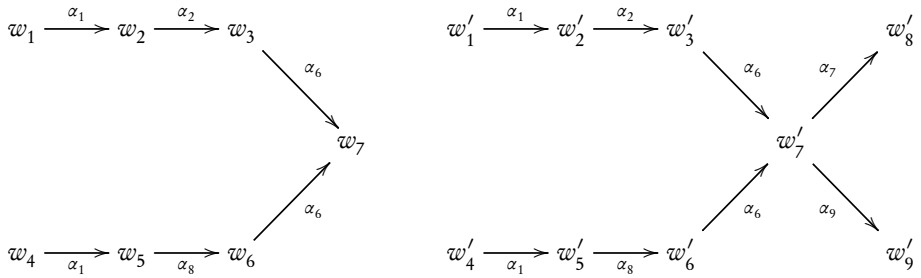


Figure III.4

III.3

observation the following holds.

$$\Lambda_{\mathscr{R}}, \mathsf{seen}(\alpha_1, 1) \wedge \mathsf{seen}(\alpha_6, 2) \mid\!\approx^{\mathsf{sk}}_{P} (\mathsf{goal}(p_3) \wedge (\mathsf{belief}(p_4) \wedge \mathsf{belief}(p_2))) \\ \vee (\mathsf{goal}(p_5) \wedge (\mathsf{belief}(p_6) \wedge \mathsf{belief}(p_7)))$$

Two models $\mathfrak{M}'''$ and $\mathfrak{M}''''$ for which it holds that $\mathrm{W}_{\mathfrak{M}'''} = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$ and $\mathrm{W}_{\mathfrak{M}''''} = \{w'_1, w'_2, w'_3, w'_4, w'_5, w'_6, w'_7, w'_8, w'_9\}$ are depicted in Figure III.4. Assume that holds $\{w_2, w'_2, w_5, w'_5\} \models \mathbf{Obs}(\alpha_1)$, $\{w_7, w'_7\} \models \mathbf{Obs}(\alpha_6)$, $\{w_3, w'_3\} \models \mathbf{Missed}(\alpha_2)$, and $\{w_6, w'_6\} \models \mathbf{Missed}(\alpha_8)$. Furthermore, let $\{w_1, w'_1\} \models \mathbf{G}(p_3) \wedge \mathbf{B}(p_4 \wedge p_2)$ and $\{w_4, w'_4\} \models \mathbf{G}(p_5) \wedge \mathbf{B}(p_6 \wedge p_7)$, noting in regard to the figure that, in accordance with the presentation of this example in the previous chapter, it is taken that $\alpha_2 = \mathsf{pickup\_paper}$, $\alpha_7 = \mathsf{squash\_bug}$, $\alpha_8 = \mathsf{pickup\_vase}$, and $\alpha_9 = \mathsf{arrange\_flowers}$. See that then $\{\mathfrak{M}''', \mathfrak{M}''''\} \subseteq \mathscr{D}^{\mathsf{sk}}_{P}$, because $\{w_7, w'_7\} \models \mathbf{Partial\text{-}Obs}(\alpha_1; \alpha_6, \alpha_1; \alpha_2; \alpha_6) \wedge \mathbf{Partial\text{-}Obs}(\alpha_1; \alpha_6, \alpha_1; \alpha_8; \alpha_6)$, and furthermore $\{w_7, w'_7\} \models \langle \alpha_1; \alpha_2; \alpha_6^- \rangle (\mathbf{G}(p_3) \wedge \mathbf{B}(p_4 \wedge p_2)) \wedge \langle \alpha_1; \alpha_8; \alpha_6^- \rangle (\mathbf{G}(p_5) \wedge \mathbf{B}(p_6 \wedge p_7))$, and both $\{w_7, w'_7\} \models [\alpha_1; \alpha_2; \alpha_6^-]((\mathbf{G}(p_3) \wedge \mathbf{B}(p_4 \wedge p_2)) \vee (\mathbf{G}(p_5) \wedge \mathbf{B}(p_6 \wedge p_7)))$ and $\{w_7, w'_7\} \models [\alpha_1; \alpha_8; \alpha_6^-]((\mathbf{G}(p_3) \wedge \mathbf{B}(p_4 \wedge p_2)) \vee (\mathbf{G}(p_5) \wedge \mathbf{B}(p_6 \wedge p_7)))$. Last but not least, note that $w'_8 \models \mathbf{Done}(\alpha_7)$ and $w'_9 \models \mathbf{Done}(\alpha_9)$, which is a formal reflection of the intuition, mentioned in Section 3.7 of Chapter II, that the observer considers the possibility of future actions being done by the agent if it interprets observation with respect to w'_7.

3 Plan-Based Ascription

In reasoning about an agent's observed behavior, a model of the agent's mental state can be formed on grounds of the agent's rules. Such rules, as defined in Definition II.3, tie together goals and beliefs, constituting the preconditions under which rules can be adopted, with behavioral recipes (plans) that, given truth of the belief condition, should result in achievement of the goal; barring unforeseen circumstances. The preceding sections focused principally on how relations between observed actions and observable sequences generated

by plans allow for inference and ascription of rules' preconditions. Underlying rationale in this case is the idea that if observed actions allow for defeasibly inferring that the agent selected some plan because of the application of some rule, then possibly the corresponding preconditions applied. This section develops the idea that if it can be defeasibly inferred that an agent selected a particular plan, then it is also justifiable to ascribe the goals and beliefs which this agent should have had if its observed actions did indeed arise from execution of that particular plan.

3.1 Ascription of Computation Sequences

Similarly to how in Section 3.3 the function *OS* was defined in order to extract observable sequences from plans, here the function *CS* is defined that extracts *computation sequences* of plans, as follows.

Definition III.19 (computation sequences)**.** *The function* $CS : \mathscr{L}_\Pi \longrightarrow \wp(\mathscr{L}_\Pi)$*, which translates plans to sets of observable sequences, is defined as follows, given* $\alpha \in \mathsf{Act}$*,* $\phi \in \{\mathbf{B}(\psi), \neg\mathbf{B}(\psi), \mathbf{G}(\psi), \neg\mathbf{G}(\psi) \mid \psi \in \mathscr{L}_0\}$*, and* $\pi, \pi' \in \mathscr{L}_\Pi$*.*

$$
\begin{aligned}
CS(\alpha) &= \{\alpha\} \\
CS(\phi?) &= \{\phi?\} \\
CS(\pi;\pi') &= CS(\pi) \bullet CS(\pi') \\
CS(\pi + \pi') &= CS(\pi) \cup CS(\pi') \\
CS(\pi^*) &= \bigcup_{n \in \mathbb{N}_0} CS(\pi^n) \\
&\quad \textit{where } \pi^n = \pi;\pi^{n-1} \textit{ and } \pi^0 = \mathbf{B}(\top)?
\end{aligned}
$$

The composition operator $\bullet : \wp(\mathscr{L}_\Pi) \times \wp(\mathscr{L}_\Pi) \longrightarrow \wp(\mathscr{L}_\Pi)$ *is defined as follows.*

$$
\begin{aligned}
\Pi \bullet \Pi' &= \{\pi;\pi' \mid \pi \in \Pi, \pi' \in \Pi'\} && \textit{if } \Pi \neq \emptyset \textit{ and } \Pi' \neq \emptyset \\
\Pi \bullet \Pi' &= \Pi && \textit{if } \Pi \neq \emptyset \textit{ and } \Pi' = \emptyset \\
\Pi \bullet \Pi' &= \Pi' && \textit{if } \Pi = \emptyset \textit{ and } \Pi' \neq \emptyset \\
\Pi \bullet \Pi' &= \emptyset && \textit{otherwise}
\end{aligned}
$$

This function is the *CS* function as defined by Harel et al. (2000), noting the exceptions that in our case the 'skip' action is $\mathbf{B}(\top)?$, and also that the range of the function is a subset of the domain, whereas Harel et al. omit sequential composition by means of ';' for elements in the range. It should be noted that the operator '$\bullet$' used in Definition III.19 is a generalization of the one used in Definition II.7, extending it from accepting sets of observable sequences to accepting sets of plans. Likewise, the structural relations on observable sequences can be generalized to plans, as follows.

Definition III.20 (extended structural relations)**.** *The partial orders prefix* $\underline{\blacktriangleright}$*, substring* $\underline{\blacktriangledown}$*, and dilution* $\underline{\odot}$*, defined in Definition II.10, are extended to* $\mathscr{L}_{\Pi}$*, as follows.*

$$\underline{\blacktriangleright} = \{(\pi,\pi),(\pi,\pi;\pi') \mid \pi,\pi' \in \mathscr{L}_{\Pi}\}$$
$$\underline{\blacktriangledown} = \{(\pi,\pi),(\pi,\pi;\pi'),(\pi,\pi';\pi),(\pi,\pi';\pi;\pi'') \mid \pi,\pi',\pi'' \in \mathscr{L}_{\Pi}\}$$
$$\underline{\odot} = \bigcup_{n\in\mathbb{N}_0} \mathsf{dil}_n(\underline{\blacktriangledown})$$

The operator dil_n *is generalized accordingly.*

$$\mathsf{dil}_n(\underline{\blacktriangledown}) = \mathsf{dil}_{n-1}(\underline{\blacktriangledown}) \cup \{(\pi;\pi'',\pi';\pi''') \mid (\pi,\pi'),(\pi'',\pi''') \in \mathsf{dil}_{n-1}(\underline{\blacktriangledown})\}$$
$$\textit{for } n > 0\textit{, where } \mathsf{dil}_0(\underline{\blacktriangledown}) = \underline{\blacktriangledown}$$

III.3

Given the function *CS* and the extended structural relations defined above, interpretations can be given of rules, similar to the dynamic interpretations of Definitions III.9, III.13, and III.14, which reflect the fact that if the agent performed actions related to an observable sequence of a particular plan, then any test actions that occurred in corresponding computation sequences must have succeeded. In order to simplify formulation of those interpretations, the function 'cpfx' is defined, similar to 'opfx' in Definition III.12.

Definition III.21. *Given a goal/belief precondition* (γ,β)*, a set of* MYAPL *rules* $\mathscr{R}$*, the function CS defined in Definition III.19, and the prefix relation* $\underline{\blacktriangleright}$ *defined in Definition III.20, the* cpfx *function maps these arguments to prefixes of computation sequences extracted from plans related to the given precondition on grounds of the rules in* $\mathscr{R}$*.*

$$\mathsf{cpfx}(\gamma,\beta,\mathscr{R}) = \{\pi \in \mathscr{L}_{\Pi} \mid \exists(n:\gamma <\!\!- \beta \mid \pi') \in \mathscr{R} \exists \pi'' \in CS(\pi') : (\pi \underline{\blacktriangleright} \pi'')\}$$

In order to distinguish plan-based ascription from the 'standard' interpretations that focus only on ascription of preconditions of rules, distinct classes of models are identified. It should be stressed here that models in those classes force the possibility that the agent performed (prefixes of) particular *computation sequences*, as opposed to just (partially observed) sequences of primitive actions. Naturally, this means that a difference exists with the interpretations of Section 2 only if those computation sequences contain unobservable actions. This occurs in the case of plans that comprise *conditionals*, i.e. plans that contain 'if-then-else' or 'while-do' constructs.

Definition III.22 (plan-based ascription — complete observation)**.** *Let* $\mathscr{R}$ *be a set of* MYAPL *rules,* $\mathscr{M}$ *the class of all models for* $\mathscr{L}_{\mathbf{M}}$*, and* msa_C *the mental state abduction function defined in Definition II.26.* $\mathscr{P}_C^{\mathscr{R}}$ *is then the class of models reflecting plan-based ascription with respect to* $\mathscr{R}$*, under the assumption of complete observation, as follows.*

$$\mathscr{P}_C^{\mathscr{R}} = \{\mathfrak{M} \in \mathscr{M} \mid \exists\delta \in \mathscr{L}_{\Delta} \exists(\gamma,\beta) \in \mathsf{msa}_C(\delta,\mathscr{R}) \exists\pi \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) \exists w \in \mathrm{W}_{\mathfrak{M}} :$$
$$OS(\pi) = \{\delta\} \quad \& \quad \mathfrak{M}, w \models \mathbf{Obs}(\delta) \wedge \langle\pi^{-}\rangle(\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\}$$

Definition III.23 (plan-based ascription — late observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\mathscr{M}$ the class of all models for $\mathscr{L}_{\mathbf{M}}$, and msa_L the mental state abduction function defined in Definition II.26. $\mathscr{P}_L^{\mathscr{R}}$ is then the class of models reflecting plan-based ascription with respect to $\mathscr{R}$, under the assumption of late observation, as follows.*

$$\mathscr{P}_L^{\mathscr{R}} = \{\mathfrak{M} \in \mathscr{M} \mid \exists \delta, \delta' \in \mathscr{L}_\Delta \exists (\gamma, \beta) \in \mathsf{msa}_L(\delta, \mathscr{R}) \exists \pi \in \mathsf{cpfx}(\gamma, \beta, \mathscr{R}) \exists w \in \mathrm{W}_{\mathfrak{M}} : \\ OS(\pi) = \{\delta'\} \quad \& \quad \mathfrak{M}, w \models \textbf{Late-Obs}(\delta, \delta') \wedge \langle \pi^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\}$$

Definition III.24 (plan-based ascription — partial observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\mathscr{M}$ the class of all models for $\mathscr{L}_{\mathbf{M}}$, and msa_P the mental state abduction function defined in Definition II.26. $\mathscr{P}_P^{\mathscr{R}}$ is then the class of models reflecting plan-based ascription with respect to $\mathscr{R}$, under the assumption of partial observation, as follows.*

$$\mathscr{P}_P^{\mathscr{R}} = \{\mathfrak{M} \in \mathscr{M} \mid \exists \delta, \delta' \in \mathscr{L}_\Delta \exists (\gamma, \beta) \in \mathsf{msa}_P(\delta, \mathscr{R}) \exists \pi \in \mathsf{cpfx}(\gamma, \beta, \mathscr{R}) \exists w \in \mathrm{W}_{\mathfrak{M}} : \\ OS(\pi) = \{\delta'\} \quad \& \quad \mathfrak{M}, w \models \textbf{Partial-Obs}(\delta, \delta') \wedge \langle \pi^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))\}$$

Note the similarity of the interpretations of Definitions III.22–III.24 to those of Definitions III.9, III.13, and III.14. The difference lies in the fact that it is required in Definitions III.22–III.24 that ascription of a goal/belief precondition (γ, β) takes place in a state preceding the computation sequence prefix π, as opposed to the requirement in the definitions of Section 2 that ascription takes place in a state preceding the prefix of an observable sequence. This difference is quite fundamental, though, as it ensures that in the case of plan-based ascription only models are considered that reflect the presumption that the agent is performing a plan, and thus provide more information than simply acknowledging that it performed certain actions. This shows through in the following theorem, proving that, given some set of rules, the models in Definitions III.22–III.24 are a subset of their counterpart defined in Definitions III.9, III.13, and III.14.

Theorem III.8. *Let $\mathscr{R}$ be a set of* MYAPL *rules, $\mathscr{D}_C^{\mathscr{R}}$, $\mathscr{D}_L^{\mathscr{R}}$, and $\mathscr{D}_P^{\mathscr{R}}$ be the classes defined in Definitions III.9, III.13 and III.14 and $\mathscr{P}_C^{\mathscr{R}}$, $\mathscr{P}_L^{\mathscr{R}}$, and $\mathscr{P}_P^{\mathscr{R}}$ the classes defined in Definitions III.22, III.23 and III.24, respectively. It then holds that*

$$\forall X \in \{C, L, P\} : \qquad \mathscr{P}_X^{\mathscr{R}} \subseteq \mathscr{D}_X^{\mathscr{R}}$$

Proof. Let $X = C$, and consider Definitions III.9 and III.22. Take some model $\mathfrak{M}$ such that $\mathfrak{M} \in \mathscr{P}_C^{\mathscr{R}}$, observing that then for some $\delta \in \mathscr{L}_\Delta$ and $(\gamma, \beta) \in \mathsf{msa}_C(\delta, \mathscr{R})$ holds that $\exists w \in \mathrm{W}_{\mathfrak{M}} : (\mathfrak{M}, w \models \textbf{Obs}(\delta) \wedge \langle \pi^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)))$ for some $\pi \in \mathsf{cpfx}(\gamma, \beta, \mathscr{R})$ for which holds $OS(\pi) = \{\delta\}$. It then follows that $\mathfrak{M}, w \models \langle \delta^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$, so that $\mathfrak{M} \in \mathscr{D}_C^{\mathscr{R}}$. Likewise, consider Definitions III.13 and III.14, observing that in case $X = L$ (or $X = P$) it follows from $\mathfrak{M}, w \models \textbf{Late-Obs}(\delta, \delta') \wedge \langle \pi^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$ (or $\mathfrak{M} \models \textbf{Partial-Obs}(\delta, \delta') \wedge \langle \pi^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$) and $OS(\pi) = \{\delta'\}$ that $\mathfrak{M}, w \models \langle \delta'^- \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$, so that $\mathfrak{M} \in \mathscr{D}_L^{\mathscr{R}}$ (or $\mathfrak{M} \in \mathscr{D}_P^{\mathscr{R}}$). □

Along the lines of Theorem III.7, the following holds.

Theorem III.9. *Let $\mathscr{R}$ be a set of* MYAPL *rules, with respect to which $\mathscr{P}_C^{\mathscr{R}}$ is the class of models defined by Definition III.22, $\mathscr{P}_L^{\mathscr{R}}$ the class of models defined by Definition III.23, and $\mathscr{P}_P^{\mathscr{R}}$ the class of models defined by Definition III.24. It then holds that*

$$\mathscr{P}_C^{\mathscr{R}} \subseteq \mathscr{P}_L^{\mathscr{R}} \subseteq \mathscr{P}_P^{\mathscr{R}}$$

Proof. Consider Definitions III.22, III.23, and III.24, and note that they differ in the fact that the preconditions of the implications which they respectively enforce are **Obs**(δ), **Late-Obs**(δ, δ'), and **Partial-Obs**(δ, δ'), for particular $\delta, \delta' \in \mathscr{L}_\Delta$. Because $\models$ **Obs**$(\delta) \rightarrow$ **Late-Obs**(δ, δ) it straightforwardly follows that $\mathscr{P}_C^{\mathscr{R}} \subseteq \mathscr{P}_L^{\mathscr{R}}$, since any model which satisfies the conditions of Definitions III.22 satisfies those of III.23. Likewise, from the fact that $\models$ **Late-Obs**$(\delta, \delta') \rightarrow$ **Partial-Obs**(δ, δ') follows $\mathscr{P}_L^{\mathscr{R}} \subseteq \mathscr{P}_P^{\mathscr{R}}$, because any model which satisfies the conditions of Definitions III.23 satisfies those of III.24. □

III.3

The above shows that classes of models reflecting plan-based ascription can be identified for distinct perceptory conditions, which show much of the same properties as did the classes that we identified in Section 2 to reflect ascription based on observable sequences. In that previous section, subclasses of models were furthermore identified that reflect the skeptical and credulous interpretations of mental state abduction considered in Chapter II. In the remainder of the current section, it is done likewise for models reflecting plan-based ascription, albeit with a somewhat different focus.

3.1.1 The Existentially Skeptical Observer

The classes $\mathscr{D}_X^{\mathsf{sk}}$, for $X \in \{C, L, P\}$, discussed in the previous section, can be considered dynamic counterparts of the skeptical interpretation of mental state abduction, as clarified in Section 2.4.1 and reflected by Theorems III.1, III.3, and III.4. Those classes characterize the observer which reasons about the observed agent's mental state, and it was stated that this observer can be considered 'ultimately skeptical' in the sense that it takes into account that *any* of the actions it has observed may be the first it observes as a result of the agent's intention, and that, accordingly, any mental state which can be derived as explanation (on grounds of the appropriate msa_X function) for the observed sequence of actions should be considered. Those interpretations of the msa_X functions differ in that sense from the abductive interpretations given in Chapter II — specifically, in Theorem II.2 — where the order of observed actions is fixed and it is thus given which action the observer considers as the first action of the observed sequence. In the current section, in line with the fact that the observer may consider some *particular* action as having been the first action of the agent's plan and subsequent actions as belonging to the same plan, classes of models are identified which reflect the ascription of mental states and plans on grounds of particular (fixed) observed sequences. In doing so, an 'existentially skeptical' variety of dynamic interpretations is focused upon, defined as follows.

Definition III.25 (existentially skeptical dynamic interpretation — complete observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules,* msa_C *the mental state abduction function defined in Definition II.26, and $\mathscr{P}_C^{\mathscr{R}}$ the class of models defined in Definition III.22. The class of models $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ reflecting the existentially skeptical dynamic interpretation of mental state abduction on grounds of the observed action sequence $\delta \in \mathscr{L}_\Delta$ is then as follows, under the presumption of complete observation.*

$$\begin{aligned}\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} = \{\mathfrak{M} \in \mathscr{P}_C^{\mathscr{R}} \mid \forall(\gamma,\beta) \in \mathsf{msa}_C(\delta,\mathscr{R}) \forall \pi \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) : \\ ((OS(\pi) = \{\delta\}) \quad \& \quad \mathsf{msa}_C(\delta,\mathscr{R}) = \{(\gamma_1,\beta_1),\ldots,(\gamma_n,\beta_n)\} \neq \emptyset \quad \Longrightarrow \\ \mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow (\langle \pi^- \rangle(\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)) \wedge \\ [\pi^-]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma) \wedge \mathbf{B}_n(\beta_n)))))\}\end{aligned}$$

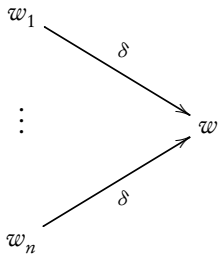
As in preceding sections, the non-triviality of this class can be shown.

III.3

Proposition III.25. *Let $\mathscr{R}$ be a set of* MYAPL *rules, and $\mathscr{P}_C^{\mathscr{R}}$ and $\mathscr{E}^{\mathsf{sk}}_{C[\![\cdot]\!]}$ as defined in Definitions III.22 and III.25 with respect to those rules, respectively. It then holds that*

$$(\mathscr{P}_C^{\mathscr{R}} \neq \emptyset) \qquad \Longleftrightarrow \qquad \exists \delta \in \mathscr{L}_\Delta : \quad (\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} \neq \emptyset)$$

Proof. ($\Rightarrow$) Take the largest $R \subseteq \mathscr{R}$ such that $R = \{1 : \gamma_1 <\!\!- \beta_1 \mid \pi_1, \ldots, n : \gamma_n <\!\!- \beta_n \mid \pi_n\}$ and for $\delta \in \mathscr{L}_\Delta$ holds $\forall m \in \{1,\ldots,n\} \forall (m : \gamma_m <\!\!- \beta_m \mid \pi_m) \in R \exists \pi'_m \in CS(\pi_m) \exists \pi''_m \in \mathscr{L}_\Pi : (\pi''_m \blacktriangleright \pi'_m)$ & $(OS(\pi''_m) = \{\delta\})$. Fix $n = |R|$ and, in regard to the model depicted below, let $\forall m \in \{1,\ldots,n\} : (\mathfrak{M}, w_m \models \mathbf{G}(\gamma_m) \wedge \mathbf{B}(\beta_m))$. Furthermore, given $\pi''_m \in \mathscr{L}_\Pi$ as some prefix of a computation sequence of the plan π_m belonging to the rule identified by m, for which holds $OS(\pi''_m) = \{\delta\}$, let $\mathfrak{M}, w \models \langle \pi''^{-}_m \rangle(\top)$.



See that then $\mathfrak{M} \in \mathscr{P}_C^{\mathscr{R}}$ following Definition III.22, and also $\mathfrak{M} \in \mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$.

($\Leftarrow$) Straightforward, as $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} \subseteq \mathscr{P}_C^{\mathscr{R}}$ for any $\delta \in \mathscr{L}_\Delta$.

□

Thus, models in $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$, for a completely observed sequence δ, share that, for each prefix related to a computation sequence on grounds of which a mental state can be ascribed, there

exists a 'path to the past' reflecting the fact that the observer maintains the existential possibility — in sense of the modality '$\langle\pi^-\rangle$' — of the agent having performed this sequence. As in the previous sections, similar classes of models can be identified for cases of incomplete observation as for the case of complete observation. In Definitions III.26 and III.27 this is done for the relevant subclasses of $\mathscr{P}_L^{\mathscr{R}}$ and $\mathscr{P}_P^{\mathscr{R}}$, where in order to realize the desired 'existentially skeptical' interpretation in the subclasses identified below, it holds for cases of incomplete observation that *all* incompletely observed prefixes, in relation to which some goal/belief pair is inferred, must lead to preceding states. Put otherwise, it cannot be the case that some such prefix is considered late or partially observed in any state, without all other such prefixes being considered likewise. This is formalized in the definitions below, which informally state that the existentially skeptical models under late (partial) observation are those which, in regard to an actually observed sequence δ, satisfy late (partial) observation of the observable part of *all* appropriate prefixes of computation sequences in case they satisfy *any* (as indicated by the series of bi-implications), and for which it furthermore holds that a 'path to the past' exists on grounds of the computation sequence prefix that is derived from the observable sequence which is late (partially) observed.

Definition III.26 (existentially skeptical dynamic interpretation — late observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, msa_L the mental state abduction function defined in Definition II.26, and $\mathscr{P}_L^{\mathscr{R}}$ the class of models defined in Definition III.23. The class of models $\mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]}$ reflecting the existentially skeptical dynamic interpretation of mental state abduction on grounds of the observed action sequence $\delta \in \mathscr{L}_\Delta$ is then as follows, under the presumption of late observation.*

$$
\begin{aligned}
\mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]} = \{\mathfrak{M} \in \mathscr{P}_L^{\mathscr{R}} \mid \forall(\gamma,\beta) \in \mathsf{msa}_L(\delta,\mathscr{R}) : \\
\{\pi' \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) \mid \exists\delta' \in OS(\pi') : (\delta \underline{\blacktriangleleft} \delta')\} = \{\pi_1,\ldots,\pi_n\} \neq \emptyset \quad \& \\
OS(\pi_1) = \{\delta_1\} \;\&\; \ldots \;\&\; OS(\pi_n) = \{\delta_n\} \qquad \Longrightarrow \\
\mathfrak{M} \models (\textbf{Late-Obs}(\delta,\delta_1) \leftrightarrow \textbf{Late-Obs}(\delta,\delta_2)) \wedge \ldots \wedge \\
(\textbf{Late-Obs}(\delta,\delta_{n-1}) \leftrightarrow \textbf{Late-Obs}(\delta,\delta_n)) \quad \& \\
(\forall\pi'' \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) : \quad OS(\pi'') = \{\delta''\} \quad \& \\
\mathsf{msa}_L(\delta,\mathscr{R}) = \{(\gamma_1,\beta_1),\ldots,(\gamma_n,\beta_n)\} \neq \emptyset \qquad \Longrightarrow \\
\mathfrak{M} \models \textbf{Late-Obs}(\delta,\delta'') \rightarrow (\langle\pi''^-\rangle(\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)) \wedge \\
[\pi''^-]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma) \wedge \mathbf{B}_n(\beta_n))))\}
\end{aligned}
$$

Note that the following definition employs $\underline{\odot}$ in place of $\underline{\blacktriangleleft}$, because of the nature of partial observation that allows the observed sequence to be a dilution of an observable sequence which is partially observed. In accordance with the fact that our focus is on the moment of observation, it is furthermore demanded that the observed sequence has a matching suffix with any observable sequence that is considered partially observed.

Definition III.27 (existentially skeptical dynamic interpretation — partial observation)**.** *Let $\mathscr{R}$ be a set of* MYAPL *rules, msa_P the mental state abduction function defined in Definition II.26, and $\mathscr{P}_P^{\mathscr{R}}$ the class of models defined in Definition III.24. The class of models*

$\mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]}$ *reflecting the existentially skeptical dynamic interpretation of mental state abduction on grounds of the observed action sequence* $\delta \in \mathscr{L}_\Delta$ *is then as follows, under the presumption of partial observation.*

$$\begin{aligned}
\mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]} = \{\mathfrak{M} \in \mathscr{P}_P^{\mathscr{R}} \mid \forall(\gamma,\beta) \in \mathsf{msa}_P(\delta,\mathscr{R}) : \\
\{\pi' \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) \mid \exists \delta', \delta'' \in \mathscr{L}_\Delta : (\delta' \in OS(\pi') \;\&\; (\delta \mathbin{\underline{\odot}} \delta') \;\&\; \\
(\delta'' \mathbin{\underline{\blacktriangleleft}} \delta) \;\&\; (\delta'' \mathbin{\underline{\blacktriangleleft}} \delta'))\} = \{\pi_1,\ldots,\pi_n\} \neq \emptyset \quad \& \\
OS(\pi_1) = \{\delta_1\} \;\&\; \ldots \;\&\; OS(\pi_n) = \{\delta_n\} \quad \Longrightarrow \\
\mathfrak{M} \models (\mathbf{Partial\text{-}Obs}(\delta,\delta_1) \leftrightarrow \mathbf{Partial\text{-}Obs}(\delta,\delta_2)) \wedge \ldots \wedge \\
(\mathbf{Partial\text{-}Obs}(\delta,\delta_{n-1}) \leftrightarrow \mathbf{Partial\text{-}Obs}(\delta,\delta_n)) \quad \& \\
(\forall \pi'' \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) : \quad OS(\pi'') = \{\delta''\} \quad \& \\
\mathsf{msa}_P(\delta,\mathscr{R}) = \{(\gamma_1,\beta_1),\ldots,(\gamma_n,\beta_n)\} \neq \emptyset \quad \Longrightarrow \\
\mathfrak{M} \models \mathbf{Partial\text{-}Obs}(\delta,\delta'') \rightarrow (\langle \pi''^{-} \rangle(\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)) \wedge \\
[\pi''^{-}]((\mathbf{G}(\gamma_1) \wedge \mathbf{B}(\beta_1)) \vee \ldots \vee (\mathbf{G}(\gamma) \wedge \mathbf{B}_n(\beta_n))))\}
\end{aligned}$$

Proofs of the non-triviality of the classes defined above, along the lines of Proposition III.25, are omitted for the sake of brevity, and it is assumed that the proof given of the aforementioned proposition, as well as similar one in preceding sections, suffices to convince.

3.2 Inconclusiveness

Models in the classes $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$, $\mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]}$, and $\mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]}$ in some cases, depending on the underlying set of rules and $\delta \in \mathscr{L}_\Delta$, force inconclusiveness in ascription. In particular, this occurs with rules that force ascription of contradictory literals on which the agent presumedly — given the observed actions and matching plans — performed successful test actions. This is illustrated as follows, reflecting the intuition that an observer reasoning with such models is forced to admit, given the constraints set by its interpretation of the agent's rules, that it cannot be certain about some aspects of the agent's mental state, if it follows up on its grounds to ascribe particular goals or beliefs.

Proposition III.26. *Let* $\mathbf{R}$ *be the domain of* MYAPL *rules, and* $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ *the class of models identified in Definition III.25 with respect to some set of rules* $\mathscr{R} \subseteq \mathbf{R}$ *and sequence of actions* $\delta \in \mathscr{L}_\Delta$*. It then holds, recalling* $\mathbf{BInc}(p) \triangleq \mathbf{B}(p) \wedge \mathbf{B}(\sim p)$*, that*

$$\begin{aligned}
\exists \mathscr{R} \subseteq \mathbf{R} \exists \delta \in \mathscr{L}_\Delta \exists \phi \in \mathscr{L}_0 : \\
\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} \models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\phi)
\end{aligned}$$

Proof. Let $\mathscr{R} = \{1 : q \mathrel{<\!\!-} r \mid a; \mathbf{B}(p)?, 2 : q' \mathrel{<\!\!-} r' \mid a; \mathbf{B}(\sim p)?\}$, observing $CS(a; \mathbf{B}(p)?) = \{a; \mathbf{B}(p)?\}$ and $CS(a; \mathbf{B}(\sim p)?) = \{a; \mathbf{B}(\sim p)?\}$, and $OS(a; \mathbf{B}(p)?) = OS(a; \mathbf{B}(\sim p)?) = \{a\}$. Take any $\mathfrak{M} \in \mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ and $w \in \mathrm{W}_{\mathfrak{M}}$, and assume that $w \models \mathbf{Obs}(a)$ is the case. Note

that Definition III.25 forces $w \models \langle a; \mathbf{B}(p)?^{-}\rangle(\mathbf{G}(q) \wedge \mathbf{B}(r))$ and $w \models \langle a; \mathbf{B}(\sim p)?^{-}\rangle(\mathbf{G}(q') \wedge \mathbf{B}(r'))$. If that is so, then it must also be the case that both $w \models \mathbf{B}(p)$ and $w \models \mathbf{B}(\sim p)$, i.e. $w \models \mathbf{B}(p \wedge \sim p)$, i.e. $w \models \mathbf{BInc}(p)$. Since $\mathfrak{M}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ are arbitrary, $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} \models \mathbf{Obs}(a) \rightarrow \mathbf{BInc}(p)$ is proven for the above $\mathscr{R}$. □

Proposition III.27.

$$\exists \mathscr{R} \subseteq \mathbf{R} \exists \delta \in \mathscr{L}_{\Delta} \exists \phi \in \mathscr{L}_{0} : \\ \mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} \models \mathbf{Obs}(\delta) \rightarrow \mathbf{GInc}(\phi)$$

Proof. Straightforward, along the lines of Proposition III.26, given the set of MYAPL rules $\mathscr{R} = \{1 : q \mathrel{<\!\!-} r \mid a; \mathbf{G}(p)?, 2 : q' \mathrel{<\!\!-} r' \mid a; \mathbf{G}(\sim p)?\}$. □

For brevity, the gist of the preceding two propositions is taken together below in a single proposition for the case of late observation, and then likewise for the case of partial observation. Note again that '||' denotes disjunction in the metalanguage.

III.3

Proposition III.28.

$$\exists \mathscr{R} \subseteq \mathbf{R} \exists \delta, \delta' \in \mathscr{L}_{\Delta} \exists \phi \in \mathscr{L}_{0} : \\ \mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]} \models \mathbf{Late\text{-}Obs}(\delta, \delta') \rightarrow \mathbf{BInc}(\phi) \quad || \\ \mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]} \models \mathbf{Late\text{-}Obs}(\delta, \delta') \rightarrow \mathbf{GInc}(\phi)$$

Proof. Straightforward, along Propositions III.26 and III.27. □

Proposition III.29.

$$\exists \mathscr{R} \subseteq \mathbf{R} \exists \delta, \delta' \in \mathscr{L}_{\Delta} \exists \phi \in \mathscr{L}_{0} : \\ \mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]} \models \mathbf{Partial\text{-}Obs}(\delta, \delta') \rightarrow \mathbf{BInc}(\phi) \quad || \\ \mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]} \models \mathbf{Partial\text{-}Obs}(\delta, \delta') \rightarrow \mathbf{GInc}(\phi)$$

Proof. Straightforward, along Propositions III.26 and III.27. □

Propositions III.26–III.29 show that sets of rules exist that force inconclusiveness on part of the observer under particular interpretations. The proof mentions a non-trivial example of such a set, comprising two distinct rules which in themselves do not force inconclusiveness under the interpretations of Definitions III.22–III.24 but in combination do. A more trivial example can also be given; consider e.g. the rule $(1 : q \mathrel{<\!\!-} r \mid a; \mathbf{B}(p \wedge \sim p)?)$. It should be clear that if the observer maintains the possibility that the agent successfully performed the test action $\mathbf{B}(p \wedge \sim p)?$ then it cannot but admit that the agent believed both p and $\sim p$. In software agents whose beliefs or goals have standard logical semantics, the above rule containing an inconsistent test action should be considered insensible, though, because the corresponding plan can only fail. Reasoning about such faulty rules is not the topic of this dissertation, and it can safely be assumed, without loss of generality, that the expressions

comprising preconditions of rules, as well as the expressions which are the object of test actions in plans, are in themselves non-contradictory (i.e. it is assumed that they are satisfiable under a classical logical interpretation).

In the following proposition a class of models is defined in which the kind of inconclusiveness sketched above does not occur; it is only given as part of that proposition for illustration, as in the next section this same matter is approached differently.

Proposition III.30. *A non-trivial class of models* $\mathscr{E}^{\mathsf{sk+}}_{C[\![\delta]\!]} \subseteq \mathscr{P}^{\mathscr{R}}_{C}$ *can be identified, given a set of* MYAPL *rules* $\mathscr{R}$ *(cf. Definition III.25), as follows.*

$$\begin{aligned}\mathscr{E}^{\mathsf{sk+}}_{C[\![\delta]\!]} = \{\mathfrak{M} \in \mathscr{P}^{\mathscr{R}}_{C} \mid \forall(\gamma,\beta) \in \mathsf{msa}_C(\delta,\mathscr{R}) \forall \pi,\pi' \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) : \\ (OS(\pi) = \{\delta\} \quad \& \quad ((\pi' \underline{\blacktriangleright} \pi) \,\&\, (OS(\pi') = \{\delta\}) \Rightarrow (\pi \underline{\blacktriangleright} \pi'))) \quad \Longrightarrow \\ \mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow \langle \pi^- \rangle (\mathbf{G}(\gamma) \wedge (\mathbf{B}(\beta)))\}\end{aligned}$$

The following then holds (cf. Propositions III.26–III.27).

$$\begin{aligned}\forall \mathscr{R} \in \mathbf{R} \forall \delta \in \mathscr{L}_{\Delta} \forall \phi \in \mathscr{L}_0 : \\ \mathscr{E}^{\mathsf{sk+}}_{C[\![\delta]\!]} \not\models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\phi) \qquad \& \qquad \mathscr{E}^{\mathsf{sk+}}_{C[\![\delta]\!]} \not\models \mathbf{Obs}(\delta) \rightarrow \mathbf{GInc}(\phi)\end{aligned}$$

Proof. The difference between $\mathscr{E}^{\mathsf{sk+}}_{C[\![\delta]\!]}$ and $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ as identified in Definition III.25 lies in the fact that above the requirement $\forall \pi,\pi' \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) : (\pi' \underline{\blacktriangleright} \pi \;\&\; OS(\pi') = \{\delta\}) \Rightarrow (\pi \underline{\blacktriangleright} \pi')$ ensures that π is the *shortest* prefix of computation sequences of plans accompanying γ,β for which still holds that $OS(\pi) = \{\delta\}$. In effect, this means that $\pi = \pi'';\alpha$ for some $\alpha \in \mathsf{Act}$ and (possibly empty) $\pi'' \in \mathscr{L}_{\Pi}$. Observe that the proof of Proposition III.26 depends on the fact that in models $\mathfrak{M} \in \mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ holds that the test actions, which are the suffix of any computation sequence prefix of which the observable sequence is δ, succeed in states $w \in \mathrm{W}_{\mathfrak{M}}$ for which holds $w \models \mathbf{Obs}(\delta)$. In case of test actions on contradictory propositions (from different rules) this forces inconclusiveness. The shortest such prefix, however, does not have a suffix of test actions, so that inconclusiveness is not forced. □

The above proposition shows that under a variant of the class of models identified in Definition III.25 inconclusiveness in ascription does not necessarily occur in the way shown by Propositions III.26–III.29. This class of models $\mathscr{E}^{\mathsf{sk+}}_{C[\![\delta]\!]} \subseteq \mathscr{P}^{\mathscr{R}}_{C}$ defined as part of Proposition III.30 provides a 'slightly more skeptical' interpretation than the class of models $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$, because ascription of facts pertaining to the observed agent's mental state is restricted to only those needed to account for the agent's hitherto observed behavior, so that the observer can be said to prefer the shortest matching prefix as basis for ascription. This alternative interpretation has the benefit that it does not force inconclusiveness on part of the observer in cases where this did occur under the interpretation it replaces. It is our conjecture that this holds similarly for cases of incomplete observation, but because this is not the main focus of this section formal mention of this fact is omitted. Instead, because there are also cases in which inconclusiveness is unavoidable, it is those cases which are focused upon next.

3.2.1 Presuming Goal Achievement

In agent programming, goals are a construct for implementation of proactive behavior, and in the language MYAPL rules of the form $n : \gamma <- \beta \mid \pi$ specify that the plan π can be selected by the agent to achieve the goal γ if it believes the fact β to be the case. Accordingly, if the agent indeed has the goal γ and belief β at some point and selects the plan π in order to achieve this goal, it can be presumed to *believe* that γ is the case after performing π successfully. This is formalized as follows.

Definition III.28 (presumed goal achievement)**.** *Given a set of* MYAPL *rules $\mathscr{R}$ and $\mathscr{M}$ as the class of all models for $\mathscr{L}_{\mathbf{M}}$, $\mathscr{G}^{\mathscr{R}}$ is the class of models reflecting presumption of the agent's belief in goal achievement, as follows.*

$$\forall (n : \gamma <- \beta \mid \pi) \in \mathscr{R} :$$
$$\mathscr{G}^{\mathscr{R}} = \{\mathfrak{M} \in \mathscr{M} \mid \mathfrak{M} \models (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)) \rightarrow [\pi]\mathbf{B}(\gamma)\}$$

III.3

It may be reasonable in some cases for the observer to treat the agent as if it believes γ after performing π in the situation sketched above, even if it is not known whether this is actually the case. This holds, for example, if it is known (e.g. because the agent's design specification is available) that the agent should have been programmed to behave in a certain way to achieve some goal, *without it having explicit representations of goals or beliefs*. In such a case, the above presumption of goal achievement can be a useful way to reason about the agent's subsequent behavior. This line of reasoning can be compared to Dennett's claim for the legitimacy of the 'intentional stance', as discussed in Section 1.1 of Chapter I, the success of which depends mostly on attributing mental states to the agent that result in successful prediction of its future behavior. In cases where agents have an explicit belief-type representation of goals they have achieved — as in the case of 2APL agents (Dastani, 2008) — this legitimacy is beyond question, in the sense that the attributed belief can be verified. Note, though, that the presumption of goal achievement, as defined in Definition III.28, is really to be regarded as a *heuristic* for interpreting agents' behavior, because it depends on the implementation of the agent whether or not this property actually holds. In any case, given the observer's presumption that the agent believes the facts it had as goals to achieve to hold after completing related plans, it is in some cases unavoidable that the observer is inconclusive about the agent's mental state. In order to make it easier to see that this is so, consider first the following propositions, which show general relations between plans and their computation sequences.

Proposition III.31.

$$\forall \pi \in \mathscr{L}_{\Pi} \forall \pi' \in CS(\pi) \forall \phi \in \mathscr{L}_{\mathbf{M}} :$$
$$\models [\pi]\phi \rightarrow [\pi']\phi$$

Proof. Take any model $\mathfrak{M}$, state $w \in \mathrm{W}_{\mathfrak{M}}$, and expressions $\pi \in \mathscr{L}_{\Pi}$ and $\phi \in \mathscr{L}_{\mathbf{M}}$. Semantics dictate that $\mathfrak{M}, w \models [\pi]\phi$ if $\forall (w, w') \in \varrho(\pi) : (w' \models \phi)$. From $\pi' \in CS(\pi)$ follows $\varrho(\pi') \subseteq \varrho(\pi)$, since π' represents a single possible execution of π. Thus, if it holds that

$\forall(w,w') \in \varrho(\pi) : (w' \models \phi)$, it also holds that $\forall(w,w'') \in \varrho(\pi') : (w'' \models \phi)$, so that $\mathfrak{M}, w \models [\pi']\phi$. Observe that $\mathfrak{M}$ and w are arbitrary, proving the claim. □

Proposition III.32.

$$\forall \pi \in \mathscr{L}_{\Pi} \forall \pi' \in CS(\pi) \forall \phi \in \mathscr{L}_{\mathbf{M}} : \\ \models [\pi^{-}]\phi \rightarrow [\pi'^{-}]\phi$$

Proof. Along the lines of the proof of Proposition III.31. □

Proposition III.33.

$$\forall \pi \in \mathscr{L}_{\Pi} \forall \pi' \in CS(\pi) \forall \phi \in \mathscr{L}_{\mathbf{M}} : \\ \models \langle \pi' \rangle \phi \rightarrow \langle \pi \rangle \phi$$

Proof. Take any model $\mathfrak{M}$, state $w \in \mathrm{W}_{\mathfrak{M}}$, and expressions $\pi \in \mathscr{L}_{\Pi}$ and $\phi \in \mathscr{L}_{\mathbf{M}}$. Semantics dictate that $\mathfrak{M}, w \models \langle \pi' \rangle \phi$ if $\exists(w,w') \in \varrho(\pi') : (w' \models \phi)$. From $\pi' \in CS(\pi)$ follows $\varrho(\pi') \subseteq \varrho(\pi)$, since π' represents a single possible execution of π. Thus, if it holds that $\exists(w,w') \in \varrho(\pi') : (w' \models \phi)$, it also holds that $\exists(w,w'') \in \varrho(\pi) : (w'' \models \phi)$, so that $\mathfrak{M}, w \models \langle \pi \rangle \phi$. Observe that $\mathfrak{M}$ and w are arbitrary, proving the claim. □

Proposition III.34.

$$\forall \pi \in \mathscr{L}_{\Pi} \forall \pi' \in CS(\pi) \forall \phi \in \mathscr{L}_{\mathbf{M}} : \\ \models \langle \pi'^{-} \rangle \phi \rightarrow \langle \pi^{-} \rangle \phi$$

Proof. Along the lines of the proof of Proposition III.33. □

In case it is assumed by the observer that the agent believes the facts it had as goal after successfully using an applicable rule, observe that the following holds, where $\overline{\psi}$ denotes the literal inversion operator of ψ defined in Section 1 of this chapter (recalling that $\overline{p} = {\sim}p$ and $\overline{{\sim}p} = p$).

Theorem III.10. *Let $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ be the class of models identified in Definition III.25 for some $\delta \in \mathscr{L}_{\Delta}$ and $\mathscr{G}^{\mathscr{R}}$ as defined in Definition III.28, both with respect to the set of* MyAPL *rules $\mathscr{R}$, and let $\mathscr{G}\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} = \mathscr{G}^{\mathscr{R}} \cap \mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ denote the intersection of those classes. Furthermore, let* $\mathsf{Lit} = \{p, {\sim}p \mid p \in \mathsf{Atom}\}$. *It then holds that*

$$\forall(n : \gamma <\!\!- \beta \mid \pi), (m : \gamma' <\!\!- \beta' \mid \pi') \in \mathscr{R} \forall \delta \in \mathscr{L}_{\Delta} \forall \psi \in \mathsf{Lit} : \\ (\{\gamma\} \models \psi \;\&\; \{\gamma'\} \models \overline{\psi} \;\&\; \delta \in (OS(\pi) \cap OS(\pi'))) \implies \\ \mathscr{G}\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} \models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\psi)$$

Proof. Take any two $(n : \gamma <\!\!- \beta \mid \pi), (m : \gamma' <\!\!- \beta' \mid \pi') \in \mathscr{R}$ and note that the MYAPL syntax of Definition II.3 defines γ, γ' to be conjoined literals. Assume that for some $\psi \in \mathsf{Lit}$ holds $\{\gamma\} \models \psi$ and $\{\gamma'\} \models \overline{\psi}$. Furthermore, take any $\delta \in \mathscr{L}_\Delta$ and assume that both $\delta \in OS(\pi)$ and $\delta \in OS(\pi')$. Observe that it must then be the case that $\exists \pi'' \in CS(\pi) : (OS(\pi'') = \{\delta\})$ and $\exists \pi''' \in CS(\pi') : (OS(\pi''') = \{\delta\})$. Take any $\mathfrak{M} \in \mathscr{G}\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ and $w \in \mathrm{W}_{\mathfrak{M}}$, and assume $w \models \mathbf{Obs}(\delta)$. From Definition III.25 then follows $w \models \langle \pi''^{-} \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$, as well as $w \models \langle \pi'''^{-} \rangle (\mathbf{G}(\gamma') \wedge \mathbf{B}(\beta'))$. Proposition III.34 shows that it then also holds that $w \models \langle \pi^{-} \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)) \wedge \langle \pi'^{-} \rangle (\mathbf{G}(\gamma') \wedge \mathbf{B}(\beta'))$, and Definition III.22 entails $\mathscr{G}\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]} \models ((\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta)) \rightarrow [\pi]\mathbf{B}(\gamma)) \wedge ((\mathbf{G}(\gamma') \wedge \mathbf{B}(\beta')) \rightarrow [\pi']\mathbf{B}(\gamma'))$. Thus, it follows that $w \models \langle \pi^{-} \rangle [\pi] (\mathbf{B}(\gamma))$ and $w \models \langle \pi'^{-} \rangle [\pi'] (\mathbf{B}(\gamma'))$, so that $w \models \mathbf{B}(\gamma) \wedge \mathbf{B}(\gamma')$, i.e. $w \models \mathbf{B}(\psi) \wedge \mathbf{B}(\overline{\psi})$, i.e. $w \models \mathbf{BInc}(\psi)$. Note that $\mathfrak{M} \in \mathscr{G}\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ are arbitrary, proving the claim. □

III.3

Theorem III.11. *Let $\mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]}$ be the class of models identified in Definition III.26 for some $\delta \in \mathscr{L}_\Delta$ and $\mathscr{G}^{\mathscr{R}}$ as defined in Definition III.28, both with respect to the set of* MYAPL *rules $\mathscr{R}$, and let $\mathscr{G}\mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]} = \mathscr{G}^{\mathscr{R}} \cap \mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]}$ denote the intersection of those classes. Furthermore, let $\mathsf{Lit} = \{p, \sim p \mid p \in \mathsf{Atom}\}$. It then holds that*

$$\begin{gathered} \forall (n : \gamma <\!\!- \beta \mid \pi), (m : \gamma' <\!\!- \beta' \mid \pi') \in \mathscr{R} \forall \delta, \delta' \in \mathscr{L}_\Delta \forall \psi \in \mathsf{Lit} : \\ (\{\gamma\} \models \psi \;\&\; \{\gamma'\} \models \overline{\psi} \;\&\; \delta' \in (OS(\pi) \cap OS(\pi'))) \implies \\ \mathscr{G}\mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]} \models \mathbf{Late\text{-}Obs}(\delta, \delta') \rightarrow \mathbf{BInc}(\psi) \end{gathered}$$

Proof. Along the lines of the proof of Theorem III.10, accounting for late observation. □

Theorem III.12. *Let $\mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]}$ be the class of models identified in Definition III.27 for some $\delta \in \mathscr{L}_\Delta$ and $\mathscr{G}$ as defined in Definition III.28, both with respect to the set of* MYAPL *rules $\mathscr{R}$, and let $\mathscr{G}\mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]} = \mathscr{G}^{\mathscr{R}} \cap \mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]}$ denote the intersection of those classes. Furthermore, let $\mathsf{Lit} = \{p, \sim p \mid p \in \mathsf{Atom}\}$. It then holds that*

$$\begin{gathered} \forall (n : \gamma <\!\!- \beta \mid \pi), (m : \gamma' <\!\!- \beta' \mid \pi') \in \mathscr{R} \forall \delta, \delta' \in \mathscr{L}_\Delta \forall \psi \in \mathsf{Lit} : \\ (\{\gamma\} \models \psi \;\&\; \{\gamma'\} \models \overline{\psi} \;\&\; \delta' \in (OS(\pi) \cap OS(\pi'))) \implies \\ \mathscr{G}\mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]} \models \mathbf{Partial\text{-}Obs}(\delta, \delta') \rightarrow \mathbf{BInc}(\psi) \end{gathered}$$

Proof. Along the lines of Theorem III.10, accounting for partial observation. □

Thus, as the above theorems show, if the observer presumes the agent to believe in having achieved its goals after finishing its plans, as in Definition III.28, then this necessarily leads to inconclusiveness in regard to the agent's beliefs, given particular sets of underlying MYAPL rules. This is a logical consequence of the observer being so bold as to ascribe to the agent any mental state derivable from explaining observed actions, along with furthermore

ascribing belief in goal achievement solely on grounds of those observed actions. This is illustrated more poignantly in the example given in Section 3.3, but first in the next section focus is on the fact that if the observer employs other information — specifically, its own beliefs — as grounds for ascription, then inconclusiveness need not occur.

3.2.2 Existentially Skeptical Ascription, Grounded in Beliefs

As illustrated in preceding sections, the observer's presumption of the agent believing to have achieved its goals can force inconclusiveness in the models $\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ that reflect the existentially skeptic interpretation of a particular $\delta \in \mathscr{L}_{\Delta}$. This is not an error, but simply results from the fact that the observer makes attributions solely on grounds of observed behavior. If the observer employs other grounds in ascription, though, then that kind of inconclusiveness can be ruled out. This is illustrated as follows with a class of models in which the observer employs its *own* beliefs in ascription in relation to those it ascribes to the agent; i.e. the observer bases ascription on what it presumes that both itself and the agent believe. It should be noted that the term 'belief' in regard to the observer is restricted to those expressions of $\mathscr{L}_{\mathbf{M}}$ without ascription operators, as our interest here is in comparing arguments to those operators (i.e. expressions from $\mathscr{L}_0$) with the propositional base level of $\mathscr{L}_{\mathbf{M}}$ that expresses particular non-mentalistic facts of which the observer is convinced. In the following definition a function $\tau_0 : \mathscr{L}_0 \longrightarrow \mathscr{L}_{\mathbf{M}}$ is therefore used, which translates $\mathscr{L}_0$-expressions to $\mathscr{L}_{\mathbf{M}}$, retaining logical connectives and replacing $\mathscr{L}_0$-negation with standard negation (i.e. $\tau_0(\sim p) = \neg p$); the translation of some $\phi \in \mathscr{L}_0$ is denoted ϕ^{τ} and interpreted as $\tau_0(\phi)$. This brings propositions constituting the agent's goal/belief preconditions to the level of the observer's 'beliefs', enabling their comparison.

Definition III.29. *Let $\mathscr{R}$ be a set of* MyAPL *rules, and $\mathscr{P}^{\mathscr{R}}_{C}$ be the class of models defined in Definition III.22. The class $\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} \subseteq \mathscr{P}^{\mathscr{R}}_{C}$ is then defined as follows.*

$$\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} = \{\mathfrak{M} \in \mathscr{P}^{\mathscr{R}}_{C} \mid \forall(\gamma,\beta) \in \mathsf{msa}_C(\delta,\mathscr{R}) \forall \pi \in \mathsf{cpfx}(\gamma,\beta,\mathscr{R}) : ((OS(\pi) = \{\delta\}) \implies \mathfrak{M} \models \mathbf{Obs}(\delta) \rightarrow (\langle \pi^{-} \rangle(\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))) \leftrightarrow (\gamma^{\tau} \wedge [\delta^{-}](\neg(\gamma^{\tau}) \wedge \beta^{\tau})))\}$$

Informally, the reasoning process of the observer modeled by $\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$ can be summarized as follows: it attributes the execution of a rule with some goal/belief precondition to the agent on grounds of its observed actions, if and only if it is itself convinced of the fact that the agent's goal is achieved after the observed actions, as well as that before those actions its goal was not achieved, and its belief precondition was fulfilled.

Theorem III.13. *Let $\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$ be the class of models defined in Definition III.29 and $\mathscr{G}^{\mathscr{R}}$ as defined in Definition III.28, both with respect to the set of* MyAPL *rules $\mathscr{R}$, and let $\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} = \mathscr{G}^{\mathscr{R}} \cap \mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$ denote the intersection of those classes. Furthermore, let* $\mathsf{Lit} = \{p, \sim p \mid p \in$

Atom}. *It then holds that*

$$\forall(n:\gamma <\!\!- \beta \mid \pi),(m:\gamma' <\!\!- \beta' \mid \pi') \in \mathscr{R} \forall \delta \in \mathscr{L}_{\Delta} \forall \psi \in \mathsf{Lit}:$$
$$(\{\gamma\} \models \psi \;\&\; \{\gamma'\} \models \overline{\psi} \;\&\; \delta \in (OS(\pi) \cap OS(\pi'))) \implies$$
$$\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} \not\models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\psi)$$

Proof. (*by contradiction*) Assume the preconditions of the '$\implies$'-implication in this theorem to be true, and $\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} \models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\psi)$ to be the case. Given that in general holds $\not\models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\psi)$ for any $\psi \in \mathsf{Lit}$, it must be the case that $\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} \models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\psi)$ holds because of the properties of models in $\mathscr{G}^{\mathscr{R}} \cap \mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$. Take any $p, {\sim}p \in \mathsf{Lit}$ and recall that by definition $\mathbf{BInc}(p) \equiv (\mathbf{B}(p) \wedge \mathbf{B}({\sim}p))$. Furthermore, take any $(n:\gamma <\!\!- \beta \mid \pi),(m:\gamma' <\!\!- \beta' \mid \pi') \in \mathscr{R}$ such that $\{\gamma\} \models p$ and $\{\gamma'\} \models {\sim}p$, and $\delta \in (OS(\pi) \cap OS(\pi'))$. Consider $\mathscr{G}^{\mathscr{R}}$ and $\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$ and see that if $\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} \models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(p)$ is to hold then it must be the case that $\forall \mathfrak{M} \in \mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} \forall w \in \mathrm{W}_{\mathfrak{M}} : (w \models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(p))$. Take any $\mathfrak{M} \in \mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ such that $w \models \mathbf{Obs}(\delta)$ and see that it must then follow that $w \models \mathbf{B}(p)$. This means that, on grounds of $\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$, for some $\pi'' \in CS(\pi)$ such that $OS(\pi'') = \{\delta\}$ holds $w \models \langle \pi''^{-} \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$, so that from Proposition III.34 follows $w \models \langle \pi^{-} \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$. Given the definition of $\mathscr{G}^{\mathscr{R}}$ then follows $w \models \langle \pi^{-} \rangle [\pi] \mathbf{B}(\gamma)$, so that $w \models \mathbf{B}(p)$. Observe that $w \models \langle \pi''^{-} \rangle (\mathbf{G}(\gamma) \wedge \mathbf{B}(\beta))$ furthermore entails $w \models \gamma^{\tau} \wedge [\delta^{-}](\neg(\gamma^{\tau}) \wedge \beta^{\tau})$ on grounds of $\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$. From $w \models \mathbf{BInc}(p)$ it also holds that $w \models \mathbf{B}({\sim}p)$. Analogous to the above, see that then for some $\pi''' \in CS(\pi')$ such that $OS(\pi''') = \{\delta\}$ holds $w \models \langle \pi'''^{-} \rangle (\mathbf{G}(\gamma') \wedge \mathbf{B}(\beta'))$, so that from Proposition III.34 follows $w \models \langle \pi'^{-} \rangle (\mathbf{G}(\gamma') \wedge \mathbf{B}(\beta'))$. Given the definition of $\mathscr{G}^{\mathscr{R}}$ then follows $w \models \langle \pi'^{-} \rangle [\pi'] \mathbf{B}(\gamma')$, i.e. $w \models \mathbf{B}({\sim}p)$ as $\{\gamma'\} \models {\sim}p$. Observe that $w \models \langle \pi'''^{-} \rangle (\mathbf{G}(\gamma') \wedge \mathbf{B}(\beta'))$ furthermore entails $w \models \gamma'^{\tau} \wedge [\delta^{-}](\neg(\gamma'^{\tau}) \wedge \beta'^{\tau})$ on grounds of $\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$. Because of $\{\gamma\} \models p$ and $\{\gamma'\} \models {\sim}p$ it follows that $\{\gamma^{\tau}\} \models p$ and $\{\gamma'^{\tau}\} \models \neg p$. This means that $w \models (p \wedge \neg p) \wedge [\delta^{-}](\neg p \wedge p)$, which, however, is inconsistent. Thus, $\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} \models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\psi)$ cannot hold, so $\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]} \not\models \mathbf{Obs}(\delta) \rightarrow \mathbf{BInc}(\psi)$. □

III.3

Theorem III.13 shows that in the class $\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$ inconclusiveness, as shown for $\mathscr{G}\mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$ in Theorem III.10, does not occur. Thus, by employing its own beliefs in reasoning about the observed agent's mental state, the observer has ruled out inconclusiveness, at the expense of (implicitly, by definition of $\mathscr{G}\mathscr{B}^{\mathscr{R}}_{C[\![\delta]\!]}$) having made additional assumptions; namely that itself and the agent shared beliefs about facts that are of interest for the agent in order to determine its behavior. In some cases it is justifiable for the observer to make such assumptions, but as this is a topic deserving elaborate attention in its own right, it is continued in the next chapter.

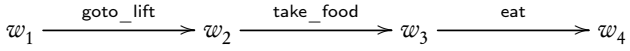


Figure III.5

3.3 Example

This example assumes the setting of the example in Section 2.5, which was first introduced in Section 3.7 of Chapter II. In Listing III.1 (p. 79) three MYAPL rules are listed, which govern the behavior of an observed agent that was stated earlier to be situated in some (virtual) room filled with objects, allowing rich variety in behavior. It is here assumed that in this room a so-called *freight lift*[2] is installed, which functions as follows. The position of the lift is either *up* or *down*, and it is either *loaded* or *empty*. It can only be accessed from the room if it is 'up', and if it is 'loaded' then food can be taken from it. The lift furthermore allows for performing an action of *operating* the lift, triggering a transition from 'up' to 'down' and from 'down' to 'up'. It is assumed that food can only be gotten if requested from the kitchen, and also only if the kitchen is serving.

Given the above, consider the set $\mathscr{R}$ consisting of the three MYAPL rules presented in Listing III.2. In illustrating plan-based ascription, which was put forward in Section 3, focus is first on rule 1. Consider to that extent the following computation sequences of the plan which is part of this rule, assuming translation to $\mathscr{L}_{\Pi}$.

$$\pi_1 = \mathsf{goto\_lift}; \mathbf{B}(\mathsf{lift\_up})?; \mathbf{B}(\mathsf{lift\_loaded})?; \mathsf{take\_food}; \mathsf{eat}$$
$$\pi_2 = \mathsf{goto\_lift}; \mathbf{B}(\mathsf{lift\_up})?; \neg\mathbf{B}(\mathsf{lift\_loaded})?; \mathsf{operate\_lift}; \mathsf{request\_food}$$
$$\pi_3 = \mathsf{goto\_lift}; \neg\mathbf{B}(\mathsf{lift\_up})?; \mathsf{operate\_lift}$$

Let the observed action be 'take_food', and let $\mathfrak{M}$ be the model depicted in Figure III.5. Furthermore, let $w_2 \models \mathbf{Missed}(\mathsf{goto\_lift})$ and $w_3 \models \mathbf{Obs}(\mathsf{take\_food})$. Verify then that, given $\delta = \mathsf{take\_food}$ and $\mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]} \subseteq \mathscr{P}^{\overline{\mathscr{R}}}_{L}$ as identified in Definition III.26, it holds that $\mathfrak{M} \in$

[2] Also called *freight elevator* or *dumbwaiter* in English; *dienstlift* in Dutch, and *Speiseaufzug* in German.

```
1:  food_eaten  <-  kitchen_serving and food_requested  |
    { goto_lift; if B(lift_up) then { if B(lift_loaded) then
      { take_food; eat } else { operate_lift; request_food } }
        else { operate_lift } }

2:  lift_up     <-  at_lift and -lift_up  |  { operate_lift }

3: -lift_up     <  -at_lift and  lift_up  |  { operate_lift }
```

Listing III.2

$\mathscr{E}^{\mathsf{sk}}_{L[\![\delta]\!]}$ if $w_1 \models \mathbf{G}(\mathsf{food\_eaten}) \wedge \mathbf{B}(\mathsf{kitchen\_serving} \wedge \mathsf{food\_requested})$ and $w_2 \models \mathbf{B}(\mathsf{lift\_up} \wedge \mathsf{lift\_loaded})$. Note that also $\mathfrak{M} \in \mathscr{E}^{\mathsf{sk}}_{P[\![\delta]\!]}$ but $\mathfrak{M} \notin \mathscr{E}^{\mathsf{sk}}_{C[\![\delta]\!]}$, because complete observation is not the case seeing that $w_3 \not\models \mathbf{Obs}(\delta)$. Assuming the distribution of atoms considered thus far, let $w_4 \models \mathbf{Obs}(\mathsf{eat})$, and observe that given $\delta' = \mathsf{take\_food;eat}$ it holds that $\mathfrak{M} \in \mathscr{E}^{\mathsf{sk}}_{L[\![\delta']\!]}$ and $\mathfrak{M} \in \mathscr{E}^{\mathsf{sk}}_{P[\![\delta']\!]}$. Consider then the class $\mathscr{G}^{\mathscr{R}}$ as given by Definition III.28, and note that $\mathfrak{M} \in (\mathscr{G}^{\mathscr{R}} \cap \mathscr{E}^{\mathsf{sk}}_{L[\![\delta']\!]})$ if $w_4 \models \mathbf{B}(\mathsf{food\_eaten})$ because w_4 is 'reachable' from w_1 through computation sequence π_1.

In a second scenario, consider the model $\mathfrak{M}'$ depicted on the left in Figure III.6. Assume that $w_7 \models \mathbf{Obs}(\mathsf{operate\_lift})$, and let $\delta'' = \mathsf{operate\_lift}$, verifying that $\mathfrak{M}' \in \mathscr{E}^{\mathsf{sk}}_{C[\![\delta'']\!]}$ if $w_5 \models \mathbf{G}(\mathsf{lift\_up}) \wedge \mathbf{B}(\mathsf{at\_lift} \wedge \mathsf{-lift\_up})$, and $w_6 \models \mathbf{G}(\mathsf{-lift\_up}) \wedge \mathbf{B}(\mathsf{at\_lift} \wedge \mathsf{lift\_up})$. Furthermore, observe that if $\mathfrak{M}' \in (\mathscr{G}^{\mathscr{R}} \cap \mathscr{E}^{\mathsf{sk}}_{C[\![\delta'']\!]})$ then $w_7 \models \mathbf{BInc}(\mathsf{lift\_up})$ as pointed out in Theorem III.10. To the right of $\mathfrak{M}'$, Figure III.6 depicts the model $\mathfrak{M}''$ with states $\{w'_6, w'_7\}$. Assume that w'_6 carries the same information as w_6 in addition to $w'_6 \models \mathsf{at\_lift} \wedge \mathsf{lift\_up}$, and that $\mathfrak{M}'' \in (\mathscr{G}^{\mathscr{R}} \cap \mathscr{B}^{\mathscr{R}}_{C[\![\delta'']\!]})$. Given that $w'_7 \models \mathbf{Obs}(\mathsf{operate\_lift}) \wedge \neg\mathsf{lift\_up}$, see that $w'_7 \models \mathbf{B}(-\mathsf{lift\_up})$ is forced. Note that $w'_7 \models \mathbf{BInc}(\mathsf{lift\_up})$ is not ruled out, though, because Theorem III.13 concerns validity in the *class* $\mathscr{B}^{\mathscr{R}}_{C[\![\delta'']\!]}$ but leaves open that $\mathbf{B}(\mathsf{lift\_up})$ holds contingently in some states belonging to models in this class.

III.3

The example presented thus far should illustrate the use of the approach presented in this chapter for modeling an observer that reasons about a software agent's observed behavior. It also shows, though, that sometimes theory and practice do not agree. Specifically, this can be seen if computation sequences π_2 and π_3 of the plan belonging to rule 1 are taken into consideration, with respect to the class defined in Definition III.28. Take $\delta''' = \mathsf{goto\_lift;operate\_lift}$ and see from Listing III.2 that $OS(\pi_3) = \{\delta'''\}$. Given some $\mathfrak{M}''' \in (\mathscr{G}^{\mathscr{R}} \cap \mathscr{E}^{\mathsf{sk}}_{C[\![\delta''']\!]})$ such that $\mathfrak{M}''', w \models \mathbf{Obs}(\delta''')$, see that $\mathfrak{M}''', w \models \mathbf{B}(\mathsf{food\_eaten})$ follows. Informally put, this occurs because δ''' is the observable part of the computation sequence

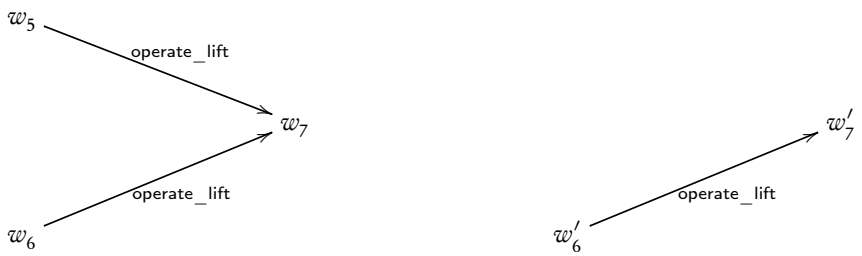


Figure III.6

π_3, such that $\mathfrak{M}''', w \models \langle \pi_3^- \rangle (\mathbf{G}(\mathsf{food\_eaten}) \wedge \mathbf{B}(\mathsf{kitchen\_serving} \wedge \mathsf{food\_requested}))$ follows and the properties of $\mathscr{G}^{\mathscr{R}}$ then force the presumption of the observer that the agent believes to have achieved its goal. This presumption is entirely misplaced, though, because the sequence π_3 does not directly result in achievement of the goal for which its originating plan was selected, although it does further its achievement because the agent operates the lift; presumedly to verify after re-application of this rule whether it contains the requested food. Note that a similar scenario occurs in relation to π_2 if the action request_food is furthermore observed.

Thus, it is in some cases not wise to simply match actions to rules and presume goal achievement. It could be said that this is then the programmer's fault, and that it should be verified whether rules for goal achievement do actually result in the achievement of goals. This argument could then be rebutted, justifying the programmer's practice in context of pragmatics of the programming language; if it is known that rules are reapplied if the goal persists then it is natural to utilize this kind of feature. A possible solution for goal attribution then is to apply the assumption of Definition III.28 only to rules with plans that are known to always result in goal achievement if performed successfully. Alternatively, one can do as in Section 3.2.2 and utilize the observer's beliefs in reasoning about goal achievement. For this, properties of the actions in themselves can be taken into consideration, which is elaborated upon in the next chapter.

4 Reflection

This chapter puts forward a propositional dynamic logic, used for modeling an observer that reasons about a BDI-based agent's observed behavior. The overall goal of the approach is thus globally the same as that of the previous chapter, the main difference being the formalism used. In contrast to the approach using nonmonotonic (abductive) reasoning taken in Chapter II, dynamic logic allows for modeling explanation of observed behavior in terms of a monotonic notion of 'past possibilities' considered by the observer. It is shown in Section 2 that this notion corresponds in certain ways to the abductive theory presented earlier; specifically, that it involves classes of models that represent dynamic interpretations of skeptical or credulous reasoning under different perceptory conditions. In this respect, the first benefit brought by the use of dynamic logic shows through, namely that it makes ascription of rule preconditions in states preceding observed actions explicit. Moreover, the PDL models not only capture actions that have been observed, but also those that (in cases of incomplete observation) have not been observed, yet can be presumed to have occurred. This insight is taken further with the notion of plan-based ascription in Section 3, based on the idea that if the agent's observed actions are presumed to stem from a computation sequence of a particular plan then any test actions occurring in this sequence must have been successful. Thus, the observer has grounds to ascribe additional facts to the agent as having been its beliefs and/or goals. It is shown that this in some cases results in inconclusiveness on part of the observer, in the sense that its assumptions about the agent's behavior force it to ascribe certain beliefs or goals, but do not allow it to rule out facts which are inconsistent. Specifically, this occurs when the agent is assumed to believe facts which it presumedly has

achieved, and it is shown that in such a case the agent can 'project' its own beliefs on the observer in order to rule out inconclusiveness.

Through the use of dynamic logic, increased expressivity is gained in comparison to the preceding chapter, as this formalism allows for modeling change brought about by actions. The foundation for this approach are the mental state abduction functions, for which this chapter presented dynamic interpretations. Those interpretations capture ascription on grounds of a match between actions and observable sequences of plans, but do not involve the semantics of actions in terms of their preconditions and effects. If the observer has the necessary information to enable reasoning about those semantics, then this can provide a useful guideline for reasoning about the observed agent's mental state in relation to observed actions, as shall be seen in the next chapter.

Chapter IV

Mindreading

"This chapter proposes a logical approach to mindreading." Without further clarification such a statement is likely to raise some eyebrows, because of the esoteric connotations the term 'mindreading' may evoke. Therefore it is explained at the beginning of this chapter how the term 'mindreading' is to be interpreted, which is, in short, as "reasoning about the behavior of others through attribution of mental states". This ability has been studied from the points of view of different scientific disciplines, including philosophy and (developmental) psychology. In Section 1 two influential accounts of mindreading are recapitulated, analyzed, and compared. This sets the stage for presenting a more formal approach to mindreading, as attempted in Section 2. Elaborate illustration of this formal approach is given in Section 3 that focuses on a well-known 'false-belief task', and this chapter concludes with the reflection given in Section 4.

1 Psychological Accounts of Mindreading

The human mind has since centuries been a source of interest for scientists from different domains. A particular functional aspect of this mind is the ability to 'make sense' of other humans, and even non-humans, by attributing to them a mental state and reasoning about their behavior on grounds of that attribution. In the course of events this ability has been referred to using (at least) the terms 'folk psychology', 'role-taking', and 'theory of mind'. These terms have particular theoretical connotations which are not generally agreed upon by the scientific community, and lately 'mindreading' has come into fashion as a more "theoretically neutral" term for the basic ability referred to by the aforementioned terms (Nichols & Stich, 2003). In this section the psychological accounts of mindreading given by Baron-Cohen (1995) and Nichols & Stich (2003) are presented in summary. Both accounts feature models of human mindreading and are of interest because they are largely based on the same underlying theory, that of Leslie (1994). These models extend upon Leslie's model in different ways so that it is considered instructive to discuss them both, in order to present a detailed picture of current psychological views on mindreading. Before doing so, however, it is first considered *why* mindreading should be studied in the first place.

1.1 Reasons to Study Mindreading

Use of the term 'mindreading' may, quite appropriately, invoke the connotation of something enigmatic and incomprehensible. It is something most of us do effortlessly on a daily basis; so effortlessly even that we may pass our lives having done so, without ever having stopped to realize that we did. In the words of Dan Sperber: "*attribution of mental states is to humans as echolocation is to the bat*" (Baron-Cohen, 1995, p. 4; quoting Sperber). A reason

to study mindreading can thus be the mere desire to gain understanding of the mechanisms underlying this mysterious ability; i.e. to see how it works. Furthermore, the ability to mindread is essential for the successful functioning of human individuals as part of a social group, a claim which is supported by evidence from pathology showing that people who lack this ability have trouble functioning in society (Baron-Cohen, 1995). Such findings add a sense of relevance and urgency to the endeavor to study and analyze mindreading, because knowledge of how humans mindread may help in alleviating, or even preventing, pathological cases.

Recent work by Nichols & Stich (2003) acknowledges the aforementioned and furthermore lists several other reasons for studying mindreading, motivated from different fields of scientific research. These reasons are here paraphrased as follows:

Behavioral ecology and cognitive ecology: Evolutionary theorists maintain that a capacity for mindreading would be fitness-enhancing, and study of mindreading capacities in different species can yield insights into the ways mindreading may have evolved through natural selection.

Anthropology and cross-cultural psychology: Evidence suggests that the thoughts and feelings of humans are influenced to a large degree by their cultural background, and it is the interest of anthropologists and cross-cultural psychologists to find out to what extent this is the case.

Developmental science: Development of mindreading is considered to be a paradigm of conceptual change, the interest of many scientists in this field, so that the early acquisition of mindreading in light of its apparent complexity spurs interest into determining the extent of innate endowment of the required concepts.

Psychopathology and clinical psychology: The term 'autism' refers to a spectrum of disorders linked to deficiencies in mindreading, and understanding of its workings may shed light on the cause of those disorders. Furthermore, detailed models of mindreading may help in improving the mindreading abilities of patients.

Philosophy: Mindreading is interesting from a philosophical point of view, amongst others, because of the 'Problem of Other Minds'. This pertains to the ontology of the mental states that are attributed to others, and the ontology of mentalistic explanations for others' behavior, the main issue (roughly) being that one can only *presume* others to have a mentalistic experience similar to one's own, but not be certain.

In addition to the above, there is (at least) one other field of scientific research interested in studying mindreading that is not mentioned by Nichols & Stich (2003), which we ourselves describe as follows.

Artificial intelligence: Mindreading is an important ability for various artificially intelligent systems interacting with humans, e.g. sociable robots (Breazeal, 2002). If such systems are to be pervasive in human society then research should be done on design and implementation of this ability. Also, results of attempting to endow artificial systems with mindreading capabilities may translate to insights in the fields mentioned by Nichols & Stich (2003).

It should be clear that the main reason for which mindreading is studied in this dissertation

is the one mentioned last; that of artificial intelligence. As stated before, this field is not mentioned by Nichols & Stich, despite the fact that past decades have witnessed significant work which can be categorized as being related to this topic (see Section 2 of Chapter VI, specifically).

In this chapter the topic of mindreading is approached formally, using the language of the previous chapter as basis, with the objective of identifying formal 'patterns' of mindreading. Having a formal language that allows for the specification of mindreading also enables, or at least facilitates, its implementation in an artificial system; be it a sociable robot or virtual character. Apart from that, formal accounts of mindreading can also aid researchers of mindreading, and possibly patients of mindreading disorders as well. Researchers may benefit from a clear and unambiguous vocabulary with which to broach the subject, whereas patients may benefit from a system of regularities helping to develop or train their mindreading skills, and both groups can benefit from software tools that support their endeavors in those respective directions. An interesting fact in this regard is that experience with teaching autistic patients to mindread indicates that best results are obtained in settings that involve high degree of structure (Howlin et al., 1999), as can be offered by a software-based approach. Moreover, there is evidence that children with autism are more inclined to interact with robotic playmates than with human ones (Feil-Seifer & Matarić, 2005; Dautenhahn & Werry, 2004), suggesting a potential application of such playmates as vehicles for teaching mindreading to autistic children.

1.2 Models of Human Mindreading

In this section two notable models of mindreading by humans are briefly discussed, with application to artificial intelligence in mind. First, the model of Baron-Cohen (1995) is discussed in Section 1.2.1. This model is selected because it is put forward by a leading authority on the subject of mindreading, and as such is often cited in the literature. Also, it focuses explicitly on aspects of mindreading which are not so apparent in the second model discussed here, that by the cognitive scientists Nichols & Stich (2003). This latter model is of interest because it is presents an integration of recent work on mindreading, attempting to position this capacity in context of a general cognitive architecture, as discussed in Section 1.2.2.

1.2.1 The Model by Baron-Cohen (1995)

The model of Baron-Cohen, depicted in Figure IV.1, comprises four separate components that Baron-Cohen considers to be essential in the human mindreading system (1995). This model has its roots in developmental (patho)psychology and was developed by Baron-Cohen, based on the model proposed by Leslie (1994) in an attempt to account for the phenomena encountered by Baron-Cohen in his capacity as a developmental psychologist specializing in autism. The four components of this model can be described as follows, paraphrasing Baron-Cohen (1995, Chapter 4):

ID (Intentionality Detector): The ID is a perceptual device that interprets motion stimuli in terms of the primitive volitional mental states of 'goal' and 'desire'.

EDD (Eye-Direction Detector): The EDD has three basic functions: detecting eyes, detecting the direction of eyes, and interpreting gaze as 'seeing'.

SAM (Shared-Attention Mechanism): The SAM builds representations of shared attention by comparing the self's perceptual state with the other's perceptual state.

ToMM (Theory-of-Mind Mechanism): The ToMM is a system for inferring the full range of mental states from behavior — that is, for employing a 'theory of mind'.

Baron-Cohen presents evidence from multiple sources for the claim that the four components of his model constitute distinct factors in the mindreading process, but mentions that the complete mindreading system most likely comprises more components than the ones he identifies. The four components he mentions are related, as shown in Figure IV.1, where the arrows indicate that ID and EDD serve as input to SAM, whose output serves as input to ToMM. In Baron-Cohen's view the components of the human mindreading system build representations of increasing complexity. Because it is of particular interest to this chapter, his view on this system — and the function of the ToMM-component specifically — is quoted below (1995, p. 51).

> *ToMM is a system for inferring the full range of mental states from behavior — that is, for employing a "theory of mind". So far, the other three mechanisms have got us to the point of being able to read behavior in terms of* volitional mental states *(desire and goal) and to read eye direction in terms of* perceptual mental states *(e.g., see). They have also got us to the point of being able to verify that different people can be experiencing these particular mental states about the same object or event (shared attention). But a theory of mind, of course, includes much more. The first thing that is still needed is a way of representing the set of* epistemic mental states *(which include pretending, thinking, knowing, believing, imagining, dreaming, guessing and deceiving). The second is a way of tying together all these mental-state concepts (the volitional, the perceptual, and the epistemic) into a*

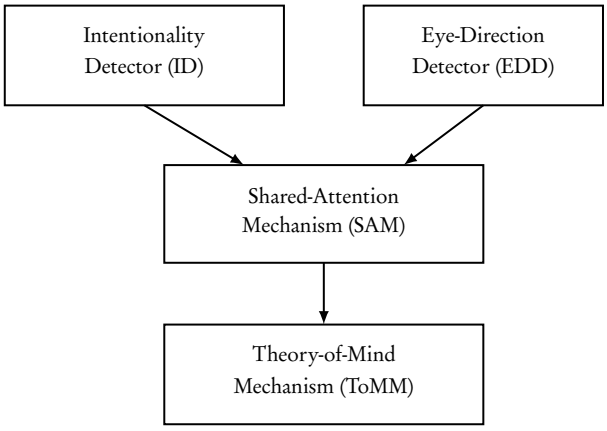


Figure IV.1: (reproduced from (Baron-Cohen, 1995))

coherent understanding of how mental states and actions are related. ToMM does just these things. It has the dual function of representing the set of epistemic mental states and turning all this mentalistic knowledge into a useful theory.

Baron-Cohen's model of human mindreading is a refinement of that of Leslie (1994), with whom he shares the view that ToMM processes so-called *M-representations*. Such representations are considered by them to be the basic type of representation used in mindreading, and have the general form [Agent-Attitude-"Proposition"], for example [John-believes-"it is raining"] or [Mary-thinks-"my marble is in the basket"] (Baron-Cohen, 1995, p. 51–52). Further on in this chapter (in Sections 1.2.3 and 1.3) this model is discussed in more detail, as is the notion of M-representations, but before doing so another model of mindreading is considered.

1.2.2 The Model by Nichols & Stich (2003)

A recent model of human mindreading, depicted in Figure IV.2, that has been put forward by Nichols & Stich (2003) attempts to place mindreading in the broader context of a human cognitive architecture. The authors are cognitive scientists (philosophers), and as such their model is not as directly linked to clinical experience as that of Baron-Cohen, which is discussed in Section 1.2.1. Instead, Nichols & Stich have performed systematic analysis of existing accounts of the phenomenon they refer to as 'mindreading', employing sources from a broad range of scientific domains. As a result of this analysis they present their own account, which attempts to integrate existing accounts into a unifying theory. In doing so, Nichols & Stich adhere to a principle of parsimony, in the sense that they try and show how cognitive mechanisms that supposedly existed before the evolutionary beginning of mindreading could have been put to use for mindreading at a later stage. The approach of Nichols & Stich (2003) is described by themselves as follows: "*Our aim is to characterize the complex and variegated skills that constitute our mindreading capacity and to begin the job of explaining how these skills are accomplished by positing a cluster of mental mechanisms that interact with one another in the ways that we will specify. [...] more often than not we will simply characterize the function of [these mechanisms], and their interactions, in as much detail as we can, and leave it to others to figure out how [they] carry out their function.*" (2003, p.10).

As should become apparent from comparison of Figures IV.1 and IV.2, the model depicted of Nichols & Stich consists of quite a few more components. These components comprise the "full third-person mindreading system" according to Nichols & Stich (2003), but are not all specialized for the task of mindreading. Following the principle of parsimony, the authors start out with an assumption about the basic human cognitive architecture and extend this architecture with components needed to account for mindreading. The basic architecture has the components whose function is described below, paraphrasing Nichols & Stich (2003):

Beliefs: Belief is a basic representational state, caused directly by perception or derived by (non-)deductive inference from existing beliefs.

Desires: Desire is a basic representational state, caused by systems that monitor bodily

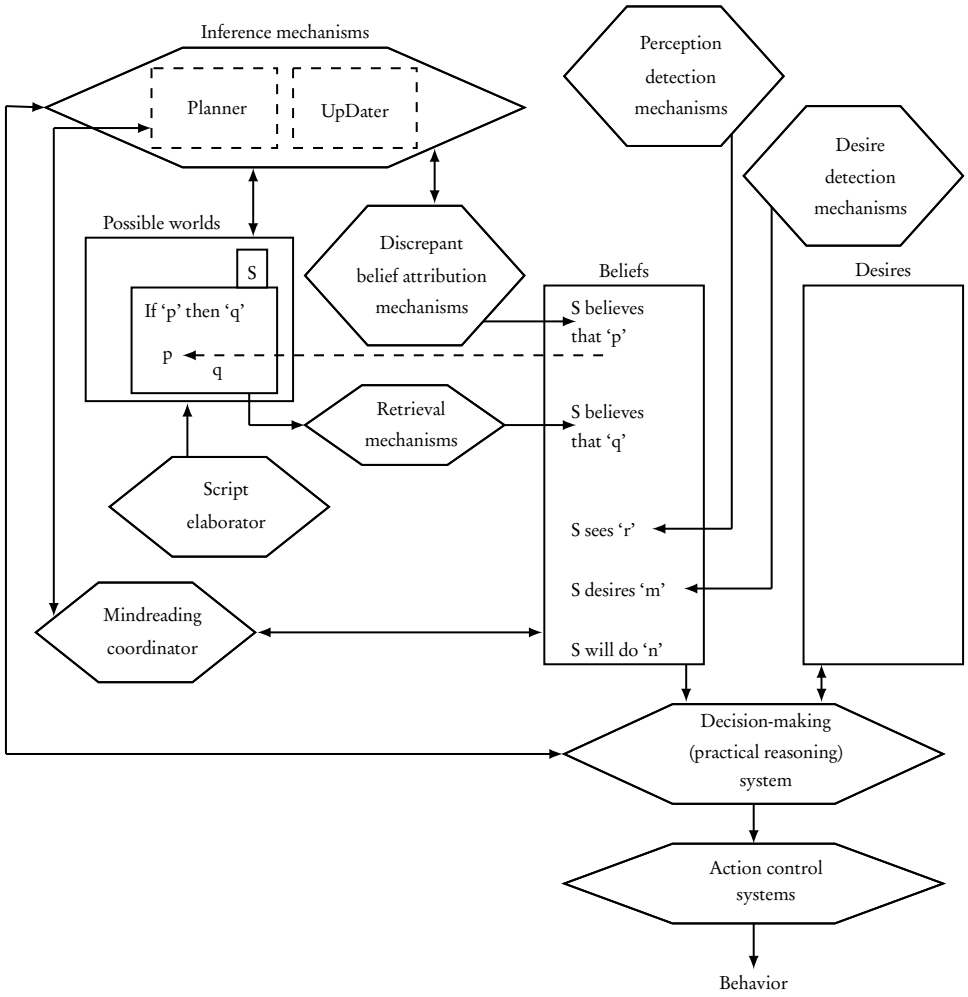


Figure IV.2: (reproduced from (Nichols & Stich, 2003))

states, or generated by a process of practical reasoning using beliefs and existing desires.

Inference mechanisms: Mechanisms for reasoning and inference capable of, amongst others, means-end reasoning in order to construct plans for goal achievement.

Decision-making (practical reasoning) system: Monitors agents' desires and goals in order to request plan construction if required, accepting plans which are compatible with existing desires and generating new (instrumental) desires for intermediate steps of the plans if necessary.

Action control systems: Systems responsible for generating physical behavior of the agent.

Given the basic architecture thus sketched, Nichols & Stich (2003) consider the phenomenon of *pretence* and identify the necessary components to extend this architecture in order for it to account for that phenomenon. Pretend play is a quite complex activity in which children agree upon the fact that certain things are the case, whereas in reality they are not. For example, a child may pretend that a banana is a telephone and treat it accordingly, or a group of children may pretend to be having a tea party and sip tea from empty cups. Nichols & Stich consider pretence to require mechanisms that are also used for mindreading, and identifying those mechanisms leads them to extend the basic cognitive architecture with the following components:

Possible worlds: The so-called 'Possible World Box' (PWB) has the function of representing what the actual world would be like if some set of assumptions were true. Inference mechanisms that operate on beliefs also operate on elements in the PWB, which are assumed to have the same representational properties as beliefs. It is hypothesized that in pretence the PWB is filled with the premise of a pretence episode, along with 'real' beliefs that are appropriately updated in regard to this premise.

UpDater: A (cluster of) mechanism(s) responsible for updating 'real' beliefs, which also works on representations in the PWB. It is hypothesized that in a pretence episode the beliefs of the pretender are filtered through the UpDater in the process of being copied to the PWB, so that they are in line with the premise of the pretence episode.

Script elaborator: A (cluster of) mechanism(s) that fills in the details of a pretence episode which cannot be inferred from the premise of this episode or the UpDater-filtered contents of the PWB.

The basic cognitive architecture is thus extended to account for pretence, but Nichols & Stich (2003) find that it must be extended even further to account for full-fledged mindreading. To this extent they employ a checklist of facts which should be explained by theories of mindreading, and put forward three distinct models of mindreading that represent three stages in the development of the human mindreader. The full (i.e. Stage 3) system is depicted in Figure IV.2, of which the following components comprise the early (Stage 1) mindreading system in addition to those already mentioned:

Desire detection mechanisms (DDM): A (cluster of) mechanism(s) for attributing desires to others, differentiated by Nichols & Stich (2003) into various strategies. One of the oldest of these strategies (in evolutionary terms) employs a variety of cues for determining a target's goals from its observed non-verbal behavior. Another strategy specializes in desire attribution based on observed facial expressions, and yet a third relies on what others *say* about the target's desires. The last strategy mentioned is one that generalizes the mindreader's own case to others, assuming them to have identical desires.

Planner: The mechanism responsible for the mindreader's own means-end reasoning is employed in mindreading for determining a plan of action that achieves the desire attributed to the observed target on grounds of the DDM.

Mindreading coordinator: This mechanism generates predictions about the target's behavior, based on the outcome of planning for the target's detected desire.

Nichols & Stich (2003) call the early mindreading system the 'Desire and Plan' system. Their claim is that it is followed in development by the 'PWB and Desire' system (Stage 2), which has the same components but differs from the Desire and Plan system mainly in the fact that the PWB and the script elaborator are utilized in mindreading. It is hypothesized that mindreaders in this stage of development use the PWB to perform *default belief attribution*, meaning that they attribute the beliefs they themselves hold to the target of mindreading. Because typically not all beliefs are shared, the model of full-fledged mindreading furthermore has the following components:

Perception detection mechanisms (PDM): In order to account for the capacity to form beliefs about a target's perceptual states, it is argued that a (cluster of) mechanism(s) must exist which utilizes information about the target and the environment to produce beliefs about the target's perception.

Discrepant belief attribution mechanisms: This (cluster of) mechanism(s) serves to 'override' the default attribution of beliefs by detecting which of the target's beliefs presumedly differ from those of the mindreader. The simplest strategy for doing so relies on negative perception attribution, which can be described as "*target did not see* X *so target does not believe/know* X". The counterpart of this strategy is positive perception attribution, and can be described as "*target saw* X *so target believes/knows* X". Also, given beliefs about the target's desires, discrepant beliefs can be determined through explanation of the target's observed behavior. Last but not least, language is a significant resource for determining targets' (discrepant) beliefs. Verbal assertion by targets of their beliefs is an obvious source for determining the beliefs of others, and third-person reporting ('gossip') is another such source.

Retrieval mechanisms: In order to reap the benefits of having an evolved model of a target's mental state in its PWB the mindreader requires mechanisms that retrieve elements from the PWB in the appropriate fashion, translating them into suitable counterparts in terms of the mindreader's own beliefs.

The above description of the components of the mindreading model posited by Nichols & Stich (2003) provides an introduction to this model, which by necessity is brief as a more thorough presentation lies outside the scope of this dissertation. For a detailed exposition the reader is referred to the original work, in which this model is given support with arguments from diverse sources ranging from empirical evidence to evolutionary accounts, and in which the model is also compared to existing alternatives. This model and that given by Baron-Cohen (Section 1.2.1) are now compared in order to emphasize differences that are important from the perspective of our A.I. approach.

1.2.3 Comparing the Models by Baron-Cohen and Nichols & Stich

In this section two models of human mindreading have been exposited, based on the original works in which they are presented. It is noteworthy in this respect that the authors of those respective works mention the model of mindreading put forward by Leslie as having largely influenced their own; because the models which we have discussed can be considered enhancements of Leslie's, the latter is not discussed in detail but only referred to where

necessary. Leslie (1994) posits the mechanisms ToMM and ToBy as specialized modules that represent a theory of mind and a theory of mechanical bodies, respectively. Baron-Cohen (1995), on the other hand, posits the mechanisms ToMM, ID, EDD, and SAM, stating that his model largely agrees with that of Leslie but differs in some respects on grounds of evidence from various sources which are left undiscussed here. Furthermore, as stated earlier, Baron-Cohen suggests that the four mechanisms he identifies present the bare minimum required for mindreading, considering it likely that there are more (1995, p. 32).

Nichols & Stich also admit strong influence by Leslie, stating the following in relation to Leslie's account of pretence (Leslie, 1987): "*part of our theory is a notational variant of part of his — and this is no accident since Leslie's work has been an important influence on our own*" (Nichols & Stich, 2003, p. 50). There are also notable differences between the respective theories, though, of which a crucial one is the fact that Leslie purports pretence-subserving representations to differ structurally from 'regular' beliefs but to be treated in the same manner, whereas Nichols & Stich take an orthogonal standpoint and claim such representations to have the same "logical form" as regular beliefs yet to be treated differently; specifically, to be stored in the PWB.

Given the theoretical foundation, in terms of Leslie's work, which the models of Baron-Cohen and Nichols & Stich largely share, it is interesting to note the way in which these models differ, yet also complement each other. The model of Baron-Cohen puts emphasis on *perception*, with which both the mechanisms ID and EDD are principally concerned. ID is hereby claimed to be multi-modal, and EDD is (naturally) concerned with visual perception. SAM is also perception-oriented as it takes its input from both ID and EDD, and according to Baron-Cohen is multi-modal with strong inclination towards the visual modality. ToMM, according to Baron-Cohen, is mainly concerned with creating M-representations (Leslie, 1994) given input from ID and EDD, mediated through SAM. About ToMM's preferred perceptual modality Baron-Cohen is not so clear, but it seems safe to assume that this is similar as for SAM. In comparison, the model of Nichols & Stich has mechanisms similar to ID and EDD, namely the DDM and PDM, and it is stated by Nichols & Stich (2003, p. 78), in regard to Baron-Cohen's model, that if something similar to ID (which they refer to as an 'agency detector') exists, it will fit comfortably with their theory. A counterpart of SAM in form of a distinct mechanism is not so obviously found in the model of Nichols & Stich (2003), as the PDM is concerned with a more low-level task — namely detecting the other's perception, as opposed to detecting shared perception — and the mechanism(s) for discrepant belief attribution are concerned with a more high-level task — namely belief attribution, which in Baron-Cohen's is left for ToMM whereas SAM creates intermediate-level representations. In any case, the account of Nichols & Stich does not seem to preclude the existence of a specialized SAM but possibly omits its mention because it focuses principally on the *cognitive* aspect of mindreading, providing a sketch of how this faculty comes into play in light of a general (human) cognitive architecture. Comparing the models, it then seems that the bulk of the model of Nichols & Stich (2003) is an explication of the ToMM as posited by Baron-Cohen (1995), whereas Baron-Cohen stresses the importance of perception and shared attention, something which is not so apparent in the account given by Nichols & Stich.

1.2.4 Remarks

The above presentation of the mindreading models of Baron-Cohen and Nichols & Stich intends to provide a frame of reference for the remainder of this chapter, and does not aim to provide an authoritative comparison of those models. Its main purpose is to provide some background to the statements put forward in later sections, and to prepare for our logical approach to formalizing (aspects of) mindreading.

1.3 Insights for Formal Approaches to Mindreading

In this section the psychological models of human mindreading that were discussed in Section 1.2 are analyzed, with the underlying motive to draw lessons for formal (i.e. mathematical, logical, computational) approaches to this topic

1.3.1 Perception

It is evident from the account of mindreading given by Baron-Cohen (1995) that perception plays a crucial role in this activity. This shows through in the presentation of his model of human mindreading discussed in Section 1.2.1 which purports that three (from a total of four) components of the mindreading system are predominantly occupied with the extraction of clues pertaining to the 'contents' of the other's mind from (mostly visual) perceptual information. Perception is also a major theme in Baron-Cohen's book on deficits in mindreading, appropriately named *Mindblindness* (1995), of which an entire chapter is devoted to the 'language of the eyes' and the importance of the visual modality in mindreading.

IV.1

Approaches to mindreading in the domain of artificial intelligence should thus also take perception into account. For embodied systems such as sociable robots (Feil-Seifer & Matarić, 2005; Dautenhahn & Werry, 2004; Breazeal, 2002) this means that those must be equipped with facilities for perception and desire detection to provide them with the capability to detect the objects of others' attention. In this sense it may be necessary to, for example, realize mechanisms for detection of the direction of others' eyes. If the mindreader is virtual and situated in a virtual environment then the requirements for performing mindreading are still similar, although it may be easier to establish the locus of attention of co-inhabitants of its virtual environment. In logical approaches like ours the matter of (shared) perception has influence of a different nature, imposing requirements for the formalism to allow expressing the (presumed) outcome of such perception. This is the case, considering that in mindreading the outcome of processes for detecting others perception/intentionality are beliefs/goals attributed to others. Later sections of this chapter (2.2.3, in particular) focus on this matter in more detail.

1.3.2 Agent Programming

The model by Nichols & Stich (2003), as recapitulated in Section 1.2.2 of this chapter, focuses principally on the capacity for mindreading in context of the cognitive architecture of a decision-making agent. It is noteworthy in this regard that this architecture — at least

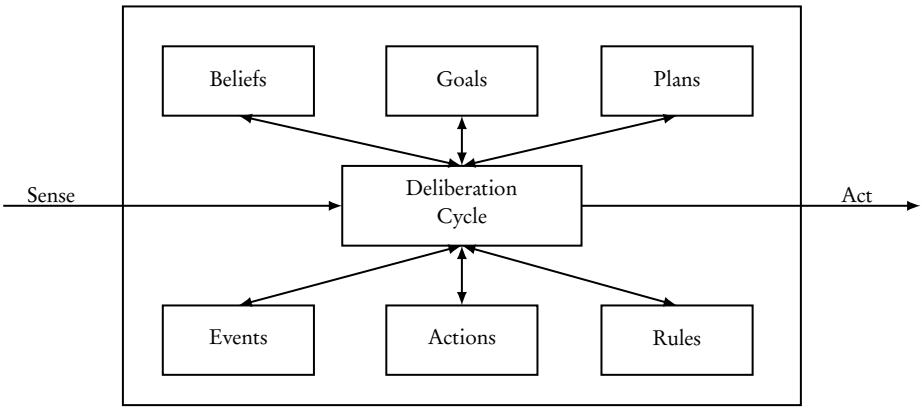


Figure IV.3: (reproduced from (Dastani, 2009))

the part of it that is not mindreading-specific — shows similarities with software agent architectures, like that of individual 2APL agents depicted in Figure IV.3. Those similarities between the basic cognitive architecture presumed by Nichols & Stich (2003) and the cognitive architecture of software agents extend beyond the surface level. A first parallel is found in the (presumed) representational capabilities of agents. Nichols & Stich take beliefs and desires as the basic representational concepts of human agents; similarly, in most agent programming approaches beliefs and goals — regarded in psychological literature as a particular sort of desires (Nichols & Stich, 2003) — are the basic representational concepts of software agents. A second parallel can be drawn in regard to agents' deliberation processes. The deliberation process of software agents is typically realized by means of a *sense-reason-act* cycle (Dastani, 2008, 2009), as indicated by Figure IV.3, in which agents take in perceptual information and process this information in order to decide their course of action. The model of Nichols & Stich (2003) features counterparts of the three phases of this cycle; namely the PDM and DDM which deal with (mindreading-specific) *sensing*, the action control systems for *acting*, and the remainder of the mechanisms which can all said to be geared towards different aspects of (partly mindreading-specific) *reasoning*.

That the aforementioned parallels exist between architectures of human and software agents is no coincidence, considering the fact that the influential philosophical Belief-Desire-Intention (BDI) account of human practical reasoning and (rational) agency by Bratman (1987) has been the object of attempts at formalization (Cohen & Levesque, 1990; Rao & Georgeff, 1991) that have laid the foundation for agent-based software architectures (Rao & Georgeff, 1995) and agent programming languages/frameworks (Hindriks, 2001; Pokahr et al., 2005; Dastani, 2008; Bordini et al., 2009). These parallels are of interest because they suggest the type of extensions required to endow BDI-based software agents with mindreading capabilities. Aside from this general insight, it is furthermore interesting to note that recent developments concerning *modularity* in agent programming (Dastani, 2009) can be fruitful with regard to realization of mindreading in A.I., which is mentioned in more detail in context of the example in Section 3.3.

1.3.3 M-representations

A central notion in many accounts of mindreading is that of *M-representations*.[1] As stated in Section 1.2.1, M-representations are constructs of the form [Agent-Attitude-"Proposition"]. Nichols & Stich, in comparing their account of pretence with that given by Leslie, observe that M-representations essentially are a notational device for *marking* 'regular' propositions and denoting that particular agents have particular propositional attitudes towards those marked propositions. In their account of mindreading, Nichols & Stich maintain that propositions serving in pretence or attribution do not differ structurally from regular propositions, but at the same time they do not see this point as a fundamental disagreement with Leslie's account that employs M-representations. Quoting Nichols & Stich (2003, p. 49): "[...] *it should be clear that the part of Leslie's theory that we have sketched so far is very similar to part of the theory that we have been defending. For, as we have noted* [earlier], *to claim that a class of representations is specially marked, and that the marking has important consequences for how the representations are treated by the cognitive system, is another way of saying that marked representations and unmarked representations are functionally different.* [...] *The representations in the Possible World Box (our theory), or within the quotation marks (in Leslie's theory) are tokens of the same types as the representations in the Belief Box (to use our preferred jargon) or in the pretender's primary representations (to use Leslie's)*". Thus, Nichols & Stich (2003) equate Leslie's marking of propositions by means of M-representations with the manner in which those propositions are treated functionally in their own account.

IV.1

It is noteworthy to recall at this point that Nichols & Stich (2003) consider *belief* and *desire* to be the two basic representational primitives which mindreaders themselves employ, and furthermore attribute to others. Thus, in terms of M-representations, Nichols & Stich essentially maintain that mindreaders employ those of the kind [i-Believes-p] and [i-Desires-p], where i is an agent and p a proposition — the interpretation of which in regard to the PWB would mean that the mindreader has the appropriate representational counterparts stored therein (i.e. stating [i-Believes-"p"] is tantamount to stating that there is a representation denoting that 'i believes p' in the PWB). In this regard it is interesting to reconsider the syntax of the language $\mathcal{L}_{\mathbf{M}}$ presented in Chapter III, and note that it provides operators that allow for expressing the basic kinds of M-representations which Leslie proposes and which both Baron-Cohen and Nichols & Stich adopt, namely the fact that some agent has a belief (by means of the **B**-operator) and the fact that some agent has a particular sort of desire, namely a goal (by means of the **G**-operator). In fact, expressions that are under the scope of the belief/goal operators can be considered as exactly the kind of "specially marked" propositions that Nichols & Stich talk about, rendering the language $\mathcal{L}_{\mathbf{M}}$ in that respect suitable for expressing mindreading based on an agent's (observed) behavior in terms of first-order M-representations. Precisely this is the subject of Section 2. Furthermore, the functional difference between marked and unmarked propositions has its reflection in the realm of agent programming, which is discussed in Section 3.3.

[1] The notion of M-representations was introduced by Leslie (1987), who originally used the term 'metarepresentation' to refer to the fact that children develop a capacity for representation which extends their capacity for 'primary' representation. However, because dispute arose on whether this term captured the intended meaning, Leslie later abandoned it in favor of the more neutral term 'M-representation' (Nichols & Stich, 2003).

1.4 Reflection

In Sections 1.1–1.3 of this chapter, accounts of human mindreading have been discussed that, in some way or other, form the basis for subsequent sections. It is noteworthy in this respect that our focus in this chapter is on providing a formal account of bottom-up mindreading; the term 'bottom-up' referring to attribution of mental states on grounds of what an observed agent "does", whereas 'top-down' refers to attribution of mental state on grounds of its "circumstances and biological needs" (Nichols & Stich, 2003, p. 144). Underlying motive is to facilitate the realization of mindreading in artificial systems. However, it should be understood that our objective is not to formally capture the intricacies of mindreading as it occurs in the human case. First of all, this is prohibited by the fact that our approach relies on the language $\mathscr{L}_{\mathbf{M}}$ used in the previous section, which is too simplistic to provide an accurate description of human-like mindreading. After all, it only has operators for the ascription of beliefs and goals, whereas predominant psychological accounts of mindreading distinguish a much broader variety of propositional attitudes, as well as emotions. If this were not a prohibiting factor, the fact would be that the logic used in this dissertation only allows for expressing so-called 'first-order beliefs': i.e. the belief of the mindreader that the observed agent has some (non-mentalistic) fact as its belief or goal. Humans typically involve higher-order nesting of propositional attitudes in their mindreading; conjecturing, for example, that person A has the belief that person B wants A to be happy. Leaving the exact magnitude of such 'nesting' that humans are capable of in the middle, it is generally agreed upon that healthy humans typically perform mindreading of an order higher than the first. Thus, a formal account aiming to capture human-like mindreading should allow for higher-order mindreading than is done here, and should also take into account the fact that mindreading takes effort and thus resources. Clearly, our account does no such thing, and thus should not be regarded as a formalization of human mindreading. Moreover, our formalism takes action occurrence in the form of unambiguous events to be the basis for mindreading, whereas in real-world domains a more refined model of action will be required. This also holds in relation to the topic of perception, which in our account is limited to attributing mental states to the agent on grounds of perception-related facts the mindreader takes to be the case, as will be seen later in this chapter.

But, one might ask, what *is* then the contribution of the mindreading account given in this chapter? It should be noted that it is principally an account of how an artificial entity in a virtual environment can perform first-order mindreading of beliefs and goals. Thus, it may contribute to the realization of software which can be said to be to some degree 'aware' of the mental states of the co-inhabitants of its environment. The fact that only first-order mindreading is considered need not be prohibitive in this respect; psychological literature takes the distinction between healthy and pathological cases of human mindreading to reside in the fact that in pathological cases no mindreading occurs whatsoever, whereas the healthy case concerns the application of some order of mindreading. Put bluntly: having some clue (if only of the first order) about others' mental states is better than having no clue at all. Also, the false-belief task example discussed in Section 3 of this chapter, based on a well-known scenario in psychological literature (Wimmer & Perner, 1983), constitutes a test that indicates quite advanced mindreading proficiency (Howlin et al., 1999).

Additionally, our approach to formalizing bottom-up mindreading of the first order can provide an inroad to tackling formalization of higher-order mindreading, if only in the sense that our framework provides a basis for extension in that direction. It likewise holds for the manner in which action observation is formalized here, which focuses on virtual environments but may also provide a basis for approaching this issue in less permissive domains. Finally, it is noteworthy that a single formalism is used to express mentalistic explanation of BDI-based agents' behavior in Chapter III, and more general mentalistic reasoning about observed behavior in the current chapter. This emphasizes the fact that explanation of observed behavior occurs in terms of the mindreader's ontology, which here employs the same "logical form" (cf. Nichols & Stich (2003)) with regard to software agents and others.

2 A Logical Account of Mindreading

In the previous section it was established that current psychological theories of mindreading agree that this is an activity comprising attribution of beliefs and desires to others, in which perception plays a significant role. In the current section our intention is to provide a (basic) logical account of this activity, by means of the formalism employed in Chapter III. In line with the accounts of the development of mindreading given by Baron-Cohen (1995) and Nichols & Stich (2003), that start with the attribution of desires as this is considered the most primitive part of mindreading from an evolutionary point of view, this section begins with remarking on the attribution of *goals* to agents. Next, the attribution of beliefs is discussed.

At this point it should be noted that our approach to formalizing mindreading focuses on identifying a variety of something which is here termed a *mindreading pattern*, abbreviated to 'MP' (plural 'MPs') for convenience. Such patterns state regularities that are presumed by the mindreader to hold between facts that do not (necessarily) pertain to the agent's mental state, and facts that do. It should be noted that those regularities are presented as logical expressions, which should be considered as *schemata* (templates) that are to be further instantiated and interpreted. The reason for this approach is that instantiation of such patterns is, typically, context-dependent; in Section 3 such a context is considered. This may seem somewhat odd, but should not be given that our concern is with the logical form ('pattern') of mindreading regularities. Also, note that formalization of those patterns in PDL allows translation to other domains, given the nature of PDL as a well-studied and common formalism for reasoning about action (Zhang & Foo, 2005; Meyer, 2000).

It can be argued that mindreading patterns, because they represent presumptions of the mindreader about agents' mental states that might be false, should be treated in a context of *nonmonotonic* reasoning (Brewka et al., 2008). This is a legitimate argument, but the topic of nonmonotonicity is not the principal focus here as it is non-trivial to relate in context of our PDL-based formalism (but see Chapters II and III for ideas). Also, it should be noted that defeasibility can be captured in different ways (Meyer & van der Hoek, 1991); using PDL, for example, it can be expressed that something holds *possibly* before or after some course of action. Moreover, we feel that expressing mindreading patterns is useful, if

only to capture their form, observing that this can be a basis for incorporating them into approaches based on (nonmonotonic) reasoning.

For stating the mindreading patterns the language $\mathscr{L}_{\mathbf{M}}$ from Chapter III is used, with slight modifications. Those modifications are principally syntactical and are employed in order to capture the fact that mindreading is a process which often concerns multiple agents. Specifically, it is the case that the operators **B** and **G** for belief and goal ascription, respectively, are extended with an agent identifier, e.g. i, to yield the operators $\mathbf{B}_i$ and $\mathbf{G}_i$. This expresses the fact that "agent i believes..." or "agent i has the goal...". Likewise, the dynamic modalities and propositions are extended with identifiers, so that $[\alpha]_i$ expresses "after action α by agent i it necessarily holds that...", $\mathbf{Done}_i(\alpha)$ expresses that "action α was done by agent i" and, similarly, $\mathbf{Obs}_i(\alpha)$ expresses that "it was observed that action α was done by agent i". The aforementioned syntactic modifications require according modification to the semantics as well. Those are not given in the same level of detail as in Chapter III, but given the fact that the modifications proposed here are not fundamental, this should not be an issue. The semantic modifications come down to the fact that accessibility relations are stated per action and per agent, so that given Ag as the set of agents the frame for the extended $\mathscr{L}_{\mathbf{M}}$ is $\mathfrak{F} = (\mathrm{W}, \{R_{\alpha.i} \mid \alpha \in \mathsf{Act} \;\&\; i \in \mathsf{Ag}\})$, where the subscript $\alpha.i$ expresses that i is the agent of action α. Observe that this modification straightforwardly subsumes the single-agent case of the previous chapter. The valuation functions are extended accordingly, being here presented as a tuple of the function interpreting the mindreader's own 'beliefs' and a set of functions related to interpretation of the mental states and actions of others. Models for $\mathscr{L}_{\mathbf{M}}$ are then tuples $\mathfrak{M} = (\mathfrak{F}, \vartheta)$, where $\vartheta = (\vartheta_{\mathsf{I}}, \{\vartheta_{\mathsf{B}.i}, \vartheta_{\mathsf{G}.i}, \vartheta_{\mathsf{O}.i}, \varrho_i \mid i \in \mathsf{Ag}\})$ and it is assumed that agent-specific expressions (e.g. $\mathbf{B}_i(p)$) are interpreted by the appropriate function (i.e. $\vartheta_{\mathsf{B}.i}$). Last but not least, it is important to mention that the property of strict sequentiality of actions (cf. Definition III.6) is maintained with extension to the actions of distinct agents, as follows.

Definition IV.1 (strict sequentiality of actions)**.** *It holds for any model* $\mathfrak{M}$ *and state* $w \in \mathrm{W}_{\mathfrak{M}}$ *for the language* $\mathscr{L}_{\mathbf{M}}$ *that*

$$\forall \alpha, \alpha' \in \mathsf{Act} \forall i, j \in \mathsf{Ag}: \\ \mathfrak{M}, w \models \mathbf{Done}_i(\alpha) \wedge \mathbf{Done}_j(\alpha') \quad \Longrightarrow \quad (\alpha = \alpha') \;\; \& \;\; (i = j)$$

The extensions to $\mathscr{L}_{\mathbf{M}}$ are thus rather harmless and merely generalize it to the multi-agent case, without fundamentally influencing the observer's 'world view'. Extensions of the formalism to concurrent action (of a single agent) and/or joint action (of multiple agents) would, in contrast, present a radical change, but those are out of the scope of the present account.

It should be noted that a superscript 'τ' is used to indicate translation of propositional expressions from $\mathscr{L}_{\mathbf{M}}$ to $\mathscr{L}_0$ by means of the (partial) function $\tau_{\mathbf{M}} : \mathscr{L}_{\mathbf{M}} \longrightarrow \mathscr{L}_0$ which retains logical connectives and replaces standard negation with $\mathscr{L}_0$-negation (i.e. $\tau_{\mathbf{M}}(\neg p) = {\sim}p$), so that for any propositional $\phi \in \mathscr{L}_{\mathbf{M}}$ holds that $\mathbf{B}(\phi^{\tau})$ and $\mathbf{G}(\phi^{\tau})$ are to be interpreted as $\mathbf{B}(\tau_{\mathbf{M}}(\phi))$ and $\mathbf{G}(\tau_{\mathbf{M}}(\phi))$, respectively. This notation is identical to that which was used in Section 3.2.2 of Chapter III to denote translation by means of the

function $\tau_0 : \mathscr{L}_0 \longrightarrow \mathscr{L}_{\mathbf{M}}$ and thus might seem confusing. It need not be, though, considering that τ_0-translation occurs outside the scope of the belief/goal ascription operators whereas $\tau_{\mathbf{M}}$-translation occurs inside this scope. For this reason the identical notations are favored for the sake of minimizing notational clutter.

2.1 Reading Goals

The models of mindreading purported by Baron-Cohen (1995) and Nichols & Stich (2003) feature mechanisms that are geared towards the detection of agents' desires: referred to as the ID (Intentionality Detector, Figure IV.1) and the DDM (Desire Detection Mechanisms, Figure IV.2). Both theories concur that those mechanisms are innate and ancient from an evolutionary point of view, citing evidence which suggests that some animals appear to employ desire detection but not other aspects of mindreading, and that infants employ desire detection at a very early age and before they employ other kinds of mindreading. The ID, as described by Baron-Cohen, is a very basic kind of mechanism, whose only task it is to detect that something is an agent (which is why Nichols & Stich refer to it as an 'agency detector'). The DDM of Nichols & Stich provide this primitive functionality of ID, amongst others, as well as the functionality to detect desire in ways which are more advanced (from an evolutionary point of view). For this reason, the following discussion of desire detection focuses principally on the functions of the DDM, as described by Nichols & Stich (2003).

IV.2

In describing the operation of the DDM, Nichols & Stich describe those as "*a cluster of mechanisms for attributing desires to others,* [that use] *a variety of cues for determining a target's goals*". Generally, those cues can be grouped into *behavioral cues* and *factual cues*, which are both the focus of subsequent sections. It is noteworthy in this regard that most cues regarding the target's goals are facts which are mediated through perception, and that the terms 'behavioral' and 'factual' in the present context point to the fact that the relevant cues are formalized as facts which (respectively) are, or are not, strictly related to the target's actions.

2.1.1 Goal Attribution on Grounds of Agents' Actions

The primary category of behavioral cues for determining a target's goals are the actions of the target itself. Nichols & Stich consider the mindreader's strategies for attribution of desires based on behavior as varying in sophistication, ranging from direct relations between action and attributed goals, to theories that serve the same purpose but take a richer variety of behavioral (and, possibly, contextual) cues into consideration (2003, p. 78–80). From the account of Nichols & Stich — and likewise from the account of Baron-Cohen, for that matter — it does not become apparent whether the mindreader, upon witnessing behavior that warrants goal attribution, attributes a goal to the agent in the state in which the behavior is observed, or before. Formalizing this sort of attribution forces one to be explicit about such matters, though, and in this respect it seems reasonable to presume to mindreader to have had the attributed goal *before* it exhibited the behavior warranting goal attribution; after all, it was this goal which supposedly initiated the agent's behavior. Furthermore, the

formal approach allows for taking actions which the agent has actually been *observed* to do as grounds for goal attribution, as well as actions which the agent is *presumed* to have done. In formalizing action-based goal attribution the latter, more general, option is chosen here, in line with the fact that it is the agent's actions in themselves that warrant attribution, irrespective of whether they have been observed or not.

Mindreading Pattern IV.1 (action-based goal attribution)**.** *Let $\alpha_1;\ldots;\alpha_n \in \mathscr{L}_\Delta$ be a sequence of (primitive) actions that warrants attribution of the goal $\psi \in \mathscr{L}_0$ to the agent $i \in \mathsf{Ag}$, under the condition that $\phi \in \mathscr{L}_\mathbf{M}$ is satisfied.*

$$(\mathbf{Done}_i(\alpha_1;\ldots;\alpha_n) \wedge \phi) \rightarrow \langle \alpha_1;\ldots;\alpha_n{}^{-}\rangle_i \mathbf{G}_i(\psi)$$

Observe also that MP IV.1 takes *sequences* of actions as grounds for goal attribution, generalizing the case in which a single action serves this purpose. This pattern also makes explicit, by means of the existential modality $\langle\cdot\rangle$, that the mindreader considers the agent to *possibly* have had the attributed goal ψ on grounds of the sequence it is presumed to have done; alternatively, the universal modality can be employed in order to formalize the fact that the mindreader considers the agent to necessarily have had that attributed goal. Furthermore, note that a contextual condition ϕ 'guards' the attribution of the sequence-related goal, formalizing the fact that the sequence of actions in some, but not all, cases warrants attribution of that goal. To see that this pattern applies to cases mentioned in the psychological literature, consider the following example, which is an elaboration of an example given by Baron-Cohen to illustrate his point that approach and avoidance are the two basic types of movement that warrant goal attribution (1995, p. 32–33).[2]

Example IV.1. *Suppose that a piece of cheese is located in some location, and that the action* 'enter' *which puts the agent* 'mickey' *in this location justifies attribution of the goal to have cheese, based on the agent approaching the cheese. This can then be expressed as follows:*

$$\mathbf{Done}_{\mathsf{mickey}}(\mathsf{enter}) \rightarrow \langle \mathsf{enter}^{-}\rangle_{\mathsf{mickey}}(\mathbf{G}_{\mathsf{mickey}}(\mathsf{have\_cheese}))$$

In regard to the above example, note that attribution of the goal to have cheese occurs on grounds of a single action and is not modulated by any contextual condition (or, equivalently, is modulated by the trivial condition $\top$). Although this comprises a very simple case of behavior-based goal attribution, Nichols & Stich (2003) maintain that the pattern of MP IV.1 constitutes the essential principle of this type of attribution, and mindreaders' skills in attribution of particular goals differs only from the basic pattern in the sense that *"strategies used to infer goals from behavior* [...] *are enriched and supplemented as children (and adults) learn more about the behavioural cues that are associated with various sorts of goals"*. In our formal approach this would be reflected by a refinement of the reasoning employed by the mindreader; i.e. more refined mindreading patterns. In fact, if one is looking for more refined patterns, the approach to attribution of mental states to an agent on

[2] Nichols & Stich (2003) mention a example of action-based goal attribution, but theirs comprises the joint action of two agents which is not expressible in our formalism, so that this single-agent example of Baron-Cohen is preferred.

grounds of its known plans, put forward in Chapter III, can be seen as a case of mindreading! There, the mindreader (observer) maintains very specific 'mindreading patterns' in regard to the rule preconditions (which, amongst others, comprise goals) that it can ascribe to an agent on grounds of its observed behavior. Those patterns derive from knowledge of the agent's rules (which couple those conditions to plans), showing how a notion of the way in which an agent's goals and beliefs serve to *generate* behavior can be used to formulate a notion of how to *explain* its observed behavior.

Generalizing the idea of the previous paragraph to some extent, it may be claimed that *any* theory of how an agent's goals and beliefs influence the selection of its behavior can be reformulated into a theory of how its observed behavior should be explained in terms of the goals and beliefs it possibly had. It is not our aim to give formal support here for this statement, which is perhaps somewhat bold, but it is our conviction that anyone looking to do so should find guidelines in terms of the material presented in Chapter III. In that chapter, significant attention was given to this topic, restricted to a 'theory' of agents' behavior in terms of their own behavior-producing rules. Focus in the next section is on goal attribution based on actions of other agents than the agent which is the object of mindreading.

2.1.2 Goal Attribution on Grounds of Third-Party Actions

At first glance it may seem strange to consider the attribution of goals to an agent based on the actions of *another* (i.e. third-party) agent. It need not always be, though, as in the case of *communicative* actions. Nichols & Stich cite verbal communication as a source of information about others' goals which is initially used undiscriminatingly by children to attribute goals to others if they claim to have those goals; a strategy which is gradually refined as children learn that not everything they are told is true (2003, p. 79). This mindreading pattern can be expressed as follows.

IV.2

Mindreading Pattern IV.2 (communication-based goal attribution)**.** *Let* $\alpha_1;\ldots;\alpha_n \in \mathscr{L}_\Delta$ *be a sequence of primitive actions that comprises communicative actions and that, if performed by some agent* $j \in \mathsf{Ag}$*, warrants the attribution of the goal* $\psi \in \mathscr{L}_0$ *to agent* $i \in \mathsf{Ag}$*, under the (propositional) condition* $\phi \in \mathscr{L}_\mathbf{M}$*.*

$$(\mathbf{Done}_j(\alpha_1;\ldots;\alpha_n) \wedge \phi) \rightarrow \langle \alpha_1;\ldots;{\alpha_n}^- \rangle_j \mathbf{G}_i(\psi)$$

Several aspects of MP IV.2 are noteworthy. First of all, it formalizes attribution of some goal $\psi \in \mathscr{L}_0$ to agent i in the state preceding the (communicative) actions of agent j. In MP IV.1 this occurred similarly with respect to attribution of a goal to the agent whose own action is presumed to have been 'done', and the underlying rationale is likewise; the motivation for agent j to perform the sequences comprising communicative actions (i.e. speech acts) was presumably based on the fact that i at that point had the goal ψ, so that goal attribution in the state preceding the sequence of actions is justified. This can be seen in an example, as follows.

Example IV.2. *Let* α *represent the communicative action (speech act) of stating* 'mickey' *to have the goal to have cheese. Goal attribution along MP IV.2 can then be expressed as follows,*

under the condition that 'minnie' *is considered reliable by the mindreader.*

$$(\mathbf{Done}_{\mathsf{minnie}}(\alpha) \wedge \mathsf{minnie\_trusted}) \rightarrow \langle \alpha^{-} \rangle_{\mathsf{minnie}}(\mathbf{G}_{\mathsf{mickey}}(\mathsf{have\_cheese}))$$

This pattern describes the sort of action-based mental state attribution described by Quaresma & Lopes (1995) and Dragoni et al. (2002), as discussed in Chapter VI on related work. In relation to MP IV.1, it is noteworthy that there the sequence of actions $\alpha_1;\ldots;\alpha_n$ may also comprise the action of the agent communicating *its own goal*. A last point to note is that the case of MP IV.2 might also warrant *second-order* attribution, seen in the fact that agent j communicating i's goal to the mindreader justifies the mindreader in attributing the communicated goal to i (as the pattern states), but moreover justifies attribution of the belief to j that the mindreader attributes this goal to i after its communicative action, as well as, perhaps, the attribution of the goal to j that the mindreader attributes this goal to i in the state preceding this communicative action. Given the fact that the language $\mathscr{L}_{\mathbf{M}}$ provides the machinery only to express first-order attribution, such matters are beyond its scope, though.

Another kind of actions of others that warrant attribution of goals to an agent is seen in the case where the agent is already attributed a desire, of which it is assumed that it is unable to pursue it until some action is performed that enables the agent to pursue its (latent) desire, adopting it as a goal. In that case the condition modulating attribution of the goal to i is some action performed by j in combination with the desire already attributed to i. The expressiveness of the language $\mathscr{L}_{\mathbf{M}}$ is a limiting factor in this case, however, as it does not provide a formal notion of 'desire'; in order to address such a mindreading pattern more expressiveness is needed. The KARO agent logic of van der Hoek et al. (1998) may serve as starting point in that respect, as it is also based on dynamic logic and provides the formal notion of a 'goal', formulated in terms of (amongst others) the primitives 'wish' (i.e. 'desire') and 'choice'. Moreover, because the operators of KARO have typical modal semantics in terms of accessibility relations, operators can be nested to express higher-order attribution.

2.1.3 Goal Attribution on Grounds of Facts

Apart from attributing goals to an agent on grounds of actions, certain facts of which the mindreader is convinced may lead it to attribute goals. This is seen clearly in the following example from the animal realm. Imagine a predatory animal (perhaps a domestic cat) that is hiding in the shade and is intently gazing at a possible prey (perhaps some small bird) sitting in the sun; unaware of being watched. It is not unreasonable in this scenario to attribute to the predator the goal of catching and eating the prey. Such examples are not restricted to the animal realm, though. An analogous example from the human realm could be that of your friend sitting on the couch and eyeing the unopened bag of chips on the table; depending on your knowledge of her preferences or typical behavior, you might attribute to her the goal of eating the chips. Regardless of the specific setting, the mindreading pattern of fact-based goal attribution takes the following form.

Mindreading Pattern IV.3 (fact-based goal attribution)**.** *Let $\phi \in \mathscr{L}_{\mathbf{M}}$ be some (propositional) condition that warrants the attribution of a goal $\psi \in \mathscr{L}_0$ to the agent $i \in \mathsf{Ag}$. The mindreading pattern of fact-based goal attribution is then as follows.*

$$\phi \rightarrow \mathbf{G}_i(\psi)$$

Observe that in the above pattern the goal $\psi \in \mathscr{L}_0$ is attributed to the agent i in the same state in which the fact $\phi \in \mathscr{L}_{\mathbf{M}}$ holds. In this sense MP IV.3 differs from MPs IV.1 and IV.2, as it does not require the occurrence of an action. Also, attribution concerns the same state as the attribution-warranting condition, whereas in the case of the action-based patterns attribution takes place in the state preceding this condition. Note, though, that the fact $\phi \in \mathscr{L}_{\mathbf{M}}$ *can* refer to actions, so that (for example) it can also be expressed that some action sequence warrants attribution of a goal in the state in which the mindreader considers it to have been done by the agent, instead of the preceding state. In fact, it could be argued that the predator in the example of the first paragraph of this section 'performs the action of doing nothing'. The point here is not that the occurrence of actions can be regarded as a fact, but that the formalism allows for making a distinction between attribution *before* some sequence of actions, and attribution 'now'; that some cases are expressible in either way is hereby less relevant.

It could be objected that MP IV.3 is too general to be useful, as it simply states that some relation may exist between a fact of which the mindreader is convinced and a goal it attributes to the agent, without giving any guidelines to the nature of those facts. Our counterargument is that it is definitely useful to make patterns like the above explicit, if only to see the difference between patterns that consist of conditions that warrant goal attribution 'now', and those that warrant goal attribution in preceding states. Recall that, after all, patterns are intended as templates, to make the form of certain generalities occurring in mindreading explicit. Those generalities are to be further instantiated for specific contexts, and in this regard it can be useful to have a mindreading pattern available as 'mold'. Examples of such generalities are given in relation to action specifications in the next section, which focuses on belief attribution but in which goal attribution is also briefly reflected upon again. Before proceeding, though, in conclusion of the current section an example is given of fact-based goal attribution.

Example IV.3. *In continuation of Example IV.1, assume that* 'mickey' *is situated in the room where the cheese is situated, and has both picked up and tasted the cheese. Also, assume that after performing the last action this agent shows an expression of disgust. The fact that the mindreader, given the aforementioned state of affairs, attributes the goal to* 'mickey' *to be rid of the cheese, can then be expressed as follows.*

$$(\mathbf{Done}_{\mathsf{mickey}}(\mathsf{taste\_cheese}) \wedge \mathsf{mickey\_looks\_disgusted}) \rightarrow \mathbf{G}_{\mathsf{mickey}}(\sim\mathsf{have\_cheese})$$

Observe that the example makes reference to the fact that an action has occurred which leads the mindreader to attribute a goal to the agent in the 'current' state, i.e. the state in which this action is considered to have been done. Also note that the example features an atomic proposition referring to an expression of emotion. Psychological evidence points

to facial expression as an important indicator of agents' desires and intentions (Nichols & Stich, 2003), which is utilized already by 18-month-old toddlers in goal attribution, and the atom 'mickey_looks_disgusted' can be thought of as the outcome of a perceptory process of detecting this agent's facial expression. The logical language $\mathscr{L}_{\mathbf{M}}$ has no primitives dedicated to formalizing emotion, and also employing logical atoms to refer to emotions in a 'true-false' fashion does no justice to their underlying complexity. The topic of emotions is thus considered out of scope for this dissertation, and the interested reader is referred to the work of Steunebrink (2010) for a head start into this direction, in the sense of a formalization of the OCC model of emotions in terms of the KARO framework by van der Hoek et al. (1998) mentioned before. Example IV.3 furthermore nicely illustrates the intended interpretations of the given patterns as heuristics. It could be the case, for example, that the mindreader mistakenly considers the agent to show an expression of disgust, or that it correctly interprets its expression but wrongly attributes the goal of not having the cheese because the agent had the goal to possess the cheese as part of a higher-level goal to bring the cheese to another agent, and was only sampling it to convince itself of the cheese's foul taste.

2.2 Reading Beliefs

This section focuses on mindreading patterns in which beliefs are attributed by the mindreader to the agent. As was the case in Section 2.1, a distinction is made between attribution on grounds of actions and attribution on grounds of facts.

2.2.1 Belief Attribution on Grounds of Agents' Actions

In this section, mindreading patterns of belief attribution on grounds of agents' actions are formalized. Similarly to goal attribution, as formalized in Section 2.1.1, it may be the case that some actions are considered by the mindreader to warrant the attribution of certain beliefs if a particular condition holds. Thus, the mindreading pattern of action-based belief attribution has a form similar to MP IV.1, as follows.

Mindreading Pattern IV.4 (action-based belief attribution)**.** *Let $\alpha_1;\ldots;\alpha_n \in \mathscr{L}_{\Delta}$ be a sequence of (primitive) actions that warrants attribution of the beliefs $\psi \in \mathscr{L}_0$ to the agent $i \in \mathsf{Ag}$, given satisfaction of the (propositional) condition $\phi \in \mathscr{L}_{\mathbf{M}}$.*

$$(\mathbf{Done}_i(\alpha_1;\ldots;\alpha_n) \wedge \phi) \rightarrow \langle \alpha_1;\ldots;\alpha_n{}^{-} \rangle_i \mathbf{B}_i(\psi)$$

In MP IV.4, as in MP IV.1, the agent i is attributed a belief ψ in the state that precedes the sequence of actions which (under some condition) warrants attribution. It can be remarked that, contrary to the case of goal attribution, it is not so obvious for belief attribution that it should pertain to the state preceding the attribution-warranting sequence, because, unlike goals, having beliefs as such does not 'drive' agents to exhibit behavior. In this respect it should be noted that, similarly to how fact-based goal attribution in Section 2.1.3 was stated to be applicable to facts pertaining to actions, and to thus warrant goal attribution in the 'current' state, this also holds for belief attribution. However, it can be perfectly natural

to attribute beliefs in states preceding action sequences of agents, as the following example illustrates in follow-up to Example IV.1.

Example IV.4. *Assume that a piece of cheese is located in some location, and that the action* 'enter' *brings* 'mickey' *to this location. If observing the agent to enter this room is sufficient grounds to ascribe to it the (possible) goal of having the cheese, as in Example IV.1, it seems reasonable to maintain the possibility that the agent had the belief that the cheese is, e.g., in the room on the table, in the state preceding the action, as follows.*

$$\mathbf{Done}_{\mathsf{mickey}}(\mathsf{enter}) \rightarrow \langle \mathsf{enter}^{-} \rangle_{\mathsf{mickey}}(\mathbf{B}_{\mathsf{mickey}}(\mathsf{table\_in\_room} \wedge \mathsf{cheese\_on\_table}))$$

Given the reference made to Example IV.1 in regard to the above example, note that if they are taken together then it is natural to consider the attribution of both the goal to have the cheese (as in Example IV.1) and the belief that the cheese is in the room on the table (as in Example IV.4). In fact, this sort of mindreading pattern is reminiscent of the attribution of beliefs and goals on grounds of rules, as in Chapter III, but this relation is not further explored here because to our opinion it adds little to the given treatment of reading goals and beliefs separately.

As remarked upon earlier (in relation to goals), formulating mindreading patterns is useful because it forces one to be explicit about the relations presupposed to hold between certain states of affairs which the mindreader considers to be the case (such as actions which have occurred and/or facts which hold true in some state), on the one hand, and mental states (beliefs and goals) which the mindreader attributes to agents, on the other. So far, though, only logical schemata/templates for mindreading have been identified (the 'mindreading patterns') that are to be interpreted as possible ways in which the mindreader can attribute mental states to agents. None of those schemata has been presented as logically valid; they merely capture the basic form of attribution of goals and beliefs to agents.

In the case of belief attribution it is sometimes possible, though, to identify mindreading patterns of which it can be argued that they are logically valid within a specific context. Before doing so, it seems instructive to point to the fact that the approach of belief and goal attribution in relation to observed behavior has two main 'flavors' in this dissertation. Firstly, in Chapters II and III, it is employed primarily to define a space of possibilities consisting of beliefs and goals entailed by an observed (software) agent's mental state. In this case it is *known* that the observed agent has a mental state of the attributed type, although it may not be known precisely which. Secondly, in the current chapter, this approach is employed principally as a heuristic for making sense of an observed (heterogeneous) agent's behavior, and it is *presumed* by the mindreader that the observed agent has a mental state of the attributed type. In either case, mental state attribution serves the purpose of providing the mindreader with a model of a mental state attributed to an agent that allows for reasoning about its behavior. It is noteworthy in this regard that the second flavor described above shows similarities with Dennett's account of the 'intentional strategy', which the author himself summarizes as follows in relation to beliefs (1987, p. 49, original emphasis): "*A system's beliefs are those it* ought to have, *given its perceptual capacities, its epistemic needs, and its biography. Thus, in general, its beliefs are both true and relevant to its life, and when false beliefs are attributed, special stories must be told to explain how the error resulted from the*

presence of features in the environment that are deceptive relative to the perceptual capacities of the system". In Section 2.2.3 the role of perception in belief attribution is formalized, but first focus is on the 'biography' of the agent in the sense of its (presumed) actions.

For identification of universally applicable mindreading patterns of belief attribution, it is here supposed that *action specifications* exist that describe (a subset of) the primitive actions that can be performed by the agent in terms of their preconditions and consequences (postconditions). It is noteworthy that those specifications reflect what could be called a 'theory of action semantics' that is employed *by the mindreader*, and as such may involve any expression of $\mathscr{L}_{\mathbf{M}}$ as precondition or postcondition. However, in what follows, focus is specifically on *ontic* (i.e. non-epistemic, cf. van Ditmarsch & Kooi (2008)) action specifications; specifications in terms of real-world facts, the truth of which can be established by both the mindreader and the agent through perception. That the restriction to ontic facts is relevant, follows from the fact that the language $\mathscr{L}_{\mathbf{M}}$ represents *the mindreader's beliefs*, so that it holds only for a subset of those beliefs that their subject is 'accessible' through perception for the agent as well. Of course, perception is by nature subjective, so that the beliefs of the mindreader and/or agent about an ontic fact could be false. Nevertheless, the fact that attribution of such facts as belief occurs on grounds of actions the agent presumably performed, justifies the mindreader's assumption of shared beliefs. The ontic specifications are as follows; note that they are uniformly interpreted as *deterministic action laws* (cf. Zhang & Foo (2005)) for technical simplicity.

Definition IV.2 (ontic action specifications)**.** *Let* Act *be the set of primitive actions,* $\mathsf{Atom}^{\mathsf{ont}} \subseteq \mathsf{Atom}$ *a set of ontic atoms, considered by the mindreader to be verifiable by both itself and any agent, and* $\mathsf{Lit}^{\mathsf{ont}} = \{p, \neg p \mid p \in \Phi\}$ *the corresponding set of ontic literals. Ontic action specifications then have the form* $(\mathsf{Pre}, \alpha, \mathsf{Post})$*, where* $\mathsf{Pre}, \mathsf{Post} \subseteq \mathsf{Lit}^{\mathsf{ont}}$ *and* $\alpha \in \mathsf{Act}$*. Given a set* $\mathsf{Spec}^{\mathsf{ont}}$ *of ontic action specifications, its interpretation is as follows.*

$$\forall i \in \mathsf{Ag} \forall (\mathsf{Pre}, \alpha, \mathsf{Post}) \in \mathsf{Spec}^{\mathsf{ont}} :$$
$$\models \bigwedge \mathsf{Pre} \rightarrow [\alpha]_i \bigwedge \mathsf{Post}$$

Ontic action specifications, as defined in Definition IV.2, thus represent the mindreader's view on how actions (of agents) affect the ontic state of the environment. Those specifications can be employed in reasoning about agents' behavior, which can be informally characterized — taking inspiration from the words of Dennett (1987), quoted earlier — as the attribution of those beliefs to the agent which it 'ought to have', according to the mindreader, in light of the actions it performed and those actions' ontic specifications. This is formally reflected in the following mindreading pattern.

Mindreading Pattern IV.5 (ontic attribution)**.** *Let* $(\mathsf{Pre}, \alpha, \mathsf{Post}) \in \mathsf{Spec}^{\mathsf{ont}}$ *be an ontic action specification of Definition IV.2, and let* ϕ^{τ} *denote the translation of* $\phi \in \mathscr{L}_{\mathbf{M}}$ *to* $\mathscr{L}_0$*, i.e.* $\phi^{\tau} = \tau_{\mathbf{M}}(\phi)$*. The mindreading pattern of ontic attribution is then as follows.*

$$(\mathbf{Done}_i(\alpha) \wedge [\alpha^-]_i \bigwedge \mathsf{Pre}) \rightarrow (\mathbf{B}_i((\bigwedge \mathsf{Post})^{\tau}) \wedge [\alpha^-]_i \mathbf{B}_i((\bigwedge \mathsf{Pre})^{\tau}))$$

The rationale behind MP IV.5 is that the mindreader attributes to the agent the same view on how the world is affected by its (the agent's) actions as that which it (the mindreader) itself has. Thus, if the mindreader is conclusively (i.e. with regard to each preceding state) convinced that the preconditions of the agent's action were satisfied, and therefore is convinced that the postconditions hold in the state in which the agent is considered 'done', then it conclusively attributes to the agent belief in those conditions as well. This is perhaps best seen in an example, as follows, continuing the 'cheesy' theme of preceding examples.

Example IV.5. *Consider the action* 'pickup_cheese' *that has the ontic action specification* $(\{\mathsf{cheese\_on\_table}\}, \mathsf{pickup\_cheese}, \{\neg\mathsf{cheese\_on\_table}\})$. *The instantiation of MP IV.5 with regard to this specification is then as follows for the agent* 'mickey'.

$$(\mathbf{Done}_{\mathsf{mickey}}(\mathsf{pickup\_cheese}) \wedge [\mathsf{pickup\_cheese}^{-}]_{\mathsf{mickey}}\mathsf{cheese\_on\_table}) \rightarrow$$
$$(\mathbf{B}_{\mathsf{mickey}}(\sim\mathsf{cheese\_on\_table}) \wedge [\mathsf{pickup\_cheese}^{-}]_{\mathsf{mickey}}\mathbf{B}_{\mathsf{mickey}}(\mathsf{cheese\_on\_table}))$$

Example IV.5 could be thought to be trivial because, informally speaking, if the cheese that was on the table was picked up by an agent then this agent should obviously have believed that it was on the table (otherwise it wouldn't have performed the action), and currently hold the belief that it isn't (considering the fact that its action succeeded). However, it is our conviction that this example (and, moreover, the mindreading pattern which it illustrates) is not trivial at all, as it provides a guideline for bottom-up construction of models of others' mental states (in this case beliefs) from first principles. Empirical evidence from human psychopathology indicates that individuals that fail to 'construct' and use models of others' mental states, especially in interacting with and reasoning about those others, have trouble functioning successfully in society (Baron-Cohen, 1995). Making regularities of mindreading explicit may help training, or otherwise helping, such individuals. Furthermore, it appears that the attribution of mental states to others sometimes is the only available heuristic for making sense of their behavior (Nichols & Stich, 2003; Dennett, 1987), so that we consider any guideline towards implementing this sort of heuristic to be useful from the point of view of artificial intelligence.

This section concludes with two remarks pertaining to MP IV.5 and the action specifications it relies on. First of all, it should be observed that the ontic action specifications considered thus far are 'agent-independent', i.e. the preconditions and postconditions of actions are taken to be interpretable in the sense of Definition IV.2 *regardless of the agent performing the action*. In general, this is true for some conditions but not for others: consider (in relation to the above example) the case that an agent is considered to possess the cheese (formalized as have_cheese in Example IV.1) after picking it up. Clearly, this postcondition is strictly relative to the agent performing this action! In other words, if agent i performs this action then the postcondition of the specification should entail that (ontic) proposition which represents i having the cheese. If such cases are to be reflected formally, then sets of action specifications should be stated per agent or be otherwise relativized with respect to agents, but for simplicity this step is here omitted. The second point to note pertains to the fact that similar to how MPs IV.1–IV.4 are given for sequences of actions, MP IV.5 can be extended to sequences as well. In order to illustrate this, the concept of a *chain of actions* is defined as follows.

Definition IV.3 (chain of actions)**.** *Let* $\mathsf{Spec}^{\mathsf{ont}}$ *be a set of ontic action specifications as defined in Definition IV.2, and let* $\alpha_1,\ldots,\alpha_n \in \mathsf{Act}$ *be primitive actions. The sequence of actions* $\alpha_1;\ldots;\alpha_n$ *is then a chain of actions with respect to this set of action specifications if*

$$\exists(\mathsf{Pre}_1,\alpha_1,\mathsf{Post}_1),\ldots,(\mathsf{Pre}_n,\alpha_n,\mathsf{Post}_n)\in\mathsf{Spec}^{\mathsf{ont}}\forall i\in\mathsf{Ag}:$$
$$(\models[\alpha_1]_i\bigwedge\mathsf{Pre}_2)\,\&\,\ldots\,\&\,(\models[\alpha_{n-1}]_i\bigwedge\mathsf{Pre}_n)$$

With regard to the mindreading pattern of ontic attribution, chains of actions warrant attribution throughout the chain, as the following proposition shows.

Proposition IV.1 (chain-based attribution)**.** *Let* $\mathsf{Spec}^{\mathsf{ont}}$ *be a set of ontic action specifications (Definition IV.2),* $\alpha_1;\ldots;\alpha_n\in\mathscr{L}_\Delta$ *a chain of actions with respect to* $\mathsf{Spec}^{\mathsf{ont}}$ *(Definition IV.3), and* $i\in\mathsf{Ag}$ *any agent. Let MP IV.5 be valid for* $\alpha_1,\ldots,\alpha_n\in\mathsf{Act}$ *with respect to* $\mathsf{Spec}^{\mathsf{ont}}$*, so that given* $(\mathsf{Pre}_1,\alpha_1,\mathsf{Post}_1),\ldots,(\mathsf{Pre}_n,\alpha_n,\mathsf{Post}_n)\in\mathsf{Spec}^{\mathsf{ont}}$ *it holds that*

$$\begin{aligned}\models(\mathbf{Done}_i(\alpha_1)\wedge[\alpha_1{}^-]_i(\textstyle\bigwedge\mathsf{Pre}_1))\rightarrow\\ ([\alpha_1{}^-]_i(\mathbf{B}_i((\textstyle\bigwedge\mathsf{Pre}_1)^\tau))\wedge\mathbf{B}_i((\textstyle\bigwedge\mathsf{Post}_1)^\tau\wedge(\textstyle\bigwedge\mathsf{Pre}_2)^\tau)\wedge[\alpha_2]_i(\mathbf{B}_i((\textstyle\bigwedge\mathsf{Post}_2)^\tau\\ \wedge(\textstyle\bigwedge\mathsf{Pre}_3)^\tau))\wedge\ldots\wedge[\alpha_2;\ldots;\alpha_n]_i(\mathbf{B}_i((\textstyle\bigwedge\mathsf{Post}_n)^\tau)))\end{aligned}$$

Proof. Follows straightforwardly from the interpretation of action specifications given in Definition IV.2, their chain property as defined in Definition IV.3, and the fact that MP IV.5 is valid for each of $\alpha_1,\ldots,\alpha_n\in\mathsf{Act}$. □

2.2.2 Belief Attribution on Grounds of Third-Party Actions

Similarly to the case of goal attribution, as discussed in Section 2.1.2, an important source of information about agents' beliefs are third-person reports (Nichols & Stich, 2003, p. 92); i.e. the communicative actions of (third-party) others. Thus, unsurprisingly perhaps, the mindreading pattern for communication-based belief attribution is not very different from that for communication-based goal attribution (MP IV.2).

Mindreading Pattern IV.6 (communication-based belief attribution)**.** *Let* $\alpha_1;\ldots;\alpha_n\in\mathscr{L}_\Delta$ *be a sequence of primitive actions that comprises communicative actions and that, if performed by some agent* $j\in\mathsf{Ag}$*, warrants attribution of the belief* $\psi\in\mathscr{L}_0$ *to agent* $i\in\mathsf{Ag}$*, under the condition* $\phi\in\mathscr{L}_{\mathbf{M}}$*.*

$$(\mathbf{Done}_j(\alpha_1;\ldots;\alpha_n)\wedge\phi)\rightarrow\langle\alpha_1;\ldots;\alpha_n{}^-\rangle_j\mathbf{B}_i(\psi)$$

MP IV.6 shows that, as in the case of communication-based goal attribution, belief attribution on grounds of communicative actions is taken to principally pertain to the state preceding the communicative action. Naturally, it is also the case here that beliefs can be attributed in the state in which the communicative action is considered to have been 'done'. Similarly as with goal attribution, though, this is taken to be subsumed by the

pattern of fact-based belief attribution, discussed in the next section. Apart from the aforementioned, other remarks made in Section 2.1.2 in regard to MP IV.2 also hold here in regard to MP IV.6. Specifically, it can be the case that an agent communicates its own belief, and that communication of beliefs warrants second-order attribution. Because the issues in those cases are similar as with goal attribution, they are not further discussed here. Lastly, also here an example is given.

Example IV.6. *Let α represent the communicative action (speech act) of stating* 'mickey' *to have the belief that the cheese is on the table. Belief attribution along MP IV.2 can then be expressed as follows, under the condition that* 'minnie' *is considered reliable by the mindreader.*

$$(\mathbf{Done}_{\mathsf{minnie}}(\alpha) \wedge \mathsf{minnie\_trusted}) \rightarrow \langle \alpha^{-} \rangle_{\mathsf{minnie}}(\mathbf{B}_{\mathsf{mickey}}(\mathsf{cheese\_on\_table}))$$

Belief attribution on grounds of the actions of third-party others can occur with respect to the ontic preconditions and postconditions of actions, similar to belief attribution on grounds of agents' own actions. A prerequisite in this case is that the mindreader is convinced of the fact that some agent j to which it attributes belief in the conditions associated with an agent i's action, is *aware* of the fact that i has performed this action. In terms of the language $\mathscr{L}_{\mathbf{M}}$ this is tantamount to stating that it is *believed* by j that i has done this action, and the mindreading pattern for ontic attribution on grounds of others' actions is thus as follows.

IV.2

Mindreading Pattern IV.7 (ontic attribution on grounds of third-party actions)**.** *Let, for some action α $(\mathsf{Pre}, \alpha, \mathsf{Post}) \in \mathsf{Spec}^{\mathsf{ont}}$ be its ontic specification, as defined in Definition IV.2, and let $i, j \in \mathsf{Ag}$ be agent identifiers. The mindreading pattern of ontic attribution based on actions of third-party others is then as follows.*

$$(\mathbf{Done}_i(\alpha) \wedge \mathbf{B}_j(\mathbf{Done}_i(\alpha)) \wedge [\alpha^{-}]_i \bigwedge \mathsf{Pre}) \rightarrow$$
$$(\mathbf{B}_j((\bigwedge \mathsf{Post})^{\tau}) \wedge [\alpha^{-}]_i \mathbf{B}_j((\bigwedge \mathsf{Pre})^{\tau}))$$

It is noteworthy that MP IV.7 differs from MP IV.5 in the fact that it involves distinct agents, as well as a proposition that denotes the fact that the agent to which action-based beliefs are attributed is aware of the other's action having occurred. Indeed, it seems reasonable to assume that agents are aware of the actions they have themselves performed — an assumption which is, in fact, central to MP IV.5 — as expressed in the following validity, stated here as mindreading pattern.

Mindreading Pattern IV.8 (action awareness)**.** *Let $i \in \mathsf{Ag}$ be any agent that is the object of mindreading, and $\alpha \in \mathsf{Act}$ any primitive action. The following mindreading pattern then reflects the presumption of agents' awareness of their own actions.*

$$[\alpha]_i \mathbf{B}_i(\mathbf{Done}_i(\alpha))$$

It is shown below that if MPs IV.7 and IV.8 are both applied, then MP IV.5 follows. Before doing so, however, it is pointed out that, whereas '$\mathbf{B}_i(\mathbf{Done}_i(\alpha))$' has meaning in our language, the expression '$\mathbf{B}_i(\mathbf{Obs}_i(\alpha))$' has not, because its mindreader-relative interpretation

(by means of the valuation function ϑ_{O}) makes for '$\mathbf{Obs}_i(\alpha)$' not being a regular atomic proposition of the set Atom, opposed to '$\mathbf{Done}_i(\alpha)$'.

Lemma IV.1. *Let MP IV.8 be valid for some agent $i \in \mathsf{Ag}$ and action $\alpha \in \mathsf{Act}$, so that $\models [\alpha]_i \mathbf{B}_i(\mathbf{Done}_i(\alpha))$. It then holds that if agent i is considered to have done action α then it is also considered to be aware of having done that action, i.e.*

$$\models (\mathbf{Done}_i(\alpha) \rightarrow \mathbf{B}_i(\mathbf{Done}_i(\alpha)))$$

Proof. Take any model $\mathfrak{M}$ for $\mathscr{L}_{\mathbf{M}}$ and any $w \in \mathrm{W}_{\mathfrak{M}}$ such that $w \models [\alpha]_i \mathbf{B}_i(\mathbf{Done}_i(\alpha))$ for some $i \in \mathsf{Ag}$ and $\alpha \in \mathsf{Act}$. Assume that $w \models \mathbf{Done}_i(\alpha)$. From Proposition III.13 it then follows that $w \models \langle \alpha^- \rangle_i(\top)$, so that $\exists (w', w) \in R_{\alpha.i}$. Because of $\models [\alpha]_i \mathbf{B}_i(\mathbf{Done}_i(\alpha))$ it holds that $\forall (w'', w''') \in R_{\alpha.i} : (w''' \models \mathbf{B}_i(\mathbf{Done}_i(\alpha)))$, so that also $w \models \mathbf{B}_i(\mathbf{Done}_i(\alpha)))$. Note that $\mathfrak{M}$ and $w \in \mathrm{W}_{\mathfrak{M}}$ are arbitrary, so that the claim is proven. □

Proposition IV.2. *Let MPs IV.7 and IV.8 be valid for any $i, j \in \mathsf{Ag}$ and some $\alpha \in \mathsf{Act}$ with specification $(\mathsf{Pre}, \alpha, \mathsf{Post}) \in \mathsf{Spec}^{\mathsf{ont}}$, so that $\models (\mathbf{Done}_i(\alpha) \wedge \mathbf{B}_j(\mathbf{Done}_i(\alpha)) \wedge [\alpha^-]_i \bigwedge \mathsf{Pre}) \rightarrow \mathbf{B}_j((\bigwedge \mathsf{Post})^\tau) \wedge [\alpha^-]_i \mathbf{B}_j((\bigwedge \mathsf{Pre})^\tau))$ and $\models [\alpha]_i \mathbf{B}_i(\mathbf{Done}_i(\alpha))$. It then holds that MP IV.5 is valid with respect to α, i.e.*

$$\models (\mathbf{Done}_i(\alpha) \wedge [\alpha^-]_i \bigwedge \mathsf{Pre}) \rightarrow (\mathbf{B}_i((\bigwedge \mathsf{Post})^\tau) \wedge [\alpha^-]_i \mathbf{B}_i((\bigwedge \mathsf{Pre})^\tau))$$

Proof. Let α be an action with respect to which MPs IV.7 and IV.8 are valid for any $i, j \in \mathsf{Ag}$. Observe that this validity also holds in case that $i = j$, so that $\models (\mathbf{Done}_i(\alpha) \wedge \mathbf{B}_i(\mathbf{Done}_i(\alpha)) \wedge [\alpha^-]_i \bigwedge \mathsf{Pre}) \rightarrow \mathbf{B}_i((\bigwedge \mathsf{Post})^\tau) \wedge [\alpha^-]_i \mathbf{B}_i((\bigwedge \mathsf{Pre})^\tau))$ holds, and from Lemma IV.1 follows $\models (\mathbf{Done}_i(\alpha) \wedge [\alpha^-]_i \bigwedge \mathsf{Pre}) \rightarrow \mathbf{B}_i((\bigwedge \mathsf{Post})^\tau) \wedge [\alpha^-]_i \mathbf{B}_i((\bigwedge \mathsf{Pre})^\tau))$. Note that this holds for any $i \in \mathsf{Ag}$, so the claim is proven. □

Even though MP IV.5 is essentially subsumed by MPs IV.7 and IV.8, as Proposition IV.2 shows, it is still considered worthwhile to state it explicitly because it succinctly expresses the case of ontic attribution based on an agent's own actions, and also because stating it brings to light the assumption which underlies this pattern, namely that agents are implicitly assumed to be aware of the actions that they have performed themselves. It should be noted that this assumption may not always be justified, so that bringing it to light is useful for consideration of cases in which it is, or isn't. As before, MP IV.7 is perhaps best illustrated with an example.

Example IV.7. *Consider the action* '$\mathsf{pickup\_cheese}$' *that has the ontic action specification* $(\{\mathsf{cheese\_on\_table}\}, \mathsf{pickup\_cheese}, \{\neg \mathsf{cheese\_on\_table}\})$. *The instantiation of MP IV.7 for agents* 'mickey' *and* 'minnie' *with regard to this specification then could be as follows.*

$$\begin{aligned}
&(\mathbf{Done}_{\mathsf{mickey}}(\mathsf{pickup_cheese}) \wedge \mathbf{B}_{\mathsf{minnie}}(\mathbf{Done}_{\mathsf{mickey}}(\mathsf{pickup_cheese})) \\
&\quad \wedge [\mathsf{pickup_cheese}^-]_{\mathsf{mickey}} \mathsf{cheese_on_table}) \rightarrow \\
&\quad (\mathbf{B}_{\mathsf{minnie}}(\sim \mathsf{cheese_on_table}) \wedge [\mathsf{pickup_cheese}^-]_{\mathsf{mickey}} \mathbf{B}_{\mathsf{minnie}}(\mathsf{cheese_on_table}))
\end{aligned}$$

This simple example concisely illustrates the case of attribution based on third-party actions, but in doing so also points out a limitation of the language $\mathscr{L}_{\mathbf{M}}$ in expressing such matters. After all, it need strictly speaking not be the case that Minnie had the belief that the cheese was on the table in the state preceding Mickey's action (although the mindreader's assumption that she did can be defended, given that she is ascribed the belief that Mickey performed the action). It is more accurate to say the Minnie has the belief that the cheese was on the table before Mickey performed the action, i.e. changing the order of the operators. This, however, is not allowed by the language employed here, suggesting the need for increased expressiveness if the intricacies of such cases are to be captured.

2.2.3 Belief Attribution on Grounds of Facts

The preceding sections pertaining to attribution of beliefs have focused on actions, both those of the agent to which beliefs are attributed (Section 2.2.1) and those of third-party others (Section 2.2.2). The current section focuses on belief attribution on grounds of facts, similar to how Section 2.1.3 did with respect to goal attribution. This offers some additional possibilities due to the fact that mutual beliefs as a result of *shared perception* can be modeled, as mentioned in Section 1. This will be shown later in this section, but first focus is on 'plain' fact-based belief attribution. Just as with the fact-based goal attribution pattern stated in MP IV.3, belief attribution can occur on grounds of specific facts. For example, the mindreader may presume that if it is itself convinced that the sky is cloudy then the agent believes the sky to be cloudy as well, or presume that it believes that it will rain the next day. The following mindreading pattern allows for expressing such kinds of belief attribution.

Mindreading Pattern IV.9 (fact-based belief attribution)**.** *Let $\phi \in \mathscr{L}_{\mathbf{M}}$ be some fact that warrants the attribution of a belief $\psi \in \mathscr{L}_0$ to the agent. The mindreading pattern of fact-based belief attribution is then as follows.*

$$\phi \rightarrow \mathbf{B}(\psi)$$

MP IV.9 expresses a relation between facts the mindreader itself takes to be true and beliefs which it attributes to the agent that is the object of its mindreading. The pattern can be seen as a schema which is to be further instantiated in order to serve in 'profiling' the agent; details for such instantiation are not dealt with here. It is noteworthy in this respect that Nichols & Stich claim that a large share of the beliefs attributed to others are actually the mindreader's own beliefs, a phenomenon they refer to as *default belief attribution* (2003, p. 87–92). The authors go as far as claiming that most of the beliefs mindreaders attribute to agents are done so by default, except for some that are explicitly overridden. Disputing or supporting this claim is outside the scope of this thesis, but in any case it shows that fact-based belief attribution may play a significant role in mindreading. The overrides mentioned by Nichols & Stich can stem from the particular circumstances agent and mindreader find themselves in, or from a profile the mindreader has of the agent. In the remainder of this section this topic is discussed with respect to a specific class of overrides, namely those that are based on facts which relate to the (presumed) perception of the agent that is the object of mindreading.

The relation between 'perceiving' and 'believing' is often encountered in psychological literature on mindreading (Leslie, 1994; Baron-Cohen, 1995; Nichols & Stich, 2003), and is also encountered in colloquial settings, for example in the English saying "*seeing is believing*".[3] It is also the title of a formal approach by van Linder et al. (1997), who employ KARO in specification of informative actions, corresponding to the various ways in which agents can acquire information. This saying reflects the fact that perception is generally considered to be a trustworthy source of belief, so that if an agent perceives something to be the case then it is justified in believing that it is the case. In the case of belief attribution the reliability of perception as a source of beliefs can be exploited by the mindreader, if it is convinced that the agent is in a position to perceive particular facts. The *presumed* perception of the agent is then taken to provide the agent with information regarding the veracity of those facts, warranting the mindreader to attribute belief to the agent based on the mindreader's own information about those facts. This is comparable to the pattern expressed by MP IV.7, where it is required that the mindreader must be convinced of the fact that the agent to which it attributes beliefs is *aware* that some other agent performed the action on grounds of which attribution occurs. Note that the fact that beliefs, attributed on grounds of agents' (presumed) perception, are in a sense more justifiable than goals attributed on those grounds, because in the latter case the attributed goals concern facts which are not (yet) true in the world and can thus not (yet) be verified by the mindreader — to see this, again consider the example given in the first paragraph of Section 2.1.3.

Both Baron-Cohen (1995) and Nichols & Stich (2003) provide numerous examples of perception-based mindreading, some of which involve the mindreader being convinced of the truth of particular facts on grounds of which it attributes certain beliefs to the object of its mindreading. An example given by Nichols & Stich pertaining to such cases of *positive belief attribution* goes as follows. Suppose there is a box, and that the mindreader has come to know there is a frog in the box. Furthermore, suppose that the mindreader perceives the object of its mindreading (i.e. the agent) to be looking inside the box. It is then reasonable by most accounts for the mindreader to attribute the belief to the agent that there is a frog in the box. This particular scenario can be abstracted from in order to give a formal account of positive belief attribution; specifically, it should be observed that this example involves a 'cue' or *indicator* for the agent's perception (namely the fact that it is looking inside the box), which the mindreader can itself verify and which warrants it to attribute particular (box-related) beliefs to the agent. In formalizing belief attribution on grounds of the presumed perception of the agent it is therefore assumed that a particular subset of atoms $\mathsf{Indic}_{\mathbf{B}} \subseteq \mathsf{Atom}$ serves as indicators for the agent's perception-based beliefs, and a function is defined that maps those particular atoms to the facts whose attribution they justify, which for the sake of simplicity are assumed to be literals.

Definition IV.4 (perception-based belief indication)**.** *Let* $\mathsf{Indic}_{\mathbf{B}} \subseteq \mathsf{Atom}$ *be a set of indicators for perception-based beliefs, and* $\mathsf{Lit} = \{p, \neg p \mid p \in \mathsf{Atom}\}$ *the literals of* $\mathscr{L}_{\mathbf{M}}$*. The perception-based belief indication function* $\mathsf{perc}_{\mathbf{B}} : \mathsf{Indic}_{\mathbf{B}} \longrightarrow (\mathsf{Ag} \times \wp(\mathsf{Lit}))$ *then maps indicators to a tuple of an agent and a set of literals of which they indicate perception-based belief.*

[3] The Dutch equivalent of this saying is "*eerst zien, dan geloven*".

The function defined in Definition IV.4 thus provides a source of information that the mindreader can employ in attribution of beliefs to the agent, similar to the set of ontic action specifications in Sections 2.2.1 and 2.2.2. In the case of action-based belief attribution the mindreader verifies whether it considers the preconditions of actions to hold conclusively (i.e. in all states preceding the action); in the case of perception-based belief attribution the mindreader employs its model of the facts about which the agent presumedly has gained information through its perception. Using the perception-based belief indication function defined in Definition IV.4 the mindreader can then attribute particular beliefs to the agent on grounds of its own beliefs (represented by the propositional subset of $\mathscr{L}_{\mathbf{M}}$), as follows.

Mindreading Pattern IV.10 (perception-based belief attribution)**.** *Let* $\mathsf{Indic}_{\mathbf{B}} \subseteq \mathsf{Atom}$ *be a set of indicators for perception-based beliefs,* $\mathsf{Lit} = \{p, \neg p \mid p \in \mathsf{Atom}\}$ *the literals of* $\mathscr{L}_{\mathbf{M}}$*, and* $\mathsf{perc}_{\mathbf{B}} : \mathsf{Indic}_{\mathbf{B}} \longrightarrow (\mathsf{Ag} \times \wp(\mathsf{Lit}))$ *the perception-based belief indication function given in Definition IV.4. Given* $\mathsf{perc}_{\mathbf{B}}(p) = (i, \Phi)$ *for* $p \in \mathsf{Indic}_{\mathbf{B}}$*, the mindreading pattern of perception-based belief attribution is as follows for each* $\phi \in \Phi$*.*

$$(p \wedge \phi) \rightarrow \mathbf{B}_i(\phi^{\tau})$$

Thus, MP IV.10 formalizes a pattern of mindreading in which the mindreader conclusively attributes belief in certain facts to the agent which is the object of its mindreading, based on the indication that the agent presumedly has information about those facts on grounds of its perception. It should be stressed that this pattern is instantiated by specific indicators (atoms) that are related to specific sets of facts (literals) whose perceptibility they indicate, serving as grounds on which the mindreader attributes those facts as beliefs to agent. For example, in the example given earlier of the frog in the box, the fact that the agent is looking into the box — a fact which can be verified by the mindreader — is an indicator for supposing that the (defeasible) information the mindreader has about the contents of the box also constitutes part of the agent's beliefs (which it can only be presumed to have). The aforementioned is again illustrated with an example, focusing on the latter pattern of perception-based belief attribution.

Example IV.8. *The cheese, which in Examples IV.4–IV.7 was on the table, turns smelly and is placed on the table under a non-transparent 'cheese bell'*[4] *for health and safety reasons. Assume that the fact that agent* i *is looking under the bell, formalized by the atomic proposition* $\mathsf{looking\_under\_bell}_i$*', is an indicator for beliefs that the agent* i *according to the mindreader can obtain by looking under the bell. Formally, let* $\mathsf{perc}_{\mathbf{B}}(\mathsf{looking\_under\_bell}_i) = (i, \{\mathsf{cheese\_on\_table}, \neg\mathsf{cheese\_on\_table}\})$*, yielding the following two patterns.*

$$(\mathsf{looking\_under\_bell}_i \wedge \mathsf{cheese\_on\_table}) \rightarrow \mathbf{B}_i(\mathsf{cheese\_on\_table}) \qquad \&$$
$$(\mathsf{looking\_under\_bell}_i \wedge \neg\mathsf{cheese\_on\_table}) \rightarrow \mathbf{B}_i(\sim\mathsf{cheese\_on\_table})$$

It is noteworthy that indicators of the agent's perception may in some cases be intrinsically related to action. For example, the state in which a certain fact indicates perception of the

[4] Called a *kaasstolp* in Dutch, which is a storage box for cheese.

agent (e.g. the fact that it is looking under the cheese bell) might have been achieved on grounds of a 'perceptory action' (i.e. the action of looking under the bell). This implies no lack of generality, though, because in such a case the perception indicator relevant to the perceptory action (e.g. 'look_under_bell') is simply the atom stating its 'doneness' (i.e. '**Done**(look_under_bell)').

Also, observe that indicators of agents' perception can be more complex than the set of atoms $\mathsf{Indic}_{\mathbf{B}}$ assumed here, and that likewise the indicated beliefs may be more complex than single literals. Particularly, instead of taking literals, both indicators and indicated beliefs could be taken to expressions from the propositional subset of $\mathscr{L}_{\mathbf{M}}$; the main format of the pattern remains the same, though. Lastly, it should be noted that the set of action specifications defined in Definition IV.2 as well as the belief indication function defined in Definition IV.4 are technical 'aides' that could have been omitted, because the main focus is on the mindreading patterns as schemata. However, stating the action specifications and belief indication function as extra-logical formal entities is useful, because it makes for more straightforward generalization of this approach to other domains (e.g. other action formalisms and/or implementation) and also benefits clarity of presentation.

2.3 Reading Minds

Sections 2.1 and 2.2 have focused on the attribution of goals and beliefs, which constitutes the 'core business' of mindreading (Baron-Cohen, 1995; Nichols & Stich, 2003). In doing so, it was discussed how sources of information that are verifiable by the mindreader — such as actions of agents that are the object of mindreading and actions of others, or the (perceptible) facts that hold in some state — may warrant the attribution of goals and beliefs to agents. In this section focus is on principally *unobservable* sources for inferring the mental states of agents; namely presumptions about those agents' mental states in themselves.

2.3.1 Rationality

In reading the minds of agents, a possible heuristic that mindreaders can employ is to suppose that those agents are rational. In fact, this supposition lies at the heart of the intentional strategy (i.e. adoption of the 'intentional stance'), showing through in the following quote of Dennett (1987, p. 17) in regard to the workings of the intentional strategy: "*Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent* [...]".[5] In this section it is discussed how assumptions of rationality can be formalized as mindreading patterns.

Before formalizing the assumption of rationality, though, it is worth noting that Nichols & Stich argue against what they refer to as "rationality theories", stating that, even if one is liberal in regard to assumptions concerning the mechanisms subserving rationality in mindreading, the idea that mindreading is governed by principles of rationality is still "singularly implausible" (2003, p. 143). The arguments that Nichols & Stich, drawing from their own

[5] This quote of Dennett pertains only to prediction of behavior, but his account of the intentional strategy is concerned with both explanation and prediction.

insights and those of others, bring to bear against rationality theories can be summarized as follows (2003, p. 142–148):[6]

1. It is not made clear what 'rationality' exactly means.
2. Mindreading is typically a bottom-up process, going from behavior to attributed mental state. Rationality-based accounts, on the other hand, are top-down.
3. If rationality theories of mindreading do incorporate the bottom-up aspect, then cases in which top-down and bottom-up attribution conflict still pose a problem for those theories.

The arguments of Nichols & Stich are generally sensible, and because it is our conviction that rationality most certainly *can* play role in mindreading, it is instructive to first discuss those arguments before proceeding.

The argument that it is not made clear by rationality theories how exactly 'rationality' is to be understood, is mostly an objection to the account of Dennett. Indeed, Dennett makes liberal use of this concept in sketching his account, and Nichols & Stich (2003) object that he fails to fix its meaning in a precise way. However, this argument does not hold against the formal approach to mindreading taken in this chapter, because stating the mindreading patterns is precisely what forces one to be explicit about the concept 'rationality'. Thus, it can actually be seen as an argument in favor of considering rationality in mindreading from a formal point of view, as we do, as it helps in disambiguating the concept of 'rationality' in a context of mindreading.

Nichols & Stich's second argument also specifically concerns the rationality-based account of mindreading given by Dennett, which they find to ignore the 'bottom-up' aspect of mindreading: "*a striking feature of Dennett's account of the intentional stance* [...] *is that it doesn't even* mention *how we might arrive at intentional attributions by* observing behaviour" (2003, p. 144, original emphasis). Again, it seems that this is not an argument against considering rationality-based mindreading from a formal point of view as we intend to do, because the formal approach to mindreading that we have taken thus far actually has been bottom-up in the sense that Nichols & Stich have in mind, seeing that it has focused on ways in which mental states can be attributed to agents on grounds of their behavior.

In support of their third argument against rationality theories of mindreading, Nichols & Stich refer to cases in which humans are found to have irrational desires or beliefs. The case of irrational desires is illustrated anecdotally with an example of an agent that is known to be strongly allergic to chocolate and also known to be itself aware of this fact, yet is observed to voluntarily and contentedly eat a chocolate bar. If this agent is rational, the argument goes, then it can be presumed (top-down) that it has the desire to avoid an allergic reaction and therefore not eat chocolate. But now assume that at some point it is seen to eat chocolate, so that (bottom-up) it can be attributed the desire to do so. Clearly, Nichols & Stich point out, this agent has desires which are "irrational" (i.e. conflicting; not conforming to reason)! Thus, their argument has it, the presumption of rationality is contradicted and in this case useless for mindreading. The case of agents having irrational beliefs is illustrated by Nichols & Stich in reference to the 'feminist bank teller' experiment

[6]The term 'rational', as it is used here, is to be interpreted in its broad meaning of 'conforming to reason', in line with its interpretation in the account of Nichols & Stich (2003).

by Kahneman & Tversky (1982), which they use to show that rationality theories come to attribution of conflicting beliefs.

In regard to attribution of conflicting desires or beliefs we feel, as before, that an exact account, made possible by use of formal logic, can be eye-opening. In both philosophical (Bratman, 1987) and logical (Cohen & Levesque, 1990; Rao & Georgeff, 1991; van der Hoek et al., 1998) accounts of practical reasoning, a distinction is typically made between agents' *desires* and *goals*. Desires, it is said, can be conflicting and therefore mutually incompatible, as they represent states of affairs that agents would like to be the case (possibly, to varying degrees) *but do not necessarily strive to achieve*. Goals, on the other hand, are usually taken to be a subset of desires[7] *which agents actively attempt to bring about*. Thus, on this account it is not irrational for agents to have conflicting desires (although it probably should be considered irrational for them to actually have conflicting goals). In the anecdotal example of Nichols & Stich this is seen in the fact that the agent, although it presumedly has conflicting desires (to both have eaten chocolate and to refrain from doing so), only acts on a single goal (namely to have eaten chocolate). Employing a clear notion of rationality aids in formulating precise patterns of mindreading, which may be helpful both for philosophical and psychological approaches that are aimed at providing descriptive accounts of this subject, as well as for the constructive accounts required from the point of view of artificial intelligence.

Yet, even if it were the case that agents apparently do have conflicting goals, this should not be considered an argument against maintaining the presumption of rationality in mindreading. After all, mindreading is 'in the eye of the beholder', and the fact that an agent *appears* to be irrational (or even insane) to the mindreader, does not mean that it actually is! This can be seen by making the notion of presuming rationality (in the sense of attributing consistent goals) explicit, as in the following pattern.

Mindreading Pattern IV.11 (consistency of goals)**.** *Let* Atom *be the atoms of* $\mathscr{L}_{\mathbf{M}}$*, and* $i \in \mathsf{Ag}$ *any agent. The mindreading pattern of presuming consistent goals is then formalized as follows, for any* $p \in \mathsf{Atom}$*.*

$$\neg(\mathbf{G}_i(p) \wedge \mathbf{G}_i(\sim p))$$

Similar to how it can be presumed that agents' goal are consistent it can be presumed likewise for beliefs, as follows.

Mindreading Pattern IV.12 (consistency of beliefs)**.** *Let* Atom *be the atoms of* $\mathscr{L}_{\mathbf{M}}$*, and* $i \in \mathsf{Ag}$ *any agent. The mindreading pattern of presuming consistent beliefs is then formalized as follows, for any* $p \in \mathsf{Atom}$*.*

$$\neg(\mathbf{B}_i(p) \wedge \mathbf{B}_i(\sim p))$$

At first glance it may seem straightforward that one should assume agents' goals and beliefs to be consistent; after all, virtually any account on epistemic (doxastic) logic (Meyer & van der Hoek, 1995; Blackburn et al., 2001) or practical reasoning (Cohen & Levesque,

[7] A view which is shared by Nichols & Stich (2003, p. 67–68).

1990; Rao & Georgeff, 1991; van der Hoek et al., 1998) does so, and it is likewise the case in agent programming (Bordini et al., 2007; Dastani, 2008; Hindriks, 2009). It should be observed, though, that those accounts deal with specification or description of the beliefs agents *have*, as opposed to our account that focuses on beliefs agents *have according to the mindreader*. Considered in this light it is then not so strange to allow for inconsistency of (attributed) beliefs, as it can occur that from the perspective of a mindreader an agent clearly does have inconsistent beliefs, although this agent may itself be convinced otherwise. To see this in an example, consider an agent that firmly believes the evening star and the morning star to be two distinct celestial bodies. If this agent is unaware of the fact that the terms 'morning star' and 'evening star' actually refer to the same object, the planet Venus, then there is no inconsistency in those beliefs. Yet, a mindreader that attributes to an agent the belief that the morning star and the evening star are the same object can attribute contradictory beliefs to this agent if it proclaims "I believe this celestial body to be a star" whilst pointing at Venus in the evening, and "I believe this celestial body to be a planet" whilst doing likewise next morning. This does not mean that the agent is insane; it could be the case that it knows the theory of the morning and evening star, but does not know their position at the firmament.

A similar observation can be made of goals: even though a rational agent's goals may be consistent from its own perspective, this can be different from the perspective of a mindreader. Again this is perhaps best seen in an example. Assume that some agent is performing maintenance on a bicycle, and a mindreader has observed it to browse through an instruction manual on chain lubrication. This mindreader could on grounds of that observation attribute to this agent the goal of having the chain lubricated. Next, the mindreader observes the agent to spray the chain with a chemical substance of which the mindreader knows that it removes any lubricants, leading it to attribute the goal to the agent of having the chain 'delubricated'. If asked to explain its behavior, the agent most likely would give a rational explanation. Its goal could have been to lubricate the chain, applying the chemical substance under the belief that this action would achieve that goal. Alternatively, its goal could actually have been to delubricate the chain, and the instruction manual might have contained some information on that subject which instructed use of the aforementioned substance. In the former case the reasoning error is on the agent's behalf, as it falsely believes the substance to have lubricating properties. In the latter case it is the mindreader that has reached an erroneous conclusion, wrongly attributing the goal of having the chain lubricated on grounds of its observation that the agent browsed the instruction manual.

The example of the previous paragraph indicates that, instead of stating the mindreader to attribute inconsistency to the agent, MPs IV.11 and IV.12 are perhaps better given a paraconsistent reading in terms of inconclusiveness. Of course, this interpretation was already formalized in the definitions given in Section 1.3.1 of Chapter III. It is of specific interest here, though, because presumptions of rationality as formalized in MPs IV.11 and IV.12 entail that the mindreader *cannot* be inconclusive about the agent's goals or beliefs. This is shown in the following propositions.

Proposition IV.3. *Assume that MP IV.11 is logically valid for any* $p \in \mathsf{Atom}$ *and* $i \in \mathsf{Ag}$*, so that* $\models \neg(\mathbf{G}_i(p) \wedge \mathbf{G}_i(\sim p))$*. With* $\mathscr{L}_0$ *the language defined in Definition III.1, it then holds*

that

$$\forall \phi \in \mathscr{L}_0 : \\ \models \neg \mathbf{GInc}_i(\phi)$$

Proof. (by contradiction, using structural induction) Let $\mathsf{Lit} = \{p, \sim p \mid p \in \mathsf{Atom}\}$ be the literals of $\mathscr{L}_0$, and take any $\psi \in \mathsf{Lit}$. Let $\mathfrak{M}$ be any model for $\mathscr{L}_{\mathbf{M}}$, and take any $w \in \mathrm{W}_{\mathfrak{M}}$. Observe that if $w \models \mathbf{GInc}_i(\psi)$ then $w \models \mathbf{G}_i(\psi) \wedge \mathbf{G}_i(\overline{\psi})$, i.e. $w \models \mathbf{G}_i(p) \wedge \mathbf{G}_i(\sim p)$ for some $p \in \mathsf{Atom}$. It is given, though, that $\models \neg(\mathbf{G}_i(p) \wedge \mathbf{G}_i(\sim p))$ for any $p \in \mathsf{Atom}$, so that a contradiction is found. By structural induction it can be seen from contradiction that $\models \neg \mathbf{GInc}_i(\phi)$ for any $\phi \in \mathscr{L}_0$. Assume $\not\models \neg \mathbf{GInc}_i(\phi)$, so that for some $\phi', \phi'' \in \mathscr{L}_0$ it holds that $\phi = (\phi' \wedge \phi'')$ or $\phi = (\phi' \vee \phi'')$. For some $w' \in \mathrm{W}_{\mathfrak{M}}$ to satisfy $w' \models \mathbf{GInc}_i(\phi' \wedge \phi'')$ or $w' \models \mathbf{GInc}_i(\phi' \vee \phi'')$ it must be that (at least) $w' \models \mathbf{GInc}_i(\phi')$ or $w' \models \mathbf{GInc}_i(\phi'')$, but as this holds for no literal such a w' does not exist, proving the claim. □

Proposition IV.4. *Assume that MP IV.12 is logically valid for any $p \in \mathsf{Atom}$ and $i \in \mathsf{Ag}$, so that $\models \neg(\mathbf{B}_i(p) \wedge \mathbf{B}_i(\sim p))$. With $\mathscr{L}_0$ the language defined in Definition III.1, it then holds that*

$$\forall \phi \in \mathscr{L}_0 : \\ \models \neg \mathbf{BInc}_i(\phi)$$

Proof. Straightforward, along the lines of the proof of Proposition IV.3. □

A related implication of the presumptions of goal and belief consistency is as follows.

Proposition IV.5. *Assume that MP IV.11 is valid for any $p \in \mathsf{Atom}$ and $i \in \mathsf{Ag}$. Then*

$$\forall \phi \in \mathscr{L}_0 : \\ \models \mathbf{G}_i(\phi) \leftrightarrow \mathbf{GConc}_i(\phi)$$

Proof. (⇒, using structural induction) Let $\mathsf{Lit} = \{p, \sim p \mid p \in \mathsf{Atom}\}$ be the literals of $\mathscr{L}_0$, and take any $\psi \in \mathsf{Lit}$. Let $\mathfrak{M}$ be any model for $\mathscr{L}_{\mathbf{M}}$, and take any $w \in \mathrm{W}_{\mathfrak{M}}$. Observe that if $w \models \mathbf{G}_i(\psi)$ then from $\models \neg(\mathbf{G}_i(\psi) \wedge \mathbf{G}_i(\overline{\psi}))$ follows $w \not\models \mathbf{G}_i(\overline{\psi})$. Thus, $w \models \mathbf{G}_i(\psi) \wedge \neg \mathbf{G}_i(\overline{\psi})$, i.e. $w \models \mathbf{GConc}_i(\psi)$. For the inductive step consider any $\phi, \phi' \in \mathscr{L}_0$ and see that if it holds that $w \models \mathbf{GConc}_i(\phi)$ and $w \models \mathbf{GConc}_i(\phi')$ then $w \models \mathbf{GConc}_i(\phi \wedge \phi')$, and alternatively, if $w \models \mathbf{GConc}_i(\phi)$ or $w \models \mathbf{GConc}_i(\phi')$ then $w \models \mathbf{GConc}_i(\phi \vee \phi')$.

(⇐) Straightforward, from the definition of **GConc**. □

Proposition IV.6. *Assume that MP IV.12 is valid for any $p \in \mathsf{Atom}$ and $i \in \mathsf{Ag}$. Then*

$$\forall \phi \in \mathscr{L}_0 : \\ \models \mathbf{B}_i(\phi) \leftrightarrow \mathbf{BConc}_i(\phi)$$

Proof. Straightforward, along the lines of the proof of Proposition IV.5. □

Essentially, Propositions IV.3–IV.6 show that if the mindreader adopts the rationality assumptions of goal and belief consistency, then it must do so at the expense of being able to claim inconclusiveness about those goals and beliefs. Adopting those rationality assumptions thus effectively limits the range of possibilities the mindreader considers, which then only include cases in which the mindreader is conclusively convinced of, or holds it to be unknown, whether or not an agent believes a certain fact (or has it as goal). The paraconsistent statement that it apparently believes both facts (or has them as goal) has become void, because it cannot be true. To see what is going on here, it is important to see things in perspective. Particularly, one should keep in mind that the language $\mathscr{L}_{\mathbf{M}}$ describes states of affairs *from the mindreader's perspective*, and that the mindreader presuming rationality of the agent — in the sense of its goals and beliefs being consistent — has nothing to do with whether or not the agent actually is rational.

In reflection, after having dealt with the arguments of Nichols & Stich (2003) against rationality-based theories of mindreading, we have stumbled upon some of our own. Those are restricted to the rationality assumption of agents' being attributed consistent goals and beliefs, and can be summarized by stating that a mindreader which adopts those assumptions limits the possibilities it allows with regard to the model it maintains of the mental state of the agent whose mind it is reading. Given that the attribution of conflicting beliefs or desires is interpreted as inconclusiveness of the mindreader, as opposed to insanity of the agent, it can be argued that consistency of beliefs and goals should not be presumed, although it can be argued that the mindreader must have some grounds (in the sense of observed behavior, or by reasoning from prior attribution) to attribute inconsistent beliefs or goals; i.e. there must be some *reason* for its inconclusiveness. In the words of Dennett (1987), as in the case of attribution of false beliefs, "special stories must be told" when attributing inconsistent beliefs.

Another type of rationality assumption is found in BDI-based agent programming (Dastani, 2008), underpinned by the rationale that an agent which has the goal to achieve a certain fact should not believe this fact to be already true.

Mindreading Pattern IV.13 (goal-belief dependency)**.** *Let $\mathscr{L}_0$ be the language defined in Definition III.1, $\phi \in \mathscr{L}_0$ any expression in this language, and $i \in \mathsf{Ag}$ any agent. The presumption of goal-belief rationality is formalized as follows.*

$$\mathbf{G}_i(\phi) \rightarrow \neg \mathbf{B}_i(\phi)$$

Observe that by contraposition the following holds.

Proposition IV.7. *If the pattern of goal-belief dependency, as formalized in MP IV.13, is valid, then it holds that $\models \mathbf{G}_i(\phi) \rightarrow \neg \mathbf{B}_i(\phi)$ for any $\phi \in \mathscr{L}_0$, and $i \in \mathsf{Ag}$. It then follows that*

$$\models \mathbf{B}_i(\phi) \rightarrow \neg \mathbf{G}_i(\phi)$$

Proof. Take any $\phi \in \mathscr{L}_0$, and see that stating $\models \mathbf{G}_i(\phi) \rightarrow \neg \mathbf{B}_i(\phi)$ is equivalent to stating $\models \neg \mathbf{G}_i(\phi) \vee \neg \mathbf{B}_i(\phi)$, which is equivalent to stating $\models \mathbf{B}_i(\phi) \rightarrow \neg \mathbf{G}_i(\phi)$. □

To presume that agents conform to the rationality assumption of goal-belief dependency seems sensible, as it is not reasonable for an agent to attempt achievement of some

```
Goals:    // The agent has as its goal that 'p' and 'q' are both true.
    p and q
Beliefs: // The agent believes 'p' to be true.
    p.
Plans:    // The test on whether the agent has the goal 'p' succeeds.
    { G(p) }
```

Listing IV.1

fact that it already believes to be the case. Again, though, the same remark can be made that was made earlier with respect to the rationality presumption of goal and belief consistency, which is that MP IV.13 reflects a presumption on part of the mindreader and as such has nothing to do with agents' actual rationality. Thus, it may be that the agent is rational and attempts to achieve a fact which it believes not to be the case, and the mindreader infers that the agent is acting towards achievement of this goal but believes it to be already achieved. In such a case the mindreader can conclude the agent to be irrational, but it can also conclude that the agent must (falsely) believe its goal not to have been already achieved. It seems therefore that, similar to the presumptions of consistent goals or beliefs, the justification of presuming goal-belief dependency depends on context. If, for example, a rational (in the sense of goal-belief dependency) agent is cooperative and truthfully shares its beliefs and/or goals with the mindreader so that the mindreader can be said to have *knowledge* of those facts, then instating MP IV.13 seems justified; in other cases it may not be so.

However, another caveat applies to MP IV.13. This caveat concerns the fact that even in the case of software agents that conform to the rationality assumption of goal-belief dependency, sometimes theory and practice do not completely agree. Specifically, it holds for 2APL (Dastani, 2008) agents that such agents adhere to the rationality assumption of goal-belief dependency as explicated above, yet at the same time can have goals that they already believe to have been achieved! The reason for this phenomenon, which seems somewhat contradictory, is that goals consisting of conjoined facts are not considered achieved until all conjuncts are considered achieved *at the same time*. Thus, a 2APL agent (like the one in Listing IV.1) that has the goal 'p and q' and believes 'p' but does *not* believe 'q', successfully performs an introspective test action on whether it has the goal 'p' because the goal 'p and q' is not considered achieved until both 'p' and 'q' are true at the same time. This means that querying such an agent on whether it has the goal 'p' would have it answer affirmatively, justifying a mindreader to (mistakenly) presume this agent to believe 'p' not yet to be achieved if it knows this agent to conform to the rationality assumption of goal-belief dependency. On the other hand, it is not necessarily wrong for the mindreader to presume this agent not to have the goal 'p' if the mindreader attributes the belief 'p' to this agent, because the agent should strive only for achievement of the conjunct 'q' that remains unachieved as part of its conjoined goal, and not for achievement of 'p'. Once more, this goes to show that the presumption of rationality should by no means be made lightheartedly, and that arguments can be given in favor of or against doing so, depending on context.

2.3.2 A Frame of Mind

It can be argued that beliefs and goals which are attributed to agents should be presumed to persist, unless there is reason to presume otherwise. The latter occurs if agents themselves perform actions that are presumed to affect particular attributed beliefs or goals to them, or if other agents do so. For example, an agent that has seen the contents of a box to comprise a chocolate bar can be attributed the belief that the bar is in the box until itself has taken the bar, or another agent has taken the bar, or the box is kicked over and its contents spilled over the floor, or until some other circumstance occurs that can be presumed to change the agent's belief that the chocolate bar is in the box. In the case of human agents, such circumstances may include the fact that the agent itself *presumes* that the bar is not in the box (perhaps because it was days ago since it last observed the bar to be in the box, and it considers it likely that someone has eaten it), or that it simply *forgets* that the chocolate bar is in the box. In certain contexts it is justifiable to assume that beliefs and goals only change as a (direct) result of agents' actions, an assumption which is made in the current section.

The assumption that beliefs and goals of agents persist, unless affected by actions, is similar to presumptions made in regard to the *frame problem* in artificial intelligence (Reiter, 1991). The frame problem relates to the fact that, in dynamic domains, it does not suffice to state only the way the world changes as a result of actions; one should also somehow make explicit how the world stays inert as a result of actions. A possible solution to the frame problem is to formulate so-called 'frame axioms', that can be informally described as reflecting that any fact which is not explicitly stated to change can be assumed to remain the same. The frame problem is typically presented as pertaining to intelligent agents' model of the (ontic) state of the world, but can be generalized to include beliefs and goals attributed to others. A similar characterization can then be given of presumed persistence of agent's goals and beliefs; i.e. that any attributed belief or goal which is not explicitly stated to be affected by actions can be assumed to remain the same. The 'laws of inertia' that concern beliefs and goals attributed to others can, accordingly, be regarded as *frame axioms of the mind*. This is done in the following mindreading patterns, similar to Zhang & Foo (2005).

Mindreading Pattern IV.14 (belief inertia). *Let $\alpha \in \mathsf{Act}$ be any action, $i, j \in \mathsf{Ag}$ any agents, and ψ any literal of $\mathscr{L}_0$. For some $\phi \in \mathscr{L}_{\mathbf{M}}$ let it be the case that $\models \phi \rightarrow [\alpha^-]_j \neg \mathbf{B}_i(\psi)$. The mindreading pattern of presumed belief persistence is then as follows.*

$$(\mathbf{B}_i(\psi) \wedge \neg\phi) \rightarrow [\alpha^-]_j(\mathbf{B}_i(\psi))$$

A similar law of inertia can be formulated with respect to goals.

Mindreading Pattern IV.15 (goal inertia). *Let $\alpha \in \mathsf{Act}$ be any action, $i, j \in \mathsf{Ag}$ any agents, and ψ any literal of $\mathscr{L}_0$. For some $\phi \in \mathscr{L}_{\mathbf{M}}$ let it be the case that $\models \phi \rightarrow [\alpha^-]_j \neg \mathbf{G}_i(\psi)$. The mindreading pattern of presumed belief persistence is then as follows.*

$$(\mathbf{G}_i(\psi) \wedge \neg\phi) \rightarrow [\alpha^-]_j(\mathbf{G}_i(\psi))$$

For more on the frame problem in a context of dynamic logic, see Zhang & Foo (2005). This section is now concluded with some remarks, after which an illustrative example is presented.

2.3.3 Remarks

In previous sections multiple mindreading patterns have been presented, that warrant the attribution of beliefs and goals to agents on various grounds. Those patterns can constitute the basis for a *theory of mind* — using 'theory' here in its strict logical interpretation — if they are considered as axioms of a logic for reasoning about agents' behavior. Formalizing such a theory is, however, considered out of the scope of this thesis, so that the term 'mindreading' is favored over 'theory of mind' and the interpretation of mindreading patterns is restricted to model-theoretic semantics, as well as more informal discussion of those patterns in relation to psychological models of mindreading.

Regardless of terminology, as with scientific theories, the principal concern of mental state attribution is both with *explanation* and *prediction*; particularly, explanation and prediction of agents' behavior. In the case of mindreading, mental states are the theoretical constructs which are assumed in order to account for agents' behavior, comprised of observable actions. From this perspective the preceding sections have focused on explanation, in the sense that 'laws' have been formulated (in the guise of mindreading patterns) that make claims about the unobservables. Little has been said about prediction of actions, though, because no patterns have been formulated that, given particular mental states, implicate agents' actions.[8] This may appear to be a flaw in our account because it is a credo of science, explicated by Popper, that theories should be judged by their falsifiability, meaning that one should be able to falsify the predictions made by a theory about observables with regard to actual observations. In reply to this we state that existing models of, for example, (BDI-based) agency can be used to complement an approach to behavior explanation with behavior prediction, following those models' principles on how the mental state of rational agents determines their behavior. Also, the prediction made by our approach is of a different nature; it concerns the claim that formalizing regularities of mindreading (as we have attempted) is useful in realizing software agents that have a notion of the mental states of others, and this is a claim that can, of course, be evaluated in multiple ways.

3 Example: Testing for False Beliefs

This section focuses on a classic task for understanding of the concept of 'false beliefs', which according to Dennett (1978) (commenting on Premack & Woodruff (1978)) might constitute a litmus test for determining whether an organism has a 'theory of mind'. The suggestion of Dennett has been followed by Wimmer & Perner (1983) by designing a task for testing healthy children, which was modified by Baron-Cohen et al. (1985) to be usable for testing children with autism and Down's Syndrome as well. More recently the importance of this task as determinant for theory of mind has been downplayed (Nichols & Stich, 2003), specifically with regard to the fact that failing the false-belief task does not indicate absence of theory of mind (Bloom & German, 2000) because the subject may have some grasp of the fact that others have mental state, but not be able to pass the false-belief task

[8] Perhaps MP IV.3 can be considered to concern prediction of actions, but then only in the sense that the mindreader can predict what will be the case as a result of actions, not which actions the agent will perform.

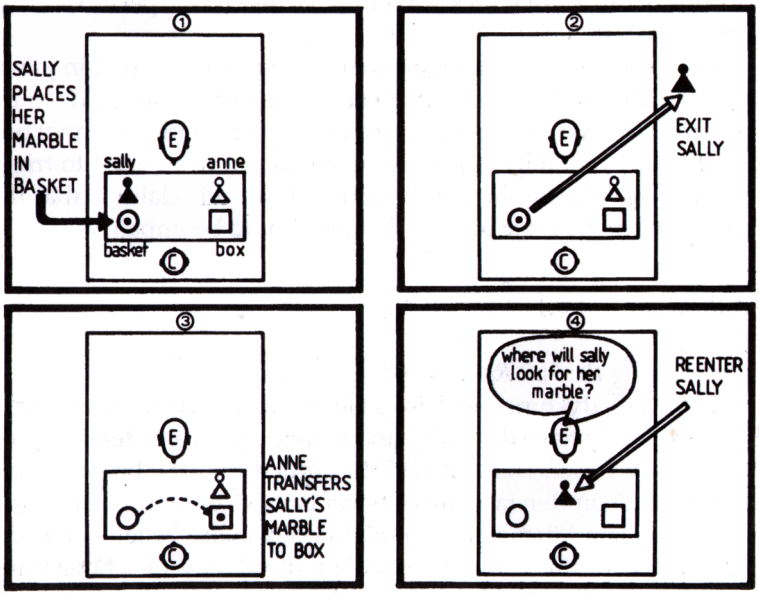


Figure IV.4: (reproduced from (Baron-Cohen, 1995))

because of its general difficulty; in this sense, the analogy to the litmus test is misleading. Nevertheless, it is generally agreed upon (Baron-Cohen, 1995; Howlin et al., 1999; Bloom & German, 2000; Nichols & Stich, 2003) that successful completion of this task is an indicator of quite advanced mindreading abilities, which is why it is of interest to us here.

The Sally-Anne false belief task as recounted by Baron-Cohen (1995) forms the basis of this section, and it is introduced in Section 3.1. In Section 3.2 it is then shown how the mindreader (i.e. the subject which is tested for understanding of false beliefs) can be modeled using the approach of Section 2 in a formalized version of this task. To this extent mindreading patterns are given, and the mindreader's 'understanding' of the progressive phases in this task is shown by means of a series of (simple) models that represent a view of the mindreader of consecutive stages in this task. Then, it is shown how this endeavour in modeling translates to implementation, yielding some insights in regard to desirable features of agent programming frameworks.

3.1 The Sally-Anne Task

The Sally-Anne task, shown pictorially in Figure IV.4, involves two main characters, Sally and Anne, and an observer (denoted 'C', for 'Child', in the figure, where 'E' denotes the experimenter). It also features objects, namely a basket, a box, and a marble. The observer is shown or told the following course of events; noting that the numbering corresponds to the stages in the figure:

1. Sally places her marble in the basket.
2. Sally leaves the 'room'.
3. In Sally's absence, Anne transfers the marble from the basket to the box.
4. Sally returns, and the observer is asked "Where will Sally look for the marble?".

If the observer is a mindreader then he/she should answer in response to the question asked at the end of the scenario that Sally will look in the basket, given the knowledge that this is where Sally placed the marble herself and the assumption that she is unaware of it having been transferred. Experimental results show that the vast majority of healthy children and children with Down's Syndrome give this answer, whereas most autistic children state that Sally will look in the box; supposedly because this is where the children themselves know the marble to be located. Thus, autistic children do not attribute the discrepant (false) belief to Sally, and fail to pass this basic mindreading test.

3.2 Modeling the Mindreader in the Sally-Anne Task

In modeling the Sally-Anne task, the following primitive actions are assumed.

$$\alpha_1 = \mathsf{put\_marble\_in\_basket}$$
$$\alpha_2 = \mathsf{take\_marble\_from\_basket}$$
$$\alpha_3 = \mathsf{put\_marble\_in\_box}$$
$$\alpha_4 = \mathsf{take\_marble\_from\_box}$$
$$\alpha_5 = \mathsf{enter\_room}$$
$$\alpha_6 = \mathsf{leave\_room}$$

Furthermore, the following atomic propositions are assumed.

$$p_1 = \mathsf{marble\_in\_basket}$$
$$p_2 = \mathsf{marble\_in\_box}$$
$$p_3 = \mathsf{anne\_has\_marble}$$
$$p_4 = \mathsf{sally\_has\_marble}$$
$$p_5 = \mathsf{anne\_in\_room}$$
$$p_6 = \mathsf{sally\_in\_room}$$

The interpretation of the above actions and propositions should be self-explanatory, and in the following sections these formal ingredients are used to present the model a mindreading observer has of the Sally-Anne scenario. We hereby slightly diverge from the scenario as it is explained by Baron-Cohen (1995), for the sake of clarity and also to illustrate particular features of our approach; those deviations will be explained when encountered. In regard to presentation it should be noted that we discuss various models that correspond to stages in the scenario, whereby focus is on individual states. Those are each equipped with an indices 'x,y', where 'x' refers to the model representing the particular stage of the

scenario to which this state belongs, corresponding to the numbers in Figure IV.4. The second index 'y' refers to the actual stage of the scenario. Thus, in reference to the figure, the state with index '3,2' corresponds to the state of affairs after Sally left the room, and forms part of the model representing the state of affairs after the mindreader has witnessed the first three actions (Sally depositing her marble, Sally leaving, and Anne transferring the marble). Accordingly, 'n,n' is the state representing 'now' in any of the models. The first deviation from the 'official' scenario is that we have opted to include an initial state '0' in the models, representing the state of affairs before Sally dropped her marble (note that the presumption of a state preceding the initial action is actually forced by our action model; Proposition III.13, specifically).

3.2.1 The Initial State of Affairs

First, the initial state $w_{0,0}$ is modeled, depicted below as model of phase '0'.

$$\bullet_{w_{0,0}}$$

It is given in the first step of the scenario in Figure IV.4 that both Sally and Anne are in the room when Sally places her marble in the basket. Thus, it is assumed with regard to the ontic state of affairs that in the preceding state both characters are in the room and that Sally is in possession of the marble, as follows.

IV.3

$$w_{0,0} \models p_4 \wedge p_5 \wedge p_6$$

With regard to the relations between ontic facts, it is defined that the mindreader takes p_1–p_4 to be mutually exclusive, as follows. It should be observed that those relations are stated as validities, but that this is only so for the scope of the example presented here; likewise for any validities mentioned in the remainder of this example.

$$\begin{aligned}
&\models p_1 \leftrightarrow \neg(p_2 \vee p_3 \vee p_4)\\
&\models p_2 \leftrightarrow \neg(p_1 \vee p_3 \vee p_4)\\
&\models p_3 \leftrightarrow \neg(p_1 \vee p_2 \vee p_4)\\
&\models p_4 \leftrightarrow \neg(p_1 \vee p_2 \vee p_3)
\end{aligned}$$

Apart from taking the aforementioned relations to hold for the ontic state of affairs, the mindreader can attribute certain beliefs by default to both Anne and Sally (identified here by subscript 'a' and 's', respectively). Particularly, it is reasonable for the mindreader to suppose that they believe themselves to be in the room if this is the case, and to believe otherwise if not. Formally, this translates to the following.

$$\begin{aligned}
&\models p_3 \rightarrow \mathbf{B}_{\mathsf{a}}(p_3) && \models p_4 \rightarrow \mathbf{B}_{\mathsf{s}}(p_4)\\
&\models \neg p_3 \rightarrow \mathbf{B}_{\mathsf{a}}(\sim p_3) && \models \neg p_4 \rightarrow \mathbf{B}_{\mathsf{s}}(\sim p_4)\\
&\models p_5 \rightarrow \mathbf{B}_{\mathsf{a}}(p_5) && \models p_6 \rightarrow \mathbf{B}_{\mathsf{s}}(p_6)\\
&\models \neg p_5 \rightarrow \mathbf{B}_{\mathsf{a}}(\sim p_5) && \models \neg p_6 \rightarrow \mathbf{B}_{\mathsf{s}}(\sim p_6)
\end{aligned}$$

Apart from attributing the belief to Anne and Sally that they themselves are in the room if the mindreader thinks this is so, the fact that the other is in the room is taken to serve as perceptory indicator for attribution of the corresponding belief to either of them. This occurs along the lines of MP IV.10 and translates to the following.

$$\models (p_5 \wedge p_6) \rightarrow \mathbf{B}_{\mathsf{a}}(p_6) \qquad \models (p_5 \wedge p_6) \rightarrow \mathbf{B}_{\mathsf{s}}(p_5)$$
$$\models (p_5 \wedge \neg p_6) \rightarrow \mathbf{B}_{\mathsf{a}}(\sim p_6) \qquad \models (\neg p_5 \wedge p_6) \rightarrow \mathbf{B}_{\mathsf{s}}(\sim p_5)$$

Thus, for example, if the mindreader observes Anne to be in the room and Sally not, then this is reason to attribute the belief to Anne that Sally is not in the room. The aforementioned patterns translates to the following belief attribution in the initial state, where it is noteworthy that the mindreader takes Sally to have the belief that she has the marble (as the mindreader is itself aware of the fact that she does), but not Anne.

$$w_{0,0} \models \mathbf{B}_{\mathsf{a}}(p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{s}}(p_4 \wedge p_5 \wedge p_6)$$

3.2.2 Sally Losing Her Marble

As Figure IV.4 shows, the first action that occurs in the scenario is that of Sally placing her marble in the basket. It is assumed here that this mindreader has an unambiguous interpretation of the observed events, such that the succession of states is linear and only a single branch exists.

$$\bullet_{w_{1,0}} \xrightarrow{\alpha_1} \bullet_{w_{1,1}}$$

It is assumed that the action of placing the marble in the basket has the effect of the marble being in the basket, so that the following holds.

$$w_{1,1} \models p_1 \wedge p_5 \wedge p_6 \wedge \mathbf{Done}_{\mathsf{s}}(\alpha_1)$$

The consequence of α_1 holds irrespective of whether it is performed by Anne or Sally. If it is assumed that the agent having possession of the marble is a precondition to this action, then attribution must be relative to the agent performing the action; here it is not. In the scenario it is Sally who performs this action, but the patterns of ontic attribution corresponding to this action are presented here for any agent $i, j \in \{\mathsf{a}, \mathsf{s}\}$.

$$\models \mathbf{Done}_i(\alpha_1) \wedge \mathbf{B}_j(\mathbf{Done}_i(\alpha_1)) \wedge [\alpha_1{}^-]_i(p_3) \rightarrow (\mathbf{B}_j(p_1) \wedge [\alpha_1{}^-]_i(\mathbf{B}_j(p_3)))$$
$$\models \mathbf{Done}_i(\alpha_1) \wedge \mathbf{B}_j(\mathbf{Done}_i(\alpha_1)) \wedge [\alpha_1{}^-]_i(p_4) \rightarrow (\mathbf{B}_j(p_1) \wedge [\alpha_1{}^-]_i(\mathbf{B}_j(p_4)))$$

It is assumed that both Anne and Sally are aware of the actions they perform, as in MP IV.8. Whether or not either of them believes the other to have done this action is taken by the mindreader to be dependent on their presence in the room, as follows.

$$\models p_5 \wedge \mathbf{Done}_{\mathsf{s}}(\alpha_1) \rightarrow \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1))$$
$$\models p_6 \wedge \mathbf{Done}_{\mathsf{a}}(\alpha_1) \rightarrow \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{a}}(\alpha_1))$$

This yields the following interpretation with regard to the beliefs the mindreader attributes to Anne and Sally.

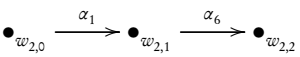
$$w_{1,0} \models \mathbf{B}_{\mathsf{a}}(p_4 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{s}}(p_4 \wedge p_5 \wedge p_6)$$
$$w_{1,1} \models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1) \wedge p_1 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1)) \wedge p_1 \wedge p_5 \wedge p_6)$$

It is noteworthy that the state $w_{1,0}$ in the model that represents the mindreader having observed Sally's action of putting the marble into the basket differs from the state $w_{0,0}$ in the model representing the initial state of affairs, even though both refer to the same phase of the scenario, albeit at different moments! This can be explained by the fact that the mindreader shares its perception of Sally's action with Anne, and updates its model of Anne's mental state with the belief that Sally possessed the marble in the initial state, which is a reasonable attribution given the fact that this fact is a precondition to Anne's subsequent observation of this action. Note that the remark made at the end of Section 2.2.2 also applies here, which is that it would be yet more accurate to state that in $w_{1,1}$ Anne has the belief that in the state preceding the action of Sally putting away the marble she had the marble in her possession, which, however, is not expressible in our language.

3.2.3 Exit Sally

Next, Sally is observed to leave the room, yielding the following model.

IV.3



The mindreader takes the ontic state of affairs in the 'current' state to be as follows.

$$w_{2,2} \models p_1 \wedge p_5 \wedge \neg p_6 \wedge \mathbf{Done}_{\mathsf{s}}(\alpha_6)$$

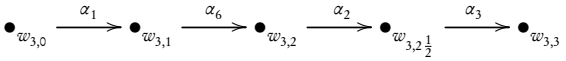
A crucial step in the false belief task, which is typically not mentioned so explicitly in psychological discussions, is the fact that its successful completion requires the presumed persistence of beliefs, along the lines of our 'frame axiom of the mind' given in MP IV.14, because a mindreader that forgets the belief about the marble's location, which was attributed to Sally, would possibly fail to correctly answer the related question. Thus, Sally's action of leaving the room is presumed by the mindreader not to affect any of her beliefs that are not directly related to this action. The belief attribution which is furthermore specifically warranted by Sally leaving the room concerns the fact that she performed that action and is therefore not in the room, which along with the presumption of belief persistence (not made explicit formally) results in the following.

$$w_{2,0} \models \mathbf{B}_{\mathsf{a}}(p_4 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{s}}(p_4 \wedge p_5 \wedge p_6)$$
$$w_{2,1} \models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1) \wedge p_1 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1)) \wedge p_1 \wedge p_5 \wedge p_6)$$
$$w_{2,2} \models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6) \wedge p_1 \wedge p_5 \wedge \sim p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6)) \wedge p_1 \wedge p_5 \wedge \sim p_6)$$

Thus, the stage is set for this scenario's main act.

3.2.4 Anne Transfers the Marble

The central action of the Sally-Anne false-belief task is Anne's transferring of the marble in Sally's absence. In our model this action is split up into two distinct actions; Anne taking the marble from the basket (the first action) and putting it into the box (second action). For consistency with the numbering scheme that takes states identified by 'n, n' to correspond to the actual state of phase n in Figure IV.4, an intermediate state is posited in between $w_{3,2}$ and $w_{3,3}$.



The actions performed by Anne result in the following ontic states of affairs.

$$w_{3,2\frac{1}{2}} \models p_3 \wedge p_5 \wedge \neg p_6 \wedge \mathbf{Done}_{\mathsf{a}}(\alpha_2)$$
$$w_{3,3} \models p_2 \wedge p_5 \wedge \neg p_6 \wedge \mathbf{Done}_{\mathsf{a}}(\alpha_3)$$

Similar to the ontic attribution patterns stated the action of putting the marble into the basket (α_1), those patterns are stated here for the actions of taking it out of the basket (α_2) and putting it into the box (α_3). Note that the postcondition of α_2 is relative to the agent performing the action, yielding the following mindreading patterns in which the identifiers 'a' and 's' are of special note.

$$\models \mathbf{Done}_{\mathsf{a}}(\alpha_2) \wedge \mathbf{B}_i(\mathbf{Done}_{\mathsf{a}}(\alpha_2)) \wedge [\alpha_1{}^-]_{\mathsf{a}}(p_1) \rightarrow (\mathbf{B}_i(p_3) \wedge [\alpha_1{}^-]_{\mathsf{a}}(\mathbf{B}_i(p_1)))$$
$$\models \mathbf{Done}_{\mathsf{s}}(\alpha_2) \wedge \mathbf{B}_i(\mathbf{Done}_{\mathsf{s}}(\alpha_2)) \wedge [\alpha_1{}^-]_{\mathsf{s}}(p_1) \rightarrow (\mathbf{B}_i(p_4) \wedge [\alpha_1{}^-]_{\mathsf{s}}(\mathbf{B}_i(p_1)))$$
$$\models \mathbf{Done}_i(\alpha_3) \wedge \mathbf{B}_j(\mathbf{Done}_i(\alpha_3)) \wedge [\alpha_3{}^-]_i(p_3) \rightarrow (\mathbf{B}_j(p_2) \wedge [\alpha_3{}^-]_i(\mathbf{B}_j(p_3)))$$
$$\models \mathbf{Done}_i(\alpha_3) \wedge \mathbf{B}_j(\mathbf{Done}_i(\alpha_3)) \wedge [\alpha_3{}^-]_i(p_4) \rightarrow (\mathbf{B}_j(p_2) \wedge [\alpha_3{}^-]_i(\mathbf{B}_j(p_4)))$$

As with the action α_1, attribution of the belief that either of the actions α_2 or α_3 has been performed by the other agent is in this scenario dependent on the presence of agents in the room, as follows.

$$\models p_5 \wedge \mathbf{Done}_{\mathsf{s}}(\alpha_2) \rightarrow \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_2))$$
$$\models p_5 \wedge \mathbf{Done}_{\mathsf{s}}(\alpha_3) \rightarrow \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_3))$$
$$\models p_6 \wedge \mathbf{Done}_{\mathsf{a}}(\alpha_2) \rightarrow \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{a}}(\alpha_2))$$
$$\models p_6 \wedge \mathbf{Done}_{\mathsf{a}}(\alpha_3) \rightarrow \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{a}}(\alpha_3))$$

Persistence of beliefs is again presumed, so that if agents are not aware of actions having been performed their corresponding attributed beliefs are not affected. This results in the

following model the mindreader has of the third stage (cf. Figure IV.4).

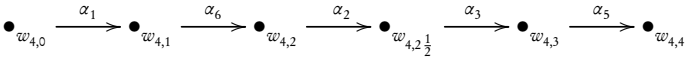
$$
\begin{aligned}
w_{3,0} &\models \mathbf{B}_{\mathsf{a}}(p_4 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{s}}(p_4 \wedge p_5 \wedge p_6) \\
w_{3,1} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1) \wedge p_1 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1)) \wedge p_1 \wedge p_5 \wedge p_6) \\
w_{3,2} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6) \wedge p_1 \wedge p_5 \wedge \sim p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6)) \wedge p_1 \wedge p_5 \wedge \sim p_6) \\
w_{3,2\frac{1}{2}} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6) \wedge p_1 \wedge p_5 \wedge \sim p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{a}}(\alpha_2)) \wedge p_3 \wedge p_5 \wedge \sim p_6) \\
w_{3,3} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6) \wedge p_1 \wedge p_5 \wedge \sim p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{a}}(\alpha_3)) \wedge p_2 \wedge p_5 \wedge \sim p_6)
\end{aligned}
$$

Thus, Sally's absence from the room leads the mindreader to presume that she is unaware of the actions performed by Anne; in fact, the presumption of belief persistence results in no change in the beliefs attributed to Sally, so that in $w_{3,3}$ she is still attributed the belief that she just performed the action of leaving the room. In some cases it may be desirable to restrict belief persistence in certain ways, e.g. to model the fact that agents (deliberately) 'forget' particular facts, but as stated before this is not our main concern. More interesting is the fact that Sally at this point — both in states $w_{3,2\frac{1}{2}}$ and $w_{3,3}$ — is attributed the belief that the marble is still in the basket, which from the perspective of the mindreader is a false belief.[9]

3.2.5 The Return of Sally

IV.3

The last action in this scenario, occurring in scene 4 of Figure IV.4, is that of Sally re-entering the room.



This action results in the following ontic state of affairs.

$$w_{4,4} \models p_2 \wedge p_5 \wedge p_6$$

The mindreading patterns presented thus far suffice to formalize the model the mindreader has at the final stage in this course of events, which is as follows.

$$
\begin{aligned}
w_{4,0} &\models \mathbf{B}_{\mathsf{a}}(p_4 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{s}}(p_4 \wedge p_5 \wedge p_6) \\
w_{4,1} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1) \wedge p_1 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_1)) \wedge p_1 \wedge p_5 \wedge p_6) \\
w_{4,2} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6) \wedge p_1 \wedge p_5 \wedge \sim p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6)) \wedge p_1 \wedge p_5 \wedge \sim p_6) \\
w_{4,2\frac{1}{2}} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6) \wedge p_1 \wedge p_5 \wedge \sim p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{a}}(\alpha_2)) \wedge p_3 \wedge p_5 \wedge \sim p_6) \\
w_{4,3} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_6) \wedge p_1 \wedge p_5 \wedge \sim p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{a}}(\alpha_3)) \wedge p_2 \wedge p_5 \wedge \sim p_6) \\
w_{4,4} &\models \mathbf{B}_{\mathsf{s}}(\mathbf{Done}_{\mathsf{s}}(\alpha_5) \wedge p_1 \wedge p_5 \wedge p_6) \wedge \mathbf{B}_{\mathsf{a}}(\mathbf{Done}_{\mathsf{s}}(\alpha_5)) \wedge p_2 \wedge p_5 \wedge p_6)
\end{aligned}
$$

[9] The term 'discrepant belief' utilized by Nichols & Stich (2003) can be considered to be more accurate than the term 'false belief' in regard to the phenomenon measured in false-belief tests, because properties of interest in such tests occur also if the mindreader is mistaken in its beliefs and the agent is actually correct. The relevant fact is that the mindreader recognizes that the agent has discrepant beliefs. However, so as not to deviate from literature on this topic, we have opted to use the term 'false belief'.

In the false-belief task as described by Baron-Cohen (1995), the mindreader is at this point questioned in regard to the place where Sally will look for her marble. It is hereby assumed that the answer to this question depends on the mindreader's model of where itself believes the marble to be and the belief about the marble's location that it attributes to Sally, which are as follows.

$$w_{4,4} \models p_2 \wedge \mathbf{B}_{\mathrm{s}}(p_1)$$

Thus, our mindreader, conceived on basis of the patterns put forward in this chapter, attributes the correct (i.e. 'false') belief to Sally, namely '$\mathbf{B}_{\mathrm{s}}(p_1)$', whereas itself believes 'p_2'. It therefore can be said to pass this test for the ability to recognize discrepant beliefs. The aforementioned patterns allow for further variation; for example, the mindreader can presume that Sally's goal initially was to have the marble in the basket (i.e. $w_{x,1} \models \langle \alpha_1{}^{-} \rangle_{\mathrm{s}}(\mathbf{G}_{\mathrm{s}}(p_1))$), or that when she returns she has the goal to have the marble again because it is asked where she will look for it (i.e. $w_{4,4} \models \mathbf{G}_{\mathrm{s}}(p_4)$). However, variations such as these are considered to be beyond the scope of this example. Instead, in subsequent sections our focus is on illustrating the usefulness of the formal model from an implementation-oriented point of view.

3.3 Implementation of the False Belief Task

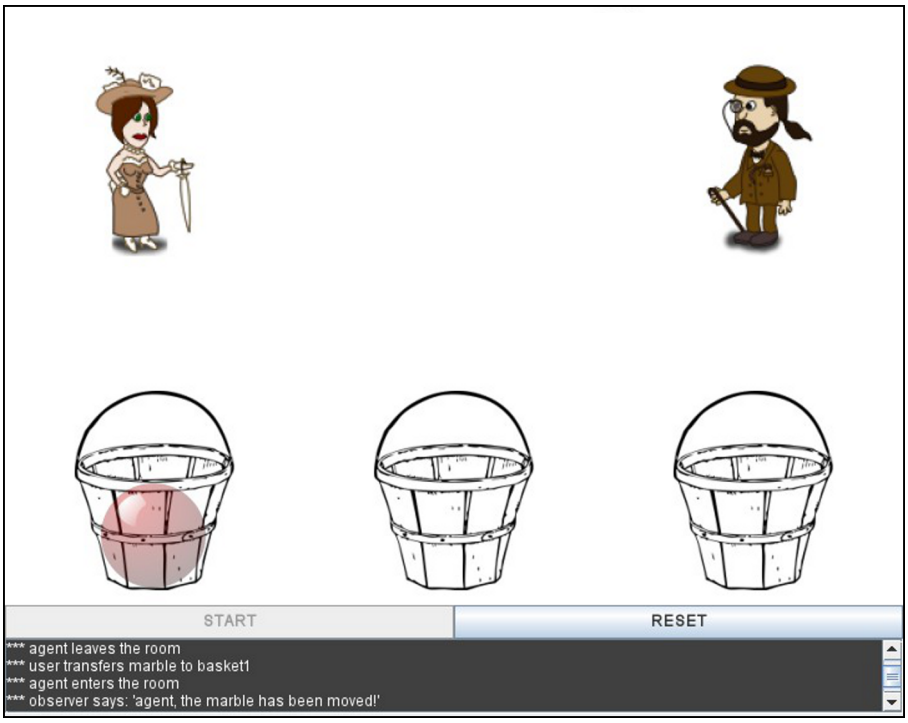
In earlier work we presented an implementation inspired by the Sally-Anne test for false beliefs using standard 2APL (Sindlar et al., 2009b), and here a more elaborate account of this scenario is presented using modular 2APL (Dastani, 2009). This implementation involves a software agent and an observer that fulfill the roles of Sally and the mindreader ('Child') in the Sally-Anne task, respectively, and features the user in the role of Anne. The scenario, of which the MAS specification is shown in Listing IV.2 and of which the graphical interface is shown in Figure IV.5, progresses as follows:

- The 'agent' reacts to the event of the user pressing the 'start' button, triggering it to deposit the (rather huge) marble into the middle basket.
- After depositing its marble, 'agent' leaves the room for 8–12 seconds.
- The environment allows for transferring the marble into any container by 'agent' or user, irrespective of whether 'agent' is present or not.
- Events generated by performance of actions by 'agent' or the user are sent to both 'agent' and 'observer', if they are present when the actions is performed and the action is not their own, representing perception of the action.

Both the 'agent' and the 'observer' are based on a template, which is shown in Listing IV.3. This template features rules for update of beliefs, and a procedural rule for responding

```
agent    : agent.2apl    @ env
observer: observer.2apl @ env
```

Listing IV.2: false-belief_test.mas



IV.3

Figure IV.5

to the event that represents the fact that the user transferred the marble to some basket. Code of the 'agent' is furthermore shown in Listing IV.4, and that of the observer in Listing IV.5. The procedural rule in the template comprises an update to a profile module called 'agent_profile', which is a feature that is offered by the modular version of 2APL. By means of modules, agents can instantiate an agent-like entity on which they can perform operations such as querying and updating beliefs or goals. Moreover, modules can be executed by agents until a certain stop condition is reached, which occurs by suspending agents' own deliberation and transferring resources to the module (Dastani, 2009).

As Listing IV.5 shows, the profile instantiated by the 'observer' is based on the same template that the 'agent' and the 'observer' itself are based on. Interestingly, the use of agent-like profile modules for modeling others is conceptually in line with the mindreading account given by Nichols & Stich (2003), as discussed in Section 1.2.2 and depicted in Figure IV.2. It was stated there that Nichols & Stich argue that mindreaders use their own inference mechanisms for reasoning about others, and store beliefs and goals attributed to others in a so-called 'Possible World Box' using the same logical form as used to represent their own beliefs and goals. In this sense, profile modules can be regarded as PWBs, given that they are operated on using agents' own resources, and comprise goals, beliefs, rules, etc. that have the same form as those which agents employ themselves. The language $\mathscr{L}_{\mathbf{M}}$ is also of note

```
// Rules of the form '{precondition} Update {postcondition}'
BeliefUpdates:
   { not in(Obj,Place) }  In(Obj,Place)     { in(Obj,Place) }
   { in(Obj,Place) }      NotIn(Obj,Place)  { not in(Obj,Place) }
   { not has(Agent,Obj) } Has(Agent,Obj)    { has(Agent,Obj) }
   { has(Agent,Obj) }     NotHas(Agent,Obj) { not has(Agent,Obj) }
PC-rules:
   // Update beliefs about the marble's location, of 'self' and in
   // 'agent_profile', if the user deposits it or picks it up
   event(action(user,put(Obj,NewPlace)),_) <- true | {
         if B(in(Obj,OldPlace)) then { NotIn(Obj,OldPlace) };
           In(Obj,NewPlace);
           if B(in(agent,room)) then {
               if agent_profile.B(in(Obj,AOldPlace)) then {
                   agent_profile.updateBB(not in(Obj,AOldPlace)) };
               agent_profile.updateBB(in(Obj,NewPlace)) } }
   event(action(user,take(Obj,NewPlace)),_) <- true | { ... }
```

Listing IV.3: template.2apl

here, because, if it is used for specification of a mindreading observer, then the operators that serve as 'markers' of propositions (as discussed in Section 1.3.3) can be taken to signify the utilization of profile modules for operating on the marked propositions. This agrees nicely with the account of mindreading given by Nichols & Stich (2003), as it shows how marked propositions are given a functionally different treatment.

Table IV.1 shows the progression of this scenario in terms of events that occur (left column), the software-based participants ('agent', 'observer') and the agent profile maintained by the observer (top row), and relevant beliefs (cell contents). As this table shows, at the end of this particular evolution of the scenario, the 'observer' has the (correct) belief that the marble is in the first basket, whereas the 'agent' believes it to be in the second basket. Notably, the observer's 'agent_profile' module contains this same belief; thus, the observer can be said to be a mindreader, given that on grounds of the witnessed events it attributes a belief to the observer 'agent' which is discrepant from its own.

	agent	**observer**	**agent_profile**
(initial state; no events processed)	has(me,marble).	in(agent,room). has(agent,marble).	
action(agent,put(marble,basket2))	in(marble,basket2).	in(agent,room). in(marble,basket2).	in(marble,basket2).
action(agent,leave)	in(marble,basket2).	in(marble,basket2).	in(marble,basket2).
action(user,put(marble,basket1))	in(marble,basket2).	in(marble,basket1).	in(marble,basket2).
action(agent,enter)	in(marble,basket2).	in(agent,room). in(marble,basket1).	in(marble,basket2).

Table IV.1: Scenario trace: events in left column, (attributed) beliefs in remaining columns.

```
Include: // The agent program subsumes the template
    template.2apl
Beliefs: // Initial belief that it possesses the marble
    has(me,marble).
PC-rules: // Agent is scripted to place the marble, exit, and reappear
    event(command(start),_) <- has(me,marble) | {
            @env(put(marble,basket2),_);
            In(marble,basket2);
            NotHas(me,marble);
            @env(leave(),_);
            @env(sleep(8,12),_);
            @env(enter(),_) }
```

Listing IV.4: agent.2apl

3.3.1 Reflection

The implementation presented above employs events that represent the occurrence of actions, and it is noteworthy that the mindreader employs those events as grounds for maintaining both its own beliefs as well as those which it attributes to the agent. In this sense the implementation is closely related to the logical model of the false belief task discussed in Section 3.2. It should also be noted that the rules for handling events comprise attribution or 'de-attribution' of beliefs, which is in line with the mindreading patterns presented earlier. In the implementation, persistence of attributed beliefs occurs by default, as facts are stored in the 'belief box' or 'goal box' of profile modules until they are explicitly manipulated. Thus, the mindreader must take action in order to ensure that facts attributed as beliefs or goals are de-attributed. The logical model can thus be seen to be a concise formal specification of the scenario, and a useful guideline for implementation.

A notable drawback of the implementation is related to the ways in which modules can be used in 2APL. Observe that the implementation of the 'observer' (cf. Listing IV.5) involves explicit updating of both the observer's own beliefs as well as the profile it maintains of the 'agent', upon reception of the event that represents the 'agent' having placed the marble somewhere, or having taken it. The semantics of those updates are quite similar, as they involve addition and deletion of the same facts in both the 'observer's own beliefs and its profile module of the 'agent'. It seems to us that the duplicity of having to state those updates separately is unnecessary, and we here give a recommendation for improving the use of modules for profiling others in 2APL. Implementation and use of 2APL modules for profiling agents would be significantly improved, in our opinion, if procedural rules could also be stored in modules and called by agents at run-time. Furthermore, employing constructs of the form 'profile(Agent).procedure(...)' instead of 'agent_profile.procedure(...)', i.e. with variable agent identifiers as argument to generic profiling features, is considered another useful improvement. The resulting changes would, in our estimation, lead to better reuse of existing code — thus decreasing the burden on the programmer (with regard to both implementation and debugging) and improving legibility of programs — in addi-

```
Include:  // The observer program subsumes the template
    template.2apl
Beliefs:  // The initial beliefs of the observer
    in(agent,room).
    has(agent,marble).
Plans:    // Creation of the 'agent_profile' module
    { create(template,agent_profile) }
// Event-based updates to the observer's own beliefs, and
// similarly to those stored in the 'agent_profile' module
PC-rules:
    event(action(agent,enter),_) <- not in(agent,room) | {
            In(agent,room) }
    event(action(agent,leave),_) <- in(agent,room) | {
            NotIn(agent,room) }
    event(action(agent,put(Obj,NewPlace)),_) <- true | {
            NotHas(agent,Obj);
            In(Obj,NewPlace);
            agent_profile.updateBB(not has(agent,Obj));
            agent_profile.updateBB(in(Obj,NewPlace)) }
    event(action(agent,take(Obj,OldPlace)),_) <- true | {
            NotIn(Obj,OldPlace);
            Has(agent,Obj);
            agent_profile.updateBB(not in(Obj,OldPlace));
            agent_profile.updateBB(has(agent,Obj)) }
```

Listing IV.5: observer.2apl

tion to a more straightforward usage of templates as basis for the program of both agents themselves, as well as any profile modules.

To see the basic idea that we have in mind in regard to the updating of beliefs by means of generic procedures in the template, consider the code of Listing IV.3. This template could, instead of the update procedures for the user's actions that are available to both 'agent' and 'observer', have a generic procedure called 'newLoc(Obj,Loc)'. That procedure should remove any stale beliefs that pertain to the location of the object instantiating 'Obj', and inserting the new belief that this object is at the location 'Loc'. This procedure could be called by any agent based on this template; specifically, it could be called by the 'observer' itself for handling the event that either the 'agent' or the user changed the location of the marble. And, more importantly, the 'observer' could call this procedure *in its profile module* of 'agent', if it were convinced that the agent also received the event (i.e. 'perceived the action', e.g. because it was present when the action was performed). Naturally, if the 'observer' itself called the procedure then its own beliefs should be affected, if it called the procedure in the profile module then the beliefs stored in the profile should be. The functioning of a profile modules as sketched above, would, in our opinion, make it better suited for profiling, and is also closer to the notion of the PWB (Nichols & Stich, 2003).

4 Reflection

This chapter focuses on 'mindreading', in the sense that this term is used in psychological literature. In Section 1 the models of human mindreading put forward by Baron-Cohen (1995) and Nichols & Stich (2003) are discussed and compared, to lay the conceptual basis for subsequent sections and to provide the foundation for insights pertaining to formal approaches to mindreading. Section 2 presents such a formal approach, utilizing the formalism presented in Chapter III, extended slightly to allow for handling the case of multiple agents, to identify so-called 'mindreading patterns' in regard to the behavior of various agents. Those patterns concern the attribution of beliefs and goals to agents, and reflect regularities that can be used in shaping virtual mindreaders. This usage is illustrated by means of the example in Section 3 that presents a formal model of the classical Sally-Anne false-belief task, followed by an implementation in modular 2APL that takes inspiration from the same theme. In the next chapter a more elaborate and technical implementation is focused upon, specific to the work on mental state abduction presented in the earlier chapters of this dissertation.

Chapter V

Implementation Using ASP

This chapter presents an implementation of mental state abduction, as formalized in Chapter II, attempting to closely follow this theoretical exposition. This implementation is realized using answer set programming, a popular paradigm for nonmonotonic logic programming which is introduced in Section 1. In Section 2 it is shown how the implementation is derived in regard to the abductive approach of Chapter II. This approach is perhaps best considered a specification for implementation, as it shows how to derive the programming clauses from logical clauses in general, instead of presenting a particular program. Nevertheless, we prefer the term 'implementation' because of the formulation in terms of actual programming clauses, and in Section 3 evaluate such implementation analytically. Section 4 concludes this chapter with a brief reflection.

1 Answer Set Programming

In this section the paradigm of answer set programming (ASP) is introduced, building on the introductory texts of Lifschitz (2008) and Gelfond (2008), the different interpretations of stable model semantics discussed by Lifschitz (2010), and the manual of the set of ASP tools that are developed at the University of Potsdam (Gebser et al., 2010).[1] ASP is a language for knowledge representation and reasoning that is based on the stable model semantics of logic programming (Gelfond, 2008). The concept of stable model semantics has its roots in research on nonmonotonic reasoning, and can be given different characterizations. Those characterizations are of interest because they can provide new perspectives on the interpretation of ASP, even though they are equivalent in regard to 'traditional' (Prolog-style) ASP programs (Lifschitz, 2010). In this chapter we define a stable model in terms of the *reduct* of a logic program, which is the preferred definition of most textbook expositions on ASP (Gelfond, 2008; Lifschitz, 2010).

The ASP approach to (nonmonotonic) reasoning is an active research area, so that as a consequence the feature set of ASP developed in research, as well as the corresponding support from ASP interpreters, are constantly evolving. In our implementation of mental state abduction using ASP, as described in this chapter, we have made use of the features of ASP on which general consensus exists (Gelfond, 2008) as much as possible, deviating from this approach only if considered prudent (which will then be mentioned in the text). Within the frame of reference presented above, the typical programming methodology of ASP is now discussed, followed by the syntax and semantics of the ASP programming language.

[1] ASP syntax has a strong connection to the Prolog language for logic programming, so that the language of ASP is sometimes referred to as 'Answer Set Prolog' (Gelfond, 2008).

1.1 Programming Methodology

A distinguishing feature of ASP programming is its two-staged nature, which at first encounter can be confusing so that it seems pertinent to explain it here. This two-stagedness is reflected in the manner in which problem solving is approached by means of ASP; in the words of Ilkka Niemelä (2010): "*Encode the problem as a theory such that solutions to the problem are given by answer sets of the theory*". The two stages are *encoding* followed by *solving*, and, accordingly, ASP systems typically consist of two separate programs: a 'grounder' and a 'solver'. Encoding of an ASP theory is done using the grounder, which is essentially a pre-processor that generates low-level representations that are accepted by the solver, from high-level code that is human-readable and allows for more convenient representation. The solver outputs the answer sets that represent possible solutions, which can be interpreted as models for the theory.

Several tools exist that allow for instantiation of an ASP system as described above; here we mention only those that are employed in realization of the implementation put forward in this chapter, which are part of the Potassco collection of ASP tools developed at the University of Potsdam, Germany (Gebser et al., 2010).[2] The Potassco system consists of a grounder called GRINGO and a solver called CLASP, also available as a monolithic program and then referred to as CLINGO. The GRINGO/CLASP system is largely compatible with LPARSE/SMODELS, another popular combination of ASP grounder and solver (Niemelä et al., 2000).

It has been explained that ASP programming typically occurs with use of separate tools providing grounding/solving functionality, by suitably encoding a theory such that solutions (models) can be found. This two-stage methodology has been referred to by Lifschitz as "generate-and-test" (2008), and can be summarized as follows:

1. Using one type of rules (so-called 'choice rules'), a set of candidate solutions is *generated*, which is typically a superset of acceptable solutions.
2. Another type of rules ('constraint rules') is used for *testing* candidate solutions to see whether they qualify as acceptable solutions.

V.1

In what follows, the term *candidate answer set* is used to refer to a candidate solution produced in the 'generation' stage, and the term *actual answer set* is used to in reference to an acceptable solution; i.e. a candidate solution that has passed the 'test' stage. The syntax and semantics of ASP are now introduced, insofar as relevant to this chapter, to set the stage for presenting our implementation that follows the above methodology.

1.2 Syntax and Semantics

In this section the syntax and semantics of ASP are introduced, focusing specifically on constructs employed in our implementation; see Gelfond (2008); Gebser et al. (2010) for more general treatises.

[2] Potassco is used because it is an actively developed project, supporting standard ASP along with a wide variety of extensions, which has consistently been a top performer in recent ASP competitions.

1.2.1 Syntax

The main ASP programming construct are *rules*, which come in different types. Those rule types are introduced and discussed further on, but first it is pointed out that ASP literals, the main building block with which rules are constructed, differ substantially from those of Prolog. Specifically, the following types of ASP literals exist:

Basic literals: Basic literals of ASP take the form $p(t_1,\ldots,t_n)$ and $-p(t_1,\ldots,t_n)$, where p is an n-ary predicate symbol with $n \geq 0$, $t_1,\ldots,t_n$ are ASP terms (i.e. constants or variables), and '–' denotes classical (strong) negation so that $-p(t_1,\ldots,t_n)$ is the negative counterpart of the ASP atom $p(t_1,\ldots,t_n)$.

Extended literals: By means of default negation 'not' extended literals of the form 'not ψ' can be formed, given that ψ is a basic literal.

Choice literals: Choice literals have the form $l\{\psi_1,\ldots,\psi_n\}u$, where $l,u \in \mathbb{N}_1$ such that $l \leq u$, and $\psi_1,\ldots,\psi_n$ are basic literals. The numbers l and u are lower and upper bounds, respectively, and a choice literal of this form indicates that any number of elements out of $\{\psi_1,\ldots,\psi_n\}$, with a minimum of l and maximum of u, should be 'chosen' for interpretation.

Having described the types of ASP literals, the syntax of different kinds of ASP rules is given; semantics is discussed further on. It should hereby be noted that a generic domain of ASP literals is assumed, which in ensuing sections that focus on the actual implementation is fixed in relation to the language $\mathscr{L}_1$ of Chapter II.

Definition V.1 (ASP syntax — implicative rule)**.** *Let ϕ be a basic or choice literal, and let $\psi_1,\ldots,\psi_n$ be basic or extended literals. The syntax of implicative ASP rules then is as follows.*

$$\phi :- \psi_1,\ldots,\psi_n.$$

ASP rules have a *head* and a *body*, as in the case of Prolog clauses, such that for rules of the form given in Definition V.1 holds that ϕ is the head and $\psi_1,\ldots,\psi_n$ is the body. Rules that have both a head and a body are referred to here as *implicative rules*, because the interpretation of $\phi :- \psi_1,\ldots,\psi_n$ is similar in spirit to that of the logical implication $\psi_1 \wedge \ldots \wedge \psi_n \rightarrow \phi$. Rules that have no body are referred to as *facts* in discussions of ASP, and rules that have no head are referred to as *constraints* (Gelfond, 2008).

Definition V.2 (ASP syntax — fact rule)**.** *Let ϕ be a basic literal or a choice literal. The syntax of ASP fact rules then is as follows.*

$$\phi.$$

Definition V.3 (ASP syntax — constraint rule)**.** *Let $\psi_1,\ldots,\psi_n$ be basic or extended literals. The syntax of ASP constraint rules then is as follows.*

$$:- \psi_1,\ldots,\psi_n.$$

For our convenience, the term 'rule' is from now on reserved exclusively for referring to the type of rules defined in Definition V.1, the term 'fact' is used for referring to the type of rules defined in Definition V.2, and the term 'constraint' is used to refer to the rules of Definition V.3, hereby following Gebser et al. (2010).

Given the above introduction of syntax, the semantics of ASP can be discussed. Before doing so, however, two important aspects of ASP should be mentioned. First of all, it is the case that ASP programs can contain variables, whereas ASP semantics are typically given for ground programs (Gelfond, 2008; Lifschitz, 2010). This is not problematic, though, because generally (and for our approach, specifically) the program with variables can be considered to be simply a shorthand for all its ground instances, as produced by the grounder (GRINGO, in our case) in the pre-processing phase. Second, a similar point can be made in regard to choice literals of the form $l\{\psi_1,\ldots,\psi_n\}u$, namely that the semantics of a program with choice literals can be given in terms of multiple programs without choice literals.

Recall that $l, u \in \mathbb{N}_1$ represent lower and upper bounds on the number of basic literals that should be 'chosen' from the set $\{\psi_1,\ldots,\psi_n\}$. This choice can be informally described as "any of between l and u literals from $\{\psi_1,\ldots,\psi_n\}$ should be in the answer set (if the body is satisfied)", where the part in parentheses applies to implicative rules but not to facts. Choice literals are a powerful generative construct, in terms of ASP methodology as discussed in Section 1.1, as they can essentially expand the single program that employs this construct to multiple programs that do not. Consider, for example, the following choice rule (where $\psi_1,\ldots,\psi_n$ are basic literals).

$$l\{\psi_1,\ldots,\psi_m\}u \mathrel{:-} \psi_{m+1},\ldots,\psi_n.$$

For any (distinct) k literals $\{\psi_1,\ldots,\psi_k\} \subseteq \{\psi_1,\ldots,\psi_m\}$, where $l \leq k \leq u$, this choice rule represents the program which contains each of the following non-choice rules. Thus, if there exist multiple instantiations of k such that $l \leq k \leq u$, then there also exist multiple non-choice programs of the above kind.

V.1

$$\begin{gathered}\psi_1 \mathrel{:-} \psi_{m+1},\ldots,\psi_n.\\ \vdots\\ \psi_k \mathrel{:-} \psi_{m+1},\ldots,\psi_n.\end{gathered}$$

The single rule $l\{\psi_1,\ldots,\psi_m\}u \mathrel{:-} \psi_{m+1},\ldots,\psi_n$ thus captures, for every distinct number k such that $l \leq k \leq u$, a number of c distinct sets of rules (i.e. programs) such that c represents the number of distinct k-element subsets of $\{\psi_1,\ldots,\psi_m\}$. The number of distinct k-element subsets of a set with m elements is $\frac{m!}{k!(m-k)!}$, so that the single rule $l\{\psi_1,\ldots,\psi_m\}u \mathrel{:-} \psi_{m+1},\ldots,\psi_n$ is semantically equivalent to $\sum_{k=l}^{u}(\frac{m!}{k!(m-k)!})$ non-choice programs of the type described above. It occurs likewise with 'choice facts' (i.e. choice literals employed as facts). The answer sets of the program with choice literals are then given by the answer sets of the corresponding non-choice programs, in the sense that every answer set

of each of the non-choice programs is an answer set of the program with choice constructs, which has no further answer sets.[3]

The use of choice rules is perhaps best illustrated by means of an example, so consider the following program, which has three answer sets: $\{a, c\}$, $\{-b, c\}$, $\{a, -b, c\}$.

$$1\{a, -b\}2 :- c. \qquad\qquad 1\{c\}1.$$

This program can be 'flattened out' to the following three programs that do not contain choice rules, where each line represents a distinct program, which have the following respective answer sets: $\{a, c\}$, $\{-b, c\}$, $\{a, -b, c\}$.

$$\begin{array}{lll} a :- c. & & c. \\ -b :- c. & & c. \\ a :- c. & -b :- c. & c. \end{array}$$

Choice rules are thus a powerful notational device, and are in fact the preferred way to represent logical disjunction (Gebser et al., 2010), as will be done later in this chapter. Because choice rules can be ignored from point of view of semantics, given the way we use them, it is possible to now turn to the standard semantics of ASP.

1.2.2 Semantics

The most widely used definition of ASP semantics is that in terms of the *reduct* of (ground) logical programs, and the discussion presented here follows Lifschitz (2010) in this respect; for a more general and formal treatment see Gelfond (2008).

Definition V.4 (reduct)**.** *Let $\mathfrak{P}$ be a ground ASP program that contains only facts or rules, and no constraints. Given that $\psi_0, \ldots, \psi_n$ are basic literals occurring in $\mathfrak{P}$, the reduct $\mathfrak{P}^\Psi$ of $\mathfrak{P}$ with respect to a set of basic literals Ψ is then obtained by*

1. *Dropping each rule of the form '$\psi_0 :- \psi_1, \ldots, \psi_m, \mathsf{not}\ \psi_{m+1}, \ldots, \mathsf{not}\ \psi_n$' from $\mathfrak{P}$ if for any $\psi \in \{\psi_{m+1}, \ldots, \psi_n\}$ holds $\psi \in \Psi$.*
2. *Dropping '$\mathsf{not}\ \psi_{m+1}, \ldots, \mathsf{not}\ \psi_n$' from rules '$\psi :- \psi_1, \ldots, \psi_m, \mathsf{not}\ \psi_{m+1}, \ldots, \mathsf{not}\ \psi_n$' that remain in $\mathfrak{P}$ after application of the previous step.*

The reduct of a ground program thus yields a program that contains only rules without extended literals in the body. This reduct is defined relative to some set of basic literals, and it is this set of literals that plays a central role in the definition of answer sets given below. Note that in this definition the terminology 'Ψ is a model of $\mathfrak{P}$' is used (Lifschitz, 2008), where Ψ is a set of basic literals and $\mathfrak{P}$ is a program; this means that if $\mathfrak{P}$ were thought of as a set of logical clauses one could write $\Psi \models \mathfrak{P}$.

[3] The above explication of choice literals is not official and simplifies matters, but is accurate for the scope of this implementation. For a more general discussion see Gebser et al. (2010) on *aggregates*.

Definition V.5 (answer set)**.** *Let Ψ be a set of basic literals, let $\mathfrak{P}^+$ be a ground program with constraints, and $\mathfrak{P}$ the corresponding program without constraints. Furthermore, let $\mathfrak{P}^\Psi$ denote the reduct of $\mathfrak{P}$ with respect to Ψ, as defined in Definition V.4. It then holds that Ψ is an answer set of $\mathfrak{P}$ if and only if*

- *Ψ is a model of $\mathfrak{P}^\Psi$.*
- *No proper subset of Ψ is a model of $\mathfrak{P}^\Psi$.*

Furthermore, it holds that Ψ is an answer set of $\mathfrak{P}^+$ if and only if Ψ is an answer set of $\mathfrak{P}$, and

- *For every constraint '$:- \psi_1, \ldots, \psi_m, \mathsf{not}\ \psi_{m+1}, \ldots, \mathsf{not}\ \psi_n$' in $\mathfrak{P}^+$ it holds that none of $\psi_1, \ldots, \psi_m$ is in Ψ, and all of $\psi_{m+1}, \ldots, \psi_n$ are in Ψ.*

Definition V.5 is also known as the *stable model semantics* of logic programs, given in terms of the reduct, where the terms 'answer set' and 'stable model' are equivalent and used interchangeably in literature. In line with the view of answer sets as models, $S \models \phi$ can be written to denote that the answer set S satisfies the expression ϕ. It should be noted that an ASP program can have more than one answer set, and that it is called *consistent* if it has at least one (Gelfond, 2008).

Definition V.6 (consistent ASP program)**.** *Let $\mathfrak{P}$ be an ASP program. $\mathfrak{P}$ is* consistent *if it has at least one answer set.*

The discussion presented here is by necessity somewhat brief and superficial; see Lifschitz (2008) for a brief introduction to ASP in general and Gelfond (2008) for a more extended one, Lifschitz (2010) for a multitude of interpretations of the concept 'stable model', and Gebser et al. (2010) for a primer on the Potassco set of ASP tools.

2 Implementation

V.2

The syntax and semantics of the answer set programming language (ASP) have been introduced and discussed in the previous section, and in the current section this language is employed for implementation of our approach to mental state abduction. The introduction of the previous section focused mostly on the technical interpretation of ASP, not so much on its conceptual interpretation. Little will also be said of that in this section, and instead the reader is referred to the literature (Gelfond, 2008; Lifschitz, 2008, 2010; Gebser et al., 2010). Nevertheless, it seems pertinent to point out that ASP is well-suited for nonmonotonic reasoning, seen in the fact that it allows for reasoning with theories that have mutually exclusive answer sets (extensions). This suitability is exploited in the current section, as will be shown.

2.1 Encoding the Abductive Theory

In Section 3.4 of Chapter II — specifically, Definitions II.9, II.12, and II.13 — we presented sets of clauses that laid the foundation of background theories for reasoning about the behavior and mental state of BDI-based agents, programmed in the MYAPL language, under

different perceptory conditions. In Section 3.5 of that same chapter the abducibles and observables were defined, and in Section 3.6 it was shown these ingredients can be put to use in abductive reasoning about agents' observed behavior. Recall that the language $\mathscr{L}_1$, defined in Section 3.4 of Chapter II, was the language of choice for representing the corpora used in abductive reasoning, formulated in terms of ground subsets of that language. This makes it quite straightforward to encode the abductive theory as a ground ASP program, and we start in doing so with the background theories.

2.1.1 Encoding the Background Theories

In Definitions II.9, II.12, and II.13 sets of clauses pertaining to complete, late, and partial observability of the actions of MYAPL agents were defined. Those clauses take the form of implications '$(\mathsf{rule}(i) \wedge \mathsf{seq}(j)) \rightarrow (\mathsf{obs}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{obs}(\alpha_n, n))$' that describe the actions that are observable if it is the case that the agent applied the rule i and the observer expects to see sequence j of the corresponding plan. It is our intention to encode those clauses as an ASP program; however, it should be observed that implications of the language $\mathscr{L}_1$ can in general not be 'reversed' without some sort of translation. There are different ways to go about making such a translation: in earlier work (Sindlar et al., 2011) we opted to encode the consequence of aforementioned implications by means of a single ASP term, and to 'decompose' this in the answer set (using variables). This approach, although convenient from a representational point of view, is not so efficient computationally. For this reason an alternative approach is taken here, which encodes the logical representations in ASP without need for decomposition.

It is here assumed that the premises and consequences of the $\mathscr{L}_1$-implications that are to be encoded in ASP have been translated to disjunctive normal form (DNF); i.e. that they are represented as disjunctions of conjunctions of literals. Given the set Lit of $\mathscr{L}_1$-literals, the DNF formula $\phi \in \mathscr{L}_1$ can then be represented as follows.

$$\begin{aligned} \phi = C_1 \vee \ldots \vee C_i, \textit{ where, given } \psi_1, \ldots, \psi_n \in \mathsf{Lit}, \\ C_1 = \psi_1 \wedge \ldots \wedge \psi_k \\ \vdots \\ C_i = \psi_m \wedge \ldots \wedge \psi_n \end{aligned}$$

The assumption that the preconditions and consequences of to-be-encoded implications are in DNF can be safely made, as every (ground) expression of $\mathscr{L}_1$ has a logically equivalent counterpart in DNF (Barwise & Etchemendy, 1992). Also note that this representational desideratum already applies to the clauses given by Definitions II.9, II.12, and II.13, because the premises and consequences of those implications contain only conjunction so that they are in DNF. Thus, the following encoding can be made, where it should be noted that ASP encodings (i.e. programs) are simply sets of ASP rules; to avoid any confusion, the '.' at the end of rules is retained.

Definition V.7 (encoding of observability)**.** *Let $C_{\mathscr{R}}$, $L_{\mathscr{R}}$, and $P_{\mathscr{R}}$ be the sets of observability clauses given by Definitions II.9, II.12, and II.13, respectively, and let $\mathfrak{P}^C_{\mathscr{R}}$, $\mathfrak{P}^L_{\mathscr{R}}$, and $\mathfrak{P}^P_{\mathscr{R}}$ be*

their respective ASP encodings, defined as follows.

$$\forall X \in \{C, L, P\}:$$
$$\mathfrak{P}^{X}_{\mathscr{R}} = \{\, n\{\mathsf{obs}(\alpha_1, 1), \ldots, \mathsf{obs}(\alpha_n, n))\}n :- \mathsf{rule}(i), \mathsf{seq}(j). \;\mid\; ((\mathsf{rule}(i) \wedge \mathsf{seq}(j)) \rightarrow (\mathsf{obs}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{obs}(\alpha_n, n))) \in X_{\mathscr{R}} \,\}$$

The above definition states that for each logical implication that describes the observability of actions, there should be a corresponding implicative rule in the ASP encoding. Informally, such rules state that *all* implied instances of obs/2 should be in the answer set if the preconditions are satisfied, and thus capture the meaning of the observability clauses of Definitions II.9, II.12, and II.13.

Apart from describing observability of actions, the background theories of Chapter II are founded on a set of clauses, defined in Definition II.15, that describe the beliefs and goals which the observed agent must have had if it applied particular rules. Observe that the postconditions of those implications are not necessarily in DNF, so that translation might be required. In regard to the encoding it should be noted that the ASP syntax does not allow nesting of the choice construct, meaning that within the brackets of '$l\{\ldots\}u$' only basic literals can occur, and not choice literals. To properly translate DNF expressions using our approach a function is therefore employed that assigns ASP atoms to expressions of $\mathscr{L}_1$, serving in the encoding as 'placeholders' for larger expressions.

Definition V.8 (placeholder function ι_{ASP})**.** *Let $\mathscr{L}_1$ be the language defined in Definition II.8, and A a set of ASP placeholders atoms. The (bijective) function $\iota_{\text{ASP}} : \mathscr{L}_1 \longrightarrow A$ then assigns placeholder atoms to expressions of $\mathscr{L}_1$.*

It is noteworthy that ASP provides two kinds of negation, 'classical' and 'default', the former of which is here employed to represent $\mathscr{L}_1$-atoms of the form $\mathsf{belief}(-p)$ and $\mathsf{goal}(-p)$. To this extent the translation function τ_{ASP} is defined as follows.

V.2

Definition V.9 (translation function τ_{ASP})**.** *The function τ_{ASP} encodes instances of the $\mathscr{L}_1$-predicates* belief/1 *and* goal/1 *in ASP, as follows.*

$$\tau_{\text{ASP}}(\mathsf{belief}(p)) = \mathsf{belief}(p) \qquad \tau_{\text{ASP}}(\mathsf{goal}(p)) = \mathsf{goal}(p)$$
$$\tau_{\text{ASP}}(\mathsf{belief}(-p)) = \text{-}\mathsf{belief}(p) \qquad \tau_{\text{ASP}}(\mathsf{goal}(-p)) = \text{-}\mathsf{goal}(p)$$

For notational convenience, ψ^{τ} is used in the following as shorthand for $\tau_{\text{ASP}}(\psi)$. Given this translation function and the placeholder function ι_{ASP} it is possible to now present the following ASP encoding of the clauses pertaining to mentalistic preconditions, which ensures that if the DNF-equivalent formula of the description of those preconditions is a conjunction then encoding takes place directly by means of a choice literal representing that conjunction, and that placeholders atoms are used to represent disjuncts if this DNF-equivalent is a disjunction. It may be noted that no mention is made of how DNF-equivalents are to be obtained, but given that this is a fairly standard operation it is here simply assumed that this is done efficiently. Furthermore, observe that the interpretation of disjuncts is, on grounds of the encoding below, 'maximal', in the sense that answer sets

(models) are computed for every satisfiable disjunct; a minimal encoding would, on the other hand, consider satisfaction of only single disjuncts. Put differently: the encoding below considers *all* models of disjunctive clauses in the observer's theory about the observed agent's mental state, not just the minimal models.

Definition V.10 (encoding of mentalistic preconditions)**.** *Let $M_{\mathscr{R}}$ be the set of clauses concerning mentalistic preconditions given by Definition II.15, and let $\phi^{\text{DNF}} \in \mathscr{L}_1$ denote a DNF-equivalent of $\phi \in \mathscr{L}_1$. $\mathfrak{P}^M_{\mathscr{R}}$ is then the ASP encoding of $M_{\mathscr{R}}$, defined as follows.*

$$
\begin{aligned}
\mathfrak{P}^M_{\mathscr{R}} = \{\ & 1\{\iota_{\text{ASP}}(\phi_1),\ldots,\iota_{\text{ASP}}(\phi_m)\}m \text{ :- } \mathsf{rule}(i).\ , \\
& x\{\psi^{\tau}_{1,1},\ldots,\psi^{\tau}_{1,x}\}x \text{ :- } \iota_{\text{ASP}}(\phi_1). \\
& \qquad\ldots, \\
& y\{\psi^{\tau}_{m,1},\ldots,\psi^{\tau}_{m,y}\}y \text{ :- } \iota_{\text{ASP}}(\phi_m). \quad \mid \\
& \quad (\mathsf{rule}(i) \rightarrow \phi) \in M_{\mathscr{R}} \ \ \&\ \ \phi^{\text{DNF}} = (\phi_1 \vee \ldots \vee \phi_m) \ \ \& \\
& \qquad \phi_1 = (\psi_{1,1} \wedge \ldots \wedge \phi_{1,x}) \ \ \&\ \ \ldots \ \ \&\ \ \phi_m = (\psi_{m,1} \wedge \ldots \wedge \phi_{m,y})\ \}
\end{aligned}
$$

The encoding of the clauses of Definition II.15 by means of Definition V.10 may seem somewhat circuitous due to its lengthy definition, but it is actually quite straightforward. This can be illustrated in an example, as follows, which also shows how the encoding of observability occurs.

Example V.1. *Let the following* MyAPL *program be the element of the singleton rule-set $\mathscr{R}$.*

$$1 : p \ \text{<-}\ q \ \text{or} \ \text{-}r \ \mid \ a; b$$

The complete observability clauses $C_{\mathscr{R}}$ and mentalistic precondition clauses $M_{\mathscr{R}}$, as obtained by Definitions II.9 and II.15, respectively, then are as follows; it is hereby assumed that $\iota_{\delta}(a; b) = 1$.

$$
\begin{aligned}
C_{\mathscr{R}} &= \{(\mathsf{rule}(1) \wedge \mathsf{seq}(1)) \rightarrow (\mathsf{obs}(a,1) \wedge \mathsf{obs}(b,2))\} \\
M_{\mathscr{R}} &= \{\mathsf{rule}(1) \rightarrow (\mathsf{goal}(p) \wedge (\mathsf{belief}(q) \vee \mathsf{belief}(\text{-}r)))\}
\end{aligned}
$$

It should be observed that the consequence '$\mathsf{goal}(p) \wedge (\mathsf{belief}(q) \vee \mathsf{belief}(\text{-}r))$' of the implication in $M_{\mathscr{R}}$ is not in DNF, as it is a top-level conjunction that contains a disjunction. By distributing $\wedge$ over $\vee$ it is brought to DNF, yielding

$$
\begin{aligned}
&(\mathsf{goal}(p) \wedge (\mathsf{belief}(q) \vee \mathsf{belief}(\text{-}r)))^{\text{DNF}} = \\
&\qquad (\mathsf{goal}(p) \wedge \mathsf{belief}(q)) \vee (\mathsf{goal}(p) \wedge \mathsf{belief}(\text{-}r))
\end{aligned}
$$

Assume in regard to assignment of placeholders that $\iota_{\text{ASP}}(\mathsf{goal}(p) \wedge \mathsf{belief}(q)) = d_1$ and $\iota_{\text{ASP}}(\mathsf{goal}(p) \wedge \mathsf{belief}(\text{-}r)) = d_2$. The encodings $\mathfrak{P}^C_{\mathscr{R}}$ and $\mathfrak{P}^M_{\mathscr{R}}$ of $C_{\mathscr{R}}$ and $M_{\mathscr{R}}$, respectively, along the lines of

Definitions V.7 and V.10, then are as follows.

$$\mathfrak{P}^{C}_{\mathscr{R}} = \{\ 2\{\mathsf{obs}(a,1) \wedge \mathsf{obs}(b,2)\}2 :- \ \mathsf{rule}(1), \mathsf{seq}(1).\ \}$$

$$\begin{aligned}\mathfrak{P}^{M}_{\mathscr{R}} = \{\ & 1\{d_1, d_2\}2 :- \ \mathsf{rule}(1).\ , \\ & 2\{\mathsf{goal}(p), \mathsf{belief}(q)\}2 :- \ d_1.\ , \\ & 2\{\mathsf{goal}(p), \text{–}\mathsf{belief}(r)\}2 :- \ d_2.\ \}\end{aligned}$$

Thus, as the above example shows, an encoding in ASP of the logical clauses pertaining to observability and mentalistic preconditions can be obtained quite straightforwardly. In Section 3.4.2 of Chapter II a third set of clauses was furthermore specified, pertaining to the deliberation of the agent, but, as will be seen in Section 2.1.3 of the present chapter, it is not necessary to encode those clauses (as such) in ASP because of the way selection of abducibles is encoded. It is, however, necessary to encode the observation of actions, and this is done in the next section.

2.1.2 Encoding Observations

It was put forward in Section 3.5.2 that the observables considered for explanation by the abductive theory are conjunctions of consecutively numbered instances of seen/2 (cf. Definition II.18). As explained in that section, the observables as such could not be accounted for by the theory, which only accounts for instances of obs/2. Nevertheless, it was opted for to represent observables as instances of seen/2, as the interpretation of this predicate makes it conceptually more appropriate for modeling actual observations than the predicate obs/2; the relation between actual observations and facts the theory can account for was established by means of a translation (cf. Definition II.19).

It turns out that the aforementioned representation of observables is convenient in regard to our implementation, as will be seen in the next section. Those observables are elements of the domain $\Omega_{\mathscr{R}}$ that consists of consecutively numbered instances of the predicate seen/2, and encoding of actual observations from the language is done by means of a single choice fact that captures those instances, as follows.

Definition V.11 (encoding of observations)**.** *Let $\Omega_{\mathscr{R}}$ be the language of observables defined in Definition II.18, and $\mathfrak{O}_{\omega}$ the ASP encoding of the observation $\omega \in \Omega_{\mathscr{R}}$, defined as follows.*

$$\mathfrak{O}_{\omega} = \{\ n\{\mathsf{seen}(\alpha_1, 1), \ldots, \mathsf{seen}(\alpha_n, n)\}n. \quad \mid \omega = \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_n, n)\ \}$$

Note that this encoding of conjoined facts by means of choice literals is in line with the encoding of conjunction in rule preconditions, as given in Section 2.1.1. Equivalently, the single choice literal '$n\{\ldots\}n$' can be replaced by the corresponding n instances of seen/2, as explained in Section 1 in regard to the interpretation of choice literals.

2.1.3 Encoding Abductive Explanation

So far, fragments of the implementation have been presented that encode the observability of an agent's actions and corresponding mental state descriptions in regard to the rule it

applied and the plan sequence it selected (Section 2.1.1), as well as fragments that encode actual observations (Section 2.1.2). It now remains to encode the fragment of the implementation that pertains to explanation of observations. In line with ASP methodology, as described in Section 1.1, this is done in generate-and-test fashion, as follows.

1. Candidate answer sets are *generated* by considering rules that could have been applied and sequences that could be seen as a result thereof.
2. It is *tested* whether candidate answer sets account for the observed actions, and those that do not are discarded.

In order to implement the first step of generating candidate answer sets, two choice literals are used that encode the consideration of valid instances of rule/1 and seq/1. It is hereby assumed that the number of rules and sequences which are considered is fixed in advance. Note that this assumption is straightforward in regard to rules but not so straightforward in regard to sequences, of which there can be a countably infinite amount if plans contain iteration. However, given that the number of actually observed actions is always known, this need not be problematic, because then in any case should those sequences be considered which account for a number of observable actions that equals the number of actually observed actions. The second step is implemented by means of a constraint, that ensures that any candidate answer sets which contain instances of seen/2 but not the corresponding instances of obs/2, are discarded. In regard to notation it should furthermore be remarked that '$p(X) : r(X)$' is a so-called *conditional* (Gebser et al., 2010) stating that instances of $p/1$ should be considered by instantiating the argument to $p/1$ in relation to its instantiation in instances of $r/1$.[4] Thus, the following encoding of abductive explanation can be given.

Definition V.12 (encoding of abductive explanation)**.** *Let $R \subseteq \mathbb{N}_1$ be the identifiers of rules considered, and $S \subseteq \mathbb{N}_1$ those of sequences considered. $\mathfrak{A}_{R,S}$ is then the ASP encoding of abductive explanation, as follows, given that $R = \{r_1, \ldots, r_m\}$ and $S = \{s_1, \ldots, s_n\}$.*

$$\begin{aligned}\mathfrak{A}_{R,S} = \{\ & \mathsf{poss\_rule}(r_1).\ , \ldots , \mathsf{poss\_rule}(r_m).\ , \\ & \mathsf{poss\_seq}(s_1).\ , \ldots , \mathsf{poss\_seq}(s_n).\ , \\ & 1\{\mathsf{rule}(R) : \mathsf{poss\_rule}(R)\}1.\ , \\ & 1\{\mathsf{seq}(S) : \mathsf{poss\_seq}(S)\}1.\ , \\ & :-\ \mathsf{seen}(A,P),\ \mathsf{not}\ \mathsf{obs}(A,P).\ \}\end{aligned}$$

The predicates poss_rule/1 and poss_seq/1 encode the identifiers of rules and sequences, respectively, which considered as possibilities for instantiating rule/1 and seq/1. By means of the choice constructs $1\{\ldots\}1$ it is then ensured that, per candidate answer set, combinations of a single instance of rule/1 and a single instance of seq/1 (as derived from poss_rule/1 and poss_seq/1) are considered, thus implicitly reflecting the deliberation clauses employed in Chapter II (cf. Definition II.16).

[4] Use of conditionals in ASP is comparable to set comprehension in mathematics (by means of ' | ').

```
1:  paper_read        <-  paper_on_shelf  |
    { goto_shelf; pickup_paper; goto_chair; sit; read }

2:  bug_squashed      <-  bug_on_table and paper_on_shelf     |
    { goto_shelf; pickup_paper; goto_table; squash_bug }

3:  flowers_arranged  <-  flowers_on_table and vase_on_shelf  |
    { goto_shelf; pickup_vase; goto_table; arrange_flowers }
```

Listing V.1

2.2 Example

In this example the setting of Section 3.7 of Chapter II is revisited, and the relevant MYAPL rules are repeated in Listing V.1. Also, for brevity the variables employed in that chapter are used once again, which were as follows with regard to actions.

$$\begin{aligned} \alpha_1 &= \mathsf{goto\_shelf} & \alpha_6 &= \mathsf{goto\_table} \\ \alpha_2 &= \mathsf{pickup\_paper} & \alpha_7 &= \mathsf{squash\_bug} \\ \alpha_3 &= \mathsf{goto\_chair} & \alpha_8 &= \mathsf{pickup\_vase} \\ \alpha_4 &= \mathsf{sit} & \alpha_9 &= \mathsf{arrange\_flowers} \\ \alpha_5 &= \mathsf{read} \end{aligned}$$

Recall that the numeration of those variables determines the numerical identifiers of the corresponding action in sequences, such that e.g. 'seq(123)' refers to the identifier of the sequence 'goto_shelf; pickup_paper; goto_chair'. Furthermore, recall that the propositional variables were as follows.

$$\begin{aligned} p_1 &= \mathsf{paper\_read} & p_5 &= \mathsf{flowers\_arranged} \\ p_2 &= \mathsf{paper\_on\_shelf} & p_6 &= \mathsf{flowers\_on\_table} \\ p_3 &= \mathsf{bug\_squashed} & p_7 &= \mathsf{vase\_on\_shelf} \\ p_4 &= \mathsf{bug\_on\_table} \end{aligned}$$

Because the cases of incomplete observation have earlier been discussed in regard to the setting of this example, for brevity only the case of complete observation is discussed here; by comparison with the occurrence of this example in Section 3.7 in Chapter II the cases of incomplete observation should be straightforward.

Given that $\mathscr{R}$ is the set of rules in Listing V.1, the ASP encoding of the background theory is as follows; where $\mathfrak{P}^C_{\mathscr{R}}$ and $\mathfrak{P}^M_{\mathscr{R}}$ are defined as in Definitions V.7 and V.10, respectively. Note that below programs are presented as lists of rules, instead of using formal set notation. In order to minimize clutter, the placeholders given by ι_{ASP}, as defined in Definition V.8, have been omitted; this is possible here seeing that the clauses of $M_{\mathscr{R}}$, underpinning $\mathfrak{P}^M_{\mathscr{R}}$, are already in DNF because they contain only a single disjunct consisting

of conjoined literals, which can be implemented straightforwardly without the need for placeholders.

$$\mathfrak{P}^{C}_{\mathscr{R}} = \left\{ \begin{array}{l} 5\{\mathsf{obs}(\alpha_1,1),\mathsf{obs}(\alpha_2,2),\mathsf{obs}(\alpha_3,3),\mathsf{obs}(\alpha_4,4),\mathsf{obs}(\alpha_5,5)\}5 \mathrel{:-} \mathsf{rule}(1),\mathsf{seq}(12345). \\ 4\{\mathsf{obs}(\alpha_1,1),\mathsf{obs}(\alpha_2,2),\mathsf{obs}(\alpha_6,3),\mathsf{obs}(\alpha_7,4)\}4 \mathrel{:-} \mathsf{rule}(2),\mathsf{seq}(1267). \\ 4\{\mathsf{obs}(\alpha_1,1),\mathsf{obs}(\alpha_8,2),\mathsf{obs}(\alpha_6,3),\mathsf{obs}(\alpha_9,4)\}4 \mathrel{:-} \mathsf{rule}(3),\mathsf{seq}(1869). \end{array} \right.$$

$$\mathfrak{P}^{M}_{\mathscr{R}} = \left\{ \begin{array}{l} 2\{\mathsf{goal}(p_1),\mathsf{belief}(p_2)\}2 \mathrel{:-} \mathsf{rule}(1). \\ 3\{\mathsf{goal}(p_3),\mathsf{belief}(p_4),\mathsf{belief}(p_2)\}3 \mathrel{:-} \mathsf{rule}(2). \\ 3\{\mathsf{goal}(p_5),\mathsf{belief}(p_6),\mathsf{belief}(p_7)\}3 \mathrel{:-} \mathsf{rule}(3). \end{array} \right.$$

Assume that the first observation is the action α_1 so that, in line with Definition V.11, $\mathfrak{O}_{\omega_1}$ is as follows, given $\omega_1 = \mathsf{seen}(\alpha_1,1)$.

$$\mathfrak{O}_{\omega_1} = \quad 1\{\mathsf{seen}(\alpha_1,1)\}1.$$

Furthermore, assume that all rules and sequences are considered as candidate hypotheses in explaining this observation action, so that, given the identifiers $R = \{1,2,3\}$ and $S = \{12345, 1267, 1869\}$, the encoding of abductive explanation is as follows; noting that '$p(a;b)$' is supported by CLASP as shorthand for '$p(a)$', '$p(b)$'.

$$\mathfrak{A}_{R,S} = \left\{ \begin{array}{l} \mathsf{poss\_rule}(1;2;3). \\ \mathsf{poss\_seq}(12345;1267;1869). \\ 1\{\mathsf{rule}(R) : \mathsf{poss\_rule}(R)\}1. \\ 1\{\mathsf{seq}(S) : \mathsf{poss\_seq}(S)\}1. \\ \mathrel{:-} \mathsf{seen}(A,P),\ \mathsf{not}\ \mathsf{obs}(A,P). \end{array} \right.$$

We have implemented this example by means of the grounder GRINGO and the solver CLASP. The output of CLASP is depicted in Figure V.1 for the abductive program $\mathfrak{P}^{C}_{\mathscr{R}} \cup \mathfrak{P}^{M}_{\mathscr{R}} \cup \mathfrak{O}_{\omega_1} \cup \mathfrak{A}_{R,S}$, which, as expected given that the plans of all three rules account for the observed action, has three answer sets (as seen in the screenshot).

Let the second observed action be α_2, and take $\omega_2 = \mathsf{seen}(\alpha_1,1) \wedge \mathsf{seen}(\alpha_2,2)$ accordingly, so that its ASP encoding is as follows.

$$\mathfrak{O}_{\omega_2} = \quad 2\{\mathsf{seen}(\alpha_1,1),\mathsf{seen}(\alpha_2,2)\}2.$$

In explaining this observation all rules and sequences are again considered, so that the program $\mathfrak{P}^{C}_{\mathscr{R}} \cup \mathfrak{P}^{M}_{\mathscr{R}} \cup \mathfrak{O}_{\omega_2} \cup \mathfrak{A}_{R,S}$ then has the answer sets depicted in Figure V.2. As expected, this program only has two answer sets: one reflecting the explanation that the agent applied rule 1, the other reflecting the explanation that it applied rule 2. It was shown in Corollary II.2 that explanation is monotonic with respect to single novel observations, and our implementation conveniently allows for utilization of this property by basing the encoding of abductive explanation, as used in explanation of novel actions, on explanation of hitherto observed actions. Note that such an approach does not change the solutions, it can

```
gringo C.lp M.lp o1.lp abdRS.lp | clasp 0

clasp version 2.0.1
Reading from stdin
Solving...
Answer: 1
poss_rule(1) poss_rule(3) poss_rule(2) poss_seq(12345) poss_seq(1869) poss_seq(1267) rule(1)
seq(12345) obs(a1,1) obs(a2,2) obs(a3,3) obs(a4,4) obs(a5,5) goal(p1) belief(p2) seen(a1,1)
Answer: 2
poss_rule(1) poss_rule(3) poss_rule(2) poss_seq(12345) poss_seq(1869) poss_seq(1267) rule(3)
seq(1869) obs(a1,1) obs(a6,3) obs(a8,2) obs(a9,4) goal(p5) belief(p6) belief(p7) seen(a1,1)
Answer: 3
poss_rule(1) poss_rule(3) poss_rule(2) poss_seq(12345) poss_seq(1869) poss_seq(1267) rule(2)
seq(1267) obs(a1,1) obs(a2,2) obs(a6,3) obs(a7,4) belief(p2) goal(p3) belief(p4) seen(a1,1)
SATISFIABLE

Models       : 3
Time         : 0.001s (Solving: 0.00s 1st Model: 0.00s Unsat: 0.00s)
CPU Time     : 0.000s
```

Figure V.1: Screenshot of CLASP output, given input to GRINGO of the program $\mathfrak{P}^{C}_{\mathcal{R}} \cup \mathfrak{P}^{M}_{\mathcal{R}} \cup \mathfrak{O}_{\omega_1} \cup \mathfrak{A}_{R,S}$, where `C.lp`, `M.lp`, `o1.lp`, and `abdRS.lp` are the corresponding encodings, and the '0' argument to CLASP states that all answer sets should be produced.

```
gringo C.lp M.lp o2.lp abdRS.lp | clasp 0

clasp version 2.0.1
Reading from stdin
Solving...
Answer: 1
poss_rule(1) poss_rule(3) poss_rule(2) poss_seq(12345) poss_seq(1869) poss_seq(1267) rule(2)
 seq(1267) obs(a1,1) obs(a2,2) obs(a6,3) obs(a7,4) belief(p2) goal(p3) belief(p4) seen(a1,1)
 seen(a2,2)
Answer: 2
poss_rule(1) poss_rule(3) poss_rule(2) poss_seq(12345) poss_seq(1869) poss_seq(1267) rule(1)
 seq(12345) obs(a1,1) obs(a2,2) obs(a3,3) obs(a4,4) obs(a5,5) goal(p1) belief(p2) seen(a1,1)
 seen(a2,2)
SATISFIABLE

Models       : 2
Time         : 0.001s (Solving: 0.00s 1st Model: 0.00s Unsat: 0.00s)
CPU Time     : 0.000s
```

Figure V.2: Screenshot of CLASP output for the program $\mathfrak{P}^{C}_{\mathcal{R}} \cup \mathfrak{P}^{M}_{\mathcal{R}} \cup \mathfrak{O}_{\omega_2} \cup \mathfrak{A}_{R,S}$.

only speed up the process of finding them. In regard to our example, this means that after having observed ω_2 only rules 1 and 2 need to be considered, along with their respective sequences. This is encoded as follows in $\mathfrak{A}_{R',S'}$, where $R' = \{1,2\}$ and $S' = \{12345, 1267\}$.

$$\mathfrak{A}_{R',S'} = \begin{cases} \mathsf{poss\_rule}(1;2). \\ \mathsf{poss\_seq}(12345;1267). \\ 1\{\mathsf{rule}(R) : \mathsf{poss\_rule}(R)\}1. \\ 1\{\mathsf{seq}(S) : \mathsf{poss\_seq}(S)\}1. \\ :- \ \mathsf{seen}(A,P), \ \mathsf{not}\ \mathsf{obs}(A,P). \end{cases}$$

Observation of action α_6 as the third action is reflected in the following ASP encoding.

$$\mathfrak{O}_{\omega_3} = \quad 3\{\mathsf{seen}(\alpha_1,1), \mathsf{seen}(\alpha_2,2), \mathsf{seen}(\alpha_6,3)\}3.$$

Accordingly, the program $\mathfrak{P}^C_{\mathfrak{R}} \cup \mathfrak{P}^M_{\mathfrak{R}} \cup \mathfrak{O}_{\omega_3} \cup \mathfrak{A}_{R',S'}$ has the answer sets depicted in Figure V.3. The example presented in this section is by necessity brief, but it should provide an idea as to how our implementation works. In the next section a more formal evaluation is given, in relation to the abductive theory of Chapter II.

3 Evaluation and Reflection

Given the approach presented in Section 2 to implementing the abductive theory of Chapter II, one might wonder whether this implementation does justice to its theoretical foundation. In this section it is shown that this is so, thus providing an evaluation of the implementation in terms of its correctness. Empirical evaluation in terms of, for example, time required to find solutions, is not given here. However, given the fact that ASP is suitable for handling large-scale problems Gelfond (2008), and that agent-based programming employs a knowledge-level representation so that useful agent behavior may be generated using relatively simple programs, our expectations in regard to performance are optimistic. In any case, focus in this section lies on formal evaluation.


```
gringo C.lp M.lp o3.lp abdRSprime.lp | clasp 0

clasp version 2.0.1
Reading from stdin
Solving...
Answer: 1
poss_rule(1) poss_rule(2) poss_seq(12345) poss_seq(1267) rule(2) seq(1267) obs(a1,1) obs(a2,2)
obs(a6,3) obs(a7,4) belief(p2) goal(p3) belief(p4) seen(a1,1) seen(a2,2) seen(a6,3)
SATISFIABLE

Models       : 1
Time         : 0.001s (Solving: 0.00s 1st Model: 0.00s Unsat: 0.00s)
CPU Time     : 0.000s
```

Figure V.3: Screenshot of CLASP output for the program $\mathfrak{P}^C_{\mathfrak{R}} \cup \mathfrak{P}^M_{\mathfrak{R}} \cup \mathfrak{O}_{\omega_3} \cup \mathfrak{A}_{R',S'}$.

3.1 Correspondence

The correspondence between the abductive approach of Chapter II and its implementation in the present chapter can be shown in terms of a correspondence between skeptical and credulous abductive inference, on the one hand, and the computation of *cautious* and *brave* consequences (Gebser et al., 2010), on the other. Skeptical and credulous abduction were defined in Definition II.24 in terms of propositions that are satisfied by all (skeptical), or at least one (credulous), of the abductive extensions; given that those exist. Cautious and brave consequences have the following definition:

Definition V.13 (cautious and brave consequences)**.** *Let $\mathfrak{P}$ be an ASP program, and $S_1, \ldots, S_n$ its answer sets.*

- *The basic literal ψ is a* cautious *consequence of $\mathfrak{P}$ if and only if $\psi \in \bigcap\{S_1, \ldots, S_n\}$.*
- *The basic literal ψ is a* brave *consequence of $\mathfrak{P}$ if and only if $\psi \in \bigcup\{S_1, \ldots, S_n\}$.*

As stated earlier, because answer sets are simply sets of literals they can be regarded as models, writing $S \models \psi$ to denote that the basic literal ψ is satisfied by the answer set S (Gelfond, 2008). This view applies to the notions of cautious and brave consequences as well, such that $\mathfrak{P} \models_{\mathsf{c}} \psi$ can be written to denote that the basic literal ψ is cautiously entailed by the program $\mathfrak{P}$ (i.e. is satisfied by all answer sets of $\mathfrak{P}$), and $\mathfrak{P} \models_{\mathsf{b}} \psi$ to denote that the basic literal ψ is bravely entailed by program $\mathfrak{P}$ (i.e. is satisfied by some answer set of $\mathfrak{P}$). And, by broadening this view somewhat, this notion of entailment can be extended to more complex logical expressions, stating, for example, $\mathfrak{P} \models_{\mathsf{c}} \phi \vee \phi'$ to denote that all answer sets of $\mathfrak{P}$ satisfy $\phi \vee \phi'$.

Given this notion of entailment by ASP programs, a correspondence can be shown to hold between skeptical/credulous abduction on grounds of an abductive theory, and cautious/brave entailment by the program encoding that theory.

Lemma V.1. *Let $\mathscr{R}$ be a set of* MyAPL *rules, $\Lambda_{\mathscr{R}}$ the corresponding abductive theory as defined in Definition II.21, $\mathscr{A}_{\mathscr{R}}$ the abducibles as defined in Definition II.17, $\Omega_{\mathscr{R}}$ the language of observables as defined in Definition II.18, and $\Gamma^{X}_{\mathscr{R}}$ the extension operator of Definition II.23 for $X \in \{C, L, P\}$ operating on $\omega \in \Omega_{\mathscr{R}}$. Furthermore, let $\mathfrak{P}^{X}_{\mathscr{R}}$ be the ASP encoding of the background theory as defined in Definition V.7 for $X \in \{C, L, P\}$, $\mathfrak{P}^{M}_{\mathscr{R}}$ the ASP encoding of preconditions as defined in Definition V.10, $\mathfrak{O}_{\omega}$ the ASP encoding of $\omega \in \Omega_{\mathscr{R}}$ as defined in Definition V.11, and $\mathfrak{A}$ the ASP encoding of abduction as defined in Definition V.12. It then holds, assuming that $\mathfrak{A}$ (at least) encompasses the identifiers of rules and sequences that account for ω, that*

$$\forall X \in \{C, L, P\}\ \forall \omega \in \Omega_{\mathscr{R}}\ \forall \psi \in \mathscr{A}_{\mathscr{R}}:$$
$$\exists \Phi \in \Gamma^{X}_{\mathscr{R}}(\omega): \quad \Phi \models \psi$$
$$\Longleftrightarrow$$
The program $\mathfrak{P}^{X}_{\mathscr{R}} \cup \mathfrak{P}^{M}_{\mathscr{R}} \cup \mathfrak{O}_{\omega} \cup \mathfrak{A}$ has an answer set S, such that $S \models \psi$

V.3

Proof. ($\Rightarrow$) Take any $X \in \{C, L, P\}$, $\omega \in \Omega_{\mathscr{R}}$, and $\psi \in \mathscr{A}_{\mathscr{R}}$. Let $\omega = \mathsf{seen}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{seen}(\alpha_m, m)$, and assume that $\Gamma^X_{\mathscr{R}}(\omega) \neq \emptyset$. Consider any $\Phi \in \Gamma^X_{\mathscr{R}}(\omega)$, and observe that from Definition II.23 of the operator $\Gamma^X_{\mathscr{R}}$ it follows that every extension contains a single abducible. In line with the premise of our proof ('$\Rightarrow$'), let $\psi = \mathsf{rule}(i) \wedge \mathsf{seq}(j)$ for some $i, j \in \mathbb{N}_1$ and assume $\psi \in \Phi$. Note that the fact that $\Phi \in \Gamma^X_{\mathscr{R}}(\omega)$ and $\psi \in \Phi$ means that the sequence $\alpha_1; \ldots; \alpha_m$ matches with some sequence j extracted from the plan of rule i; specifically, $\exists(i : \gamma <\!\!- \beta \mid \pi) \in \mathscr{R} \exists \delta \in OS(\pi) \exists \delta' \in \mathscr{L}_\Delta : (\delta' R \delta)$ & $(\iota_\delta(\delta') = j)$, where R is the relation figuring in the observability clauses of Definitions II.9, II.12, and II.13 (i.e. equivalence '$=$', suffix '$\underline{\blacktriangleleft}$', and dilution '$\underline{\odot}$', respectively). Let $\mathfrak{P} = \mathfrak{P}^X_{\mathscr{R}} \cup \mathfrak{P}^M_{\mathscr{R}} \cup \mathfrak{O}_\omega \cup \mathfrak{A}$ and note the ASP encoding of abductive explanation ($\mathfrak{A}$) given by Definition V.12, on grounds of which every abducible is considered for candidate answer sets. Specifically, consider the abducible $\psi = \mathsf{rule}(i) \wedge \mathsf{seq}(j)$, and the candidate answer set S for which holds $\{\mathsf{rule}(i), \mathsf{seq}(j)\} \subseteq S$. Let aforementioned $\delta' = \alpha_1; \ldots; \alpha_n$, and note from Definition V.7 that the fact that $(\psi \rightarrow (\mathsf{obs}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{obs}(\alpha_n, n))) \in X_{\mathscr{R}}$ implies that $\{\mathsf{obs}(\alpha_1, 1), \ldots, \mathsf{obs}(\alpha_n, n)\} \subseteq S$ holds, because $\delta' R \delta$ is given. That S is an *actual* answer set follows from the fact that $\{\mathsf{seen}(\alpha_1, 1), \ldots, \mathsf{seen}(\alpha_m, m)\} \subseteq S$ and $\{\mathsf{obs}(\alpha_1, 1), \ldots, \mathsf{obs}(\alpha_n, n)\} \models \mathsf{obs}(\alpha_1, 1) \wedge \ldots \wedge \mathsf{obs}(\alpha_m, m)$, so that it holds that S is not ruled out by the constraint rule '$:\!-\ \mathsf{seen}(A, P), \mathsf{not\ obs}(A, P).$'.

($\Leftarrow$) The proof proceeds for '$\Leftarrow$' quite similarly to '$\Rightarrow$', given that any candidate answer set S of $\mathfrak{P}^X_{\mathscr{R}} \cup \mathfrak{P}^M_{\mathscr{R}} \cup \mathfrak{O}_\omega \cup \mathfrak{A}$ contains a single instance of both $\mathsf{rule}(i)$ and $\mathsf{seq}(j)$, and the fact that S is not ruled out (i.e. is an actual answer set, as assumed for '$\Leftarrow$') means that the instances of $\mathsf{seen}/2$ encoded by $\mathfrak{O}_\omega$ agree, in regard to the constraint rule '$:\!-\ \mathsf{seen}(A, P), \mathsf{not\ obs}(A, P).$', with the instances of $\mathsf{obs}/2$ accounted for by $\mathsf{rule}(i)$ and $\mathsf{seq}(j)$ by means of the encoding of the underlying observability clauses, so that the same holds for the background theories of $\Lambda_{\mathscr{R}}$ and for some $\Phi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi \models \psi)$. □

Lemma V.1 shows that the abductive theory has an extension that entails a certain abductive hypothesis (i.e. instances of $\mathsf{rule}/1$ and $\mathsf{seq}/1$) if and only if the corresponding ASP program has an answer set that does so as well; note that an equivalent result can be obtained in terms of abductive explanation '$\mid\!\approx_X$', but the result in terms of extensions of the abductive theory is more convenient for the following theorem, which states that the implementation is correct in consideration to its theoretical underpinnings.

Theorem V.1 (correctness of implementation)**.** *Given the conditions that hold for Lemma V.1 it holds that the implementation is correct, in the following sense.*

$$\forall X \in \{C, L, P\}\ \forall \omega \in \Omega_{\mathscr{R}}\ \forall \psi \in \mathscr{A}_{\mathscr{R}} :$$
$$\Lambda_{\mathscr{R}}, \omega \mid\!\approx^{\mathsf{sk}}_X \psi \quad \Longleftrightarrow \quad \mathfrak{P}^X_{\mathscr{R}} \cup \mathfrak{P}^M_{\mathscr{R}} \cup \mathfrak{O}_\omega \cup \mathfrak{A} \models_{\mathsf{c}} \psi$$
$$\&$$
$$\Lambda_{\mathscr{R}}, \omega \mid\!\approx^{\mathsf{cr}}_X \psi \quad \Longleftrightarrow \quad \mathfrak{P}^X_{\mathscr{R}} \cup \mathfrak{P}^M_{\mathscr{R}} \cup \mathfrak{O}_\omega \cup \mathfrak{A} \models_{\mathsf{b}} \psi$$

Proof. Consider the first conjunct of '&' in the claim, and the '⇒'-direction of '⇔' for this conjunct. Take any $X \in \{C, L, P\}$ and any $\omega \in \Omega_{\mathscr{R}}$, and let $\mathfrak{P} = \mathfrak{P}^X_{\mathscr{R}} \cup \mathfrak{P}^M_{\mathscr{R}} \cup \mathfrak{O}_\omega \cup \mathfrak{A}$. Take any $\psi \in \mathscr{A}_{\mathscr{R}}$ and let $\psi = \mathsf{rule}(i) \wedge \mathsf{seq}(j)$, noting that $\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{sk}}_X \psi$ requires ψ to be part of *all* extensions. Also, $\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{sk}}_X \psi$ means that $\Gamma^X_{\mathscr{R}}(\omega) \neq \emptyset$, and given that $\forall \Phi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi \models \psi)$ it follows from Lemma V.1 that for at least some answer set S of $\mathfrak{P}$ holds $S \models \psi$. To see that this holds for *all* answer sets of $\mathfrak{P}$, consider the fact that if for some answer set S' and some $\psi' \in \mathscr{A}_{\mathscr{R}}$ such that $\psi' = (\mathsf{rule}(y) \wedge \mathsf{seq}(z))$ it were the case that $S' \models \mathsf{rule}(y)$ or $S' \models \mathsf{seq}(z)$, such that $i \neq y$ or $j \neq z$, then because of Lemma V.1 it would have to be the case that $\exists \Phi' \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi' \models \mathsf{rule}(y))$ or $\exists \Phi' \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi' \models \mathsf{seq}(z))$. This contradicts the given fact that $\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{sk}}_X \psi$ because of the constraints of Definition II.16, which state that $\neg(\mathsf{rule}(i) \wedge \mathsf{rule}(y))$ and $\neg(\mathsf{seq}(j) \wedge \mathsf{seq}(z))$. Having shown that it holds for every answer set S of $\mathfrak{P}$ that $S \models \psi$, $\mathfrak{P} \models_{\mathsf{c}} \psi$ follows because if ψ is entailed by each answer set of $\mathfrak{P}$ then it is also entailed by their intersection. In regard to the '⇐'-direction of '⇔' in the claim this occurs in similar fashion; i.e. by turning to '⇐' in Lemma V.1 and observing that if an abducible is entailed by all answer sets then it can be skeptically inferred, because '$\mid\!\!\approx^{\mathsf{sk}}_X$' considers all answer sets and the constraints of Definition II.16 rule out the possibility of multiple abducibles being satisfied by a single answer set. Thus, from assuming $\mathfrak{P} \models_{\mathsf{c}} \psi$ it can be proven that $\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{sk}}_X \psi$.

For proof of the second conjunct of '&' in the '⇒'-direction, assume $\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{cr}}_X \psi$, i.e. $\exists \Phi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi \models \psi)$, and see that then from Lemma V.1 follows that for at least some answer set S of $\mathfrak{P}$ holds $S \models \psi$. If this is the case, then the union of all answer sets of $\mathfrak{P}$ entails ψ, and $\mathfrak{P} \models_{\mathsf{b}} \psi$. Conversely, for the '⇐'-direction note that if $\mathfrak{P} \models_{\mathsf{b}} \psi$ is given then for at least some actual answer set S of $\mathfrak{P}$ holds $S \models \psi$; this must be so because it is known that $\mathsf{rule}(i) \wedge \mathsf{seq}(j) \in \mathscr{A}_{\mathscr{R}}$ so that the rule identified by i must produce a sequence identified by j. This sequence agrees with the observed actions, given that the answer set is not ruled out, so that from Lemma V.1 then follows that $\exists \Phi \in \Gamma^X_{\mathscr{R}}(\omega) : (\Phi \models \psi)$, i.e. $\Lambda_{\mathscr{R}}, \omega \mid\!\!\approx^{\mathsf{cr}}_X \psi$. □

V.3

As Theorem V.1 shows, the implementation presented in this chapter can be said to be correct, in the sense that abductive explanations which can be skeptically/credulously inferred are cautious/brave consequences, as intended. In regard to reasoning about the observed agent's mental state, it furthermore holds that every description of mentalistic preconditions in terms of belief/1 and goal/1 that is entailed by an abductive extension, on grounds of the clauses of Definition II.15, has a model in terms of an answer set that encapsulates this description's corresponding encoding on grounds of Definition V.10. It is noteworthy in regard to reasoning about the observed agent's mental state that cautious/brave reasoning (in terms of the implementation) does not relate as straightforwardly to skeptical/credulous reasoning (in terms of the abductive theory) as is the case for reasoning about abductive explanations. This is mainly so if disjunction occurs in the agent's rules, as illustrated by the following proposition which can be considered a refutation of the claim made if the biimplication were Theorem V.1 considered with respect to the observed agent's mental state. For technical convenience, let $\mathscr{L}_{bg} \subseteq \mathscr{L}_1$ be the subset of the language $\mathscr{L}_1$ that consists of

atoms belief/1 and goal/1 composed by conjunction and disjunction, and τ_{ASP} the translation function of those predicates to ASP, here accepting expressions from $\mathscr{L}_{bg}$, translating atoms as defined in Definition V.9 with preservation of the logical connectives.

Proposition V.1. *Let* **R** *be the domain of* MYAPL *rules,* $\mathscr{R} \subseteq \mathbf{R}$ *a set of such rules,* $\Lambda_{\mathscr{R}}$ *the corresponding abductive theory as defined in Definition II.21,* $X \in \{C, L, P\}$ *an indicator of a perceptory condition, and* $\Omega_{\mathscr{R}}$ *the language of observables as defined in Definition II.18. Furthermore, let* $\mathfrak{P}^X_{\mathscr{R}}$ *be the encoding of the background theory as defined in Definition V.7 for* $X \in \{C, L, P\}$, $\mathfrak{P}^M_{\mathscr{R}}$ *the encoding of preconditions as defined in Definition V.10,* $\mathfrak{O}_\omega$ *the encoding of* $\omega \in \Omega_{\mathscr{R}}$ *as defined in Definition V.11, and* $\mathfrak{A}$ *the encoding of abduction as defined in Definition V.12. It then holds, assuming that* $\mathfrak{A}$ *(at least) encompasses the identifiers of rules and sequences that account for* ω*, that*

$$\begin{aligned}
&\exists \mathscr{R} \in \mathbf{R}\ \exists X \in \{C, L, P\}\ \exists \omega \in \Omega_{\mathscr{R}}\ \exists \phi \in \mathscr{L}_{bg} : \\
&\qquad \Lambda_{\mathscr{R}} \mathrel{|\!\approx^{\mathsf{sk}}_X} \phi \qquad \& \qquad \mathfrak{P}^X_{\mathscr{R}} \cup \mathfrak{P}^M_{\mathscr{R}} \cup \mathfrak{O}_\omega \cup \mathfrak{A} \not\models_{\mathsf{c}} \tau_{\text{ASP}}(\phi) \\
&\qquad\qquad\qquad \& \\
&\qquad \Lambda_{\mathscr{R}} \not\mathrel{|\!\approx^{\mathsf{cr}}_X} \phi \qquad \& \qquad \mathfrak{P}^X_{\mathscr{R}} \cup \mathfrak{P}^M_{\mathscr{R}} \cup \mathfrak{O}_\omega \cup \mathfrak{A} \models_{\mathsf{b}} \tau_{\text{ASP}}(\phi)
\end{aligned}$$

Proof. Let $\mathscr{R} = \{1 : p \mathrel{<\!\!-} q \text{ or } r \mid a\}$ and $\mathfrak{P} = \mathfrak{P}^X_{\mathscr{R}} \cup \mathfrak{P}^M_{\mathscr{R}} \cup \mathfrak{O}_\omega \cup \mathfrak{A}$, and assume $\omega = \mathsf{seen}(a, 1)$. It follows that $\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{sk}}_X} \mathsf{rule}(1)$ and $\mathfrak{P} \models_{\mathsf{c}} \mathsf{rule}(1)$, as stated by Theorem V.1. Also, see that $\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{sk}}_X} \mathsf{belief}(q) \vee \mathsf{belief}(r)$ as a result of the belief-precondition of rule 1, given that $(\mathsf{rule}(1) \rightarrow (\mathsf{belief}(q) \vee \mathsf{belief}(r))) \in M_{\mathscr{R}}$ (cf. II.15). However, observing that the encoding of Definition V.10 ensures that $\mathfrak{P}$ has an answer set S_1 for which holds $\mathsf{belief}(q) \in S_1$ but $\mathsf{belief}(r) \notin S_1$, and another answer set S_2 for which holds $\mathsf{belief}(q) \notin S_2$ but $\mathsf{belief}(r) \in S_2$, it is the case for the set $S_\cap = \bigcap\{S_1, \ldots, S_n\}$, given that $S_1, \ldots, S_n$ are the answer sets of $\mathfrak{P}$, that $\mathsf{belief}(q), \mathsf{belief}(r) \notin S_\cap$. Thus, $\mathfrak{P} \not\models_{\mathsf{c}} \mathsf{belief}(q) \vee \mathsf{belief}(r)$. Likewise, see that $\Lambda_{\mathscr{R}}, \omega \mathrel{|\!\approx^{\mathsf{cr}}_X} \mathsf{belief}(q) \vee \mathsf{belief}(r)$, but $\Lambda_{\mathscr{R}}, \omega \not\mathrel{|\!\approx^{\mathsf{cr}}_X} \mathsf{belief}(q)$ and $\Lambda_{\mathscr{R}}, \omega \not\mathrel{|\!\approx^{\mathsf{cr}}_X} \mathsf{belief}(r)$. However, given that $S_\cup = \bigcup\{S_1, \ldots, S_n\}$, again taking $S_1, \ldots, S_n$ as the answer sets of $\mathfrak{P}$, it holds that $\mathsf{belief}(q), \mathsf{belief}(r) \in S_\cup$, so that both $\mathfrak{P} \models_{\mathsf{b}} \mathsf{belief}(q)$ and $\mathfrak{P} \models_{\mathsf{b}} \mathsf{belief}(r)$. □

The phenomenon illustrated by above proposition should not be surprising, given that cautious and brave consequences of a program are defined, respectively, in terms of the intersection and union of its answer sets. Clearly, stating that some expression is entailed by every/some abductive extension (as in the case of skeptical/credulous abduction) is different from stating that this expression is satisfied by the intersection/union of possible models of those extensions. This discrepancy between the account of Chapter II and the implementation of the present chapter can thus be attributed to slightly different notions of consequence in the light of nonmonotonicity, which (as Theorem V.1 shows) do not affect conclusions that can be drawn in regard to explanations of observed behavior, but (as Proposition V.1 shows) do affect conclusions that can be drawn about an observed agent's mental state. In any case, the implementation can be deemed satisfactory also with respect

to reasoning about the observed agent's mental state, if it is considered that any *single* answer set of the program that encodes the abductive theory is a possible model for the skeptical claims $\phi \in \mathscr{L}_{bg}$ that can be made about the agent's mental state (cf. Theorem II.2), just as that any answer set which entails '$\mathsf{rule}(n)$' is a model for the credulous claims that can be made about the observed agent's mental state in regard to the rule identified by 'n' (cf. Theorem II.3); assuming translation of those claims by means of the function τ_{ASP}.

A final note concerns the fact that the ASP encoding can be said to improve upon the approach of Chapter II by representing negative MYAPL literals as negative ASP literals, instead of an atomic representation. The classical interpretation of $\mathscr{L}_1$ in Chapter II forces any literal to be either true or false, which is not desirable for observation-based attribution of mental states. That the use of strong negation is justified can be motivated by the fact that ASP atoms and their negative counterparts are part of an answer set only if there is 'reason' for this to be so; i.e. it is not by default the case that either p or $-p$ is in the answer set (for this to be so, the closed-world assumption would have to be made explicit by means of a rule '$-p$:– not p.' (Lifschitz, 2008)). Gelfond describes this characteristic of ASP by viewing the program as instantiation of a rational reasoner, stating "the reasoner should adhere to the *rationality principle* which says that *one shall not believe anything one is not forced to believe*" (2008, original emphasis). This point of view agrees with the one taken in this dissertation on the attribution of goals and beliefs, namely that those should be attributed not out of the blue, but as a result of attempted explanation of observed actions. Thus, apart from its formal correctness shown in Theorem V.1, the implementation also correctly captures this intuition that underlies mental state abduction.

3.2 A Note on Dynamics

In the preceding sections of this chapter, focus has been on implementing the abductive approach of Chapter II. This approach deals with defeasible inference of the mental state (i.e. beliefs/goals) an observed agent had *at the point it applied a particular rule and selected a particular plan*. It was stated in Chapter II that the agent's mental state quite possibly changes as a result of the plan it executes, and PDL was employed in Chapter III for modeling this dynamic aspect. It is legitimate to ask whether dynamics can also be incorporated into an implementation along the lines presented in the current chapter, and here some effort is taken to hint at answers to this question.

V.3

A popular logic-based framework to modeling dynamics of actions is the *situation calculus* (Reiter, 1991), which is often encountered in approaches that use logic programming (Turner, 1997). Zhang & Foo (2005) provide a translation between PDL, the situation calculus, and action languages that are also used in ASP approaches to reasoning about dynamic domains (Baral & Gelfond, 2005; Gelfond, 2008). In context of the situation calculus it is natural to consider instances of our predicates $\mathsf{belief}/1$ and $\mathsf{goal}/1$ as pertaining to an initial state in which the observed agent applied a rule and selected a plan, and extend those predicates to $\mathsf{belief}/2$ and $\mathsf{goal}/2$ by adding a situation argument. In comparison; just as $\mathsf{holds}(\phi, s_0)$ in situation calculus expresses that ϕ was the case in the initial state s_0, $\mathsf{belief}(\phi, s_0)$ — or $\mathsf{goal}(\phi, s_0)$ — can be used to express that the agent had the belief — or goal — ϕ in that initial state. By means of action descriptions encoded in ASP that pertain to

dynamics of the observed agent's mental state apart from dynamics of the environment, it is then possible to express change in the observed agent's mental state as a result of its actions, as presumed by the observer. The PDL-based Chapters III and IV can be used as basis for formulating the laws of those dynamics, leaning on the relation between PDL and situation calculus (Zhang & Foo, 2005). In (Sindlar et al., 2011) some guidance is given in that direction; also, the work of Baral et al. (2010) on finding Kripke models by means of ASP can serve as inspiration in this respect.

4 Reflection

This chapter presents an implementation in answer set programming, which, as shown in Section 2 shows, is directly founded upon the abductive theory put forward Chapter II. It adheres to the generate-and-test methodology of answer set programming (Section 1), such that in the 'generate' phase candidate abductive explanations are put forward, which are then 'tested' to see whether they would, if true, account for the observed actions. The implementation is evaluated in Section 3, and formally shown to correctly implement the abductive approach.

V.4

Chapter VI

Comparison to Related Work

This dissertation, as its title says, covers an interplay of two main themes: explanation of behavior, and mental state attribution. Topics in this general area have fostered a lot of research in artificial intelligence, under such umbrella terms as *plan recognition* (Kautz, 1993; Charniak & Goldman, 1993; Bauer & Paul, 1994; Rao, 1994; Ardissono & Sestero, 1996; Albrecht et al., 1998; Carberry, 2001; Avrahami-Zilberbrand & Kaminka, 2005; Geib & Steedman, 2007; Goultiaeva & Lespérance, 2007), *intention recognition* (Kiefer & Schlieder, 2007; Pereira & Anh, 2009; Doirado & Martinho, 2010; Sadri, 2011), *mental state recognition* (Rao & Murray, 1994; Dragoni et al., 2002), *theory of mind* (Pynadath & Marsella, 2005; Bosse et al., 2007c; van Ditmarsch & Labuschagne, 2007; Harbers et al., 2009; Hoogendoorn & Soumokil, 2010; Bosse et al., 2011), *mindreading* (Bosse et al., 2007a; Hendriks, 2010), and variations thereof (Appelt & Pollack, 1992; Brafman & Tennenholtz, 1994; Quaresma & Lopes, 1995; Kott & McEneaney, 2006; Booth & Nittka, 2008; Baral et al., 2010). To our present chapter, the prolificacy of research in plan/intention recognition, and related areas, has the benefit of having already inspired others to perform surveys of existing literature in this field. For this chapter we gratefully make use of the efforts of Carberry (2001) and Sadri (2011) in this respect by building on some of their findings and insights; specifically, in the sense that the corpus of work discussed in their surveys of literature on plan/intention recognition has, to some extent, guided our own choice of literature. It is noteworthy that Sadri (2011) actually incorporates our work on the mental state abduction functions (Sindlar et al., 2008), detailed in Chapter II (specifically Section 4), into her comparison of logic-based approaches to intention recognition. Thus, it could be said that the comparison of our approach to related work, as presented in the current chapter, already has some support in terms of an existing comparison of this kind. Nevertheless, it is our goal to be original, so, even though we have undoubtedly been influenced by existing surveys and discussions of work on plan/intention recognition, the following discussion of related work represents our own views.

In this chapter focus is on work that is, in some way or other, related strongly to ours. This essentially excludes approaches that are not primarily logic-based (Charniak & Goldman, 1993; Albrecht et al., 1998; Avrahami-Zilberbrand & Kaminka, 2005; Ramírez & Geffner, 2010), because those differ fundamentally from our own in the sense that they typically require prior observations. Such approaches are therefore left largely undiscussed (barring a few exceptions). Nevertheless, it should be noted those approaches can benefit from logic-based approaches like ours, and vice versa, given that the analytical work of this dissertation can be useful in determining which correlations between observed behavior and attributed mental state to look for in a probabilistic approach, whereas our approach can benefit from refinement with probabilities in order to, e.g., facilitate selection of 'best' explanations. Of course, such crossovers are not trivial to realize. Also, because of the paradigmatic difference, approaches from decision theory (Brafman & Tennenholtz, 1994;

Gmytrasiewicz & Durfee, 1995; Pynadath & Marsella, 2005) are not discussed here.

The remainder of this chapter consists of two principal sections, the first concerning related work on intention recognition, the second focusing on work in the sphere of mind-reading, and each of those sections is set up to thematically discuss selected works in chronological order. This chapter is concluded with a brief reflection.

1 Intention Recognition

This section discusses work on intention recognition which is principally logic-based. The terms 'plan recognition' and 'intention recognition' are often used interchangeably, both referring to the fact that the behavior of an agent is explained in terms of a plan it is presumedly executing ('plan' referring to a recipe from the agent's plan library, or its mental attitude, i.e. 'intention'). Many definitions of plan/intention recognition are furthermore stated in terms of agents' goals: Carberry (2001) defines plan recognition as "[inferring] *the goals of the other person and a portion of that person's plan for achieving those goals*", and Sadri (2011) actually considers 'intention recognition' the same as 'goal recognition'. Other descriptors have also been frequently used, showing in the titles of the works discussed in this section.

1.1 Appelt & Pollack (1992)

This well-known and regularly cited paper presents an approach to *plan ascription* using weighted abduction. Appelt & Pollack distinguish between plan recognition (defined as the case where "*the observer must distinguish between the actor's goal as well as his plan*") and plan evaluation (where "*the observer is given the actor's goal*"), using the term 'plan ascription' to refer to "*the general process that subsumes plan recognition and evaluation*". The authors describe 'abduction' as "*the process of reasoning from some observations to the best explanation for them*", and in their approach use annotations (weights) on the rules of a logical theory to encode a preference ordering on models for the theory, to yield such 'best' explanations in the process of abduction. The use of weights is presented as an alternative to statistical methods, and a major advantage (according to the authors) is that weights can be employed to express domain-specific information about the likelihood that any particular proposition is true, as well as preferences for particular explanations. In this regard, the use of weights could usefully complement our approach; nevertheless, it brings with it the same issue of statistical approaches, which is that a basis is required to determine this domain-specific information.

One of the aspects in which our approach improves upon that of Appelt & Pollack (1992) is in terms of expressiveness of the formalism, especially where it concerns dynamics. It was shown in Chapter II how our approach handles abduction with respect to a first-order theory obtained from rules with plans that contain general compositional constructs like sequence, choice and iteration, and in Chapter III dynamics interpretations were given in terms of PDL models. In contrast, the approach of Appelt & Pollack is limited to actions as propositional statements, as in our first-order formalization. It remains

unclear how compositionality in plans, or unobservable test actions, should be handled. Also, the distinction between observed and observable, explicitly made in our account, is not found in that of Appelt & Pollack. Moreover, in their approach, ascription of mental states happens using operators pertaining to 'belief' and 'intention', for which no formal interpretation is given, although Cohen & Levesque (1990) are cited. Semantics of such operators in agent programming is typically more in line with the account of Rao & Georgeff (1991), however, so that it is also unclear how the approach of Appelt & Pollack relates to this domain. In summary, it would seem that our approach is more general than that of Appelt & Pollack (1992) in regard to reasoning about dynamics, noting that it could benefit from this account in regard to employing domain-specific information for selection of 'best' explanations, in cases where such information is available.

1.2 Bauer & Paul (1994)

Bauer & Paul present an abductive approach to plan recognition using modal temporal logic, augmented with probabilistic selection of explanations based on Dempster-Shafer Theory. In regard to this extension, the same remarks apply as made in Section 1.1, which is that it could contribute to an approach like ours in cases where the necessary information is available for determining domain-specific probabilities. Concerning the choice of (modal) logic for reasoning about behavior, we contend that our use of PDL is more suited than Bauer & Paul's choice of temporal logic for reasoning about plans of software agents; which, at the end of the day, are 'just' programs.

1.3 Rao & Murray (1994)

The work of Rao & Murray (1994) deals with the recognition of plans of BDI-based agents, and it is recognized by the authors that the term 'plan' concerns "*plans as abstract structures or recipes for achieving certain states of the world*", as well as "*plans as complex mental attitudes intertwined in a complex web of relationships with other mental attitudes of beliefs, desires, and intentions*". This work focuses on *reactive plan recognition*, defined as "*the use of plans as recipes to guide the recognition process and the use of plans as mental attitudes to constrain the recognition process*". In this regard, as well as the fact that the underlying formalism (Rao, 1994) is a dynamic logic, the approach of Rao & Murray (1994) and ours are closely related. It is therefore interesting to note that they make no mention of the fact that attribution of mental states, based on recognition of agents' plans, involves some form of nonmonotonicity (as we have in Chapter II) or 'past possibility' in the dynamic logic (as we have in Chapter III), except for acknowledging that "*while the executing agent can choose an applicable plan, one after the other, until one of them succeeds, the observing agent should attempt to recognize all the applicable plans simultaneously*". Rao & Murray do not delve into this matter beyond stating that "*at any point in time the current recognition trace will enable the agent to infer the beliefs, desires, and intentions of other agents*", a statement which is given no formal support. Moreover, this statement is later contradicted by the fact that mental states are taken to be sets of beliefs and intentions, of which desires are left out with the motivation that only beliefs and intentions are maintained by agents from one state to

the next. Thus, although both approaches employ some form of dynamic logic, it seems that only ours explicitly handles the fact that explaining observed behavior typically involves maintaining multiple possible explanations. Instead of focusing on this aspect, Rao & Murray do pay significant attention to how plan recognition is intertwined with plan execution, which we do not.

1.4 Geib & Steedman (2007)

Although not directly logic-based, this work of Geib & Steedman (2007) is of interest because it points out commonalities between the areas of plan recognition (PR) and natural language processing (NLP). To this extent the authors abstract away from specifics of PR/NLP, treating action sequences/natural language utterances as observations, plan structures/parse trees as explanations, and referring to the explanative process as *parsing*. In the words of the authors, "*to parse a set of observations into an explanation both PR and NLP must specify the patterns of observations they are willing to accept or the rules that govern how the combinations can be combined*". Accordingly, they refer to such rules as *grammars* in context of both PR and NLP. This approach agrees with ours, given that descriptions of plans (cf. Definition II.4) can be viewed as regular expressions over primitive actions and tests (Harel et al., 2000). Thus, an alternative way to perform mental state abduction with such plan descriptions could be to view them as grammars, interpreted such that successful parsing of observed actions allows for abduction of the mental state precondition derived from the related rule. Naturally, the matters of incomplete observation and incremental recognition must then be taken into account, which most likely is not trivial in this context.

1.5 Goultiaeva & Lespérance (2007)

The approach described in this work of Goultiaeva & Lespérance (2007) is one of the few approaches to plan recognition specific to agent programming. It comprises a formal model of plan recognition based on the situation calculus, which is meant for inclusion in the ConGolog agent programming language. This model makes use of special annotation actions to recognize ConGolog procedures, and can handle concurrent execution of actions, although it does not handle concurrency on the level of procedures. Recognition is incremental, in the sense that it can occur in-step with the agent performing its actions, and in this sense is similar to our own approach (which, as it stands, does not explicitly handle concurrency on any level — see the remarks in Section 3.5.2 of Chapter II, though). A major difference is that our approach abstracts from specifics of the agent programming language, using a general description of plans as basis for recognition. This is a useful property, as it makes it easier to apply our approach to agents programmed in other, similar, programming languages. Moreover, the model of Goultiaeva & Lespérance does not account for missing observations, thus lacking a feature which is crucial for plan recognition systems to be considered robust (Carberry, 2001; Sadri, 2011).

1.6 Pereira & Anh (2009)

Pereira & Anh (2009) present an approach to intention recognition that is of interest because it mixes probabilistic methods with logic programming. To this extent the declarative language P-LOG is used, which comprises answer set programming as foundation for a logical representation of (situation-sensitive) Causal Bayes Networks. The approach differs fundamentally from ours, mainly because of the fact that a probabilistic model is used, yet also because plan recognition does not occur with respect to a plan library. Instead, the probabilistic model is used to determine an agent's most likely goals from contextual information, such as the observed state of the environment. A planner is then used to determine the action sequences that could result in achievement of the goal; a useful property is that the account allows for 'plugging in' any kind of plan generator. This leads to a model which is substantially different from ours, but it nevertheless seems important to point out the existence of approaches that mix logic programming with probabilities to bring the best of both worlds.

2 Mindreading

This section discusses existing work that is somehow related to our logical account of mindreading, as presented in Chapter IV, and includes some work (Quaresma & Lopes, 1995; Dragoni et al., 2002) which is considered by others (Carberry, 2001; Sadri, 2011) to fall into the category of intention recognition. However, given that such work specifically deals with attribution of mental states based on single (communicative) actions and does not deal with longer-term plans, it is best compared to our account of mindreading — MPs IV.2 and IV.6, specifically — warranting its inclusion here. Due to restrictions of space and scope this section is limited to selected works, leaving out some others that may also be of interest (Gmytrasiewicz & Durfee, 1995; Wahl & Spada, 2000; Laird, 2001; van Lambalgen & Smid, 2003; Baral & Gelfond, 2005; Pynadath & Marsella, 2005; Doirado & Martinho, 2010; Hoogendoorn & Soumokil, 2010).

2.1 Quaresma & Lopes (1995)

Quaresma & Lopes (1995) propose "*a framework that supports the recognition of plans and intentions behind speech acts through abductive inferences over discourse sentences*", employing the epistemic operators of Appelt & Pollack (1992) for representing speech acts in terms of logic programming rules, based on a description of those speech acts in terms of an action language (Gelfond, 2008). Inference is based on abductive planning in terms of the event calculus, with event-related predicates (*happens*/1, *act*/2, $</2$) as the abducibles, and mental state attribution occurs through deduction based on speech act descriptions. An important difference with our approach lies in the fact that Quaresma & Lopes do not seem to consider the mindreader-relativity of mental state attribution. Consider, e.g., the '*inform(s,h,p)*' speech act, stating that speaker '*s*' informs hearer '*h*' of the proposition '*p*',

described (in the action language) as follows.

inform(s,h,p)	causes	*bel(h,bel(s,p))*
	if	*bel(s,p), bel(s,int(s,inform(s,h,p))).*

Quaresma & Lopes give the following motivation for this description: "*If a speaker believes in a proposition and intends to inform the hearer about that proposition, then the hearer will believe that the speaker believes it*" and justify it by stating that the sender '*s*' must be sincere. In our estimation, though, the above is a fundamentally incorrect description of this speech act for the purpose of mental state attribution (unless it is one's goal to model *telepathic* mindreading, but as Quaresma & Lopes motivate their approach from a perspective of 'man-machine interaction', this does not appear to be the case). Our problem with the above description relates to the inclusion of '*bel(s,p)*', which is defined as "agent *s* believes that *p* is currently true". Thus, the hearer *h* of the speech act, on the above account, only attributes the belief *p* to the sender *s* if it is *actually* the case that the sender believes *p*; a condition which is independent of the hearer (unless, of course, it has telepathic mindreading powers)![1]

From their account it appears that Quaresma & Lopes (1995) waver between 'mindreading', on the one hand, and specification of behavior of agents in a multi-agent system, on the other. In the former case it should be assumed that agents' mental state cannot be accessed, whereas in the latter this assumption can be dropped. Furthermore, it is unclear why the description of the *inform/3* speech act states that '*s*' should believe in itself having the intention to inform '*h*' of '*p*', instead of just stating '*s*' to have this intention. Similar remarks apply to other speech act descriptions presented by Quaresma & Lopes, which, as pointed out, at times confuse 'mindreading' with 'telepathy'. It is our view that approaches such as that of Quaresma & Lopes (1995) can benefit from a basis in terms of a formal account of mindreading, as we have given in Chapter IV. In turn, our approach can profit from work on reasoning about mental state attribution in dynamic domains, using logic programming rules — as Quaresma & Lopes have given on grounds of action language descriptions of speech acts, and for the implementation of which they employ the event calculus. Insights from this domain can be helpful to extend an implementation along the lines of Chapter V to also model evolution, based on observed actions, of the mental state attributed to agents, given that our `belief/1` and `goal/1` predicates are regarded as referring to the agents' mental state in an initial state (cf. Section 3.2 of Chapter V, specifically).

2.2 Dragoni et al. (2002)

VI.2

This approach of Dragoni et al. (2002) is founded upon a multi-context system, in which each context interprets a specific nesting-level of operators for belief/intention attribution. Contexts are connected through bridge rules that capture the relationships between beliefs

[1] We understand that if it is known that the sender can only perform e.g. the *inform*/3 speech act if itself believes the fact it is communicating to be true, then attribution of the communicated fact as belief to the sender is warranted. Our objection concerns the fact that a check on the *actual* belief of the sender plays a role in what is presented as a model of (non-telepathic) mindreading by the hearer.

and intentions of the hearer, and facts within a context can be manipulated by abductive update procedures. Speech acts, of the same kind as those considered by Quaresma & Lopes (1995), are represented in STRIPS-like fashion, using preconditions and postconditions. However, Dragoni et al. give a more appropriate formalization of speech act-based mental state attribution than Quaresma & Lopes do, recognizing that attribution should occur relative to the 'mindreader', and therefore not be dependent on conditions that require accessing the contents of the sender's actual mental state. Thus, the '*inform(s,h,p)*' speech act, in their account, has the immediate effect of '*h*' attributing to '*s*' belief in the truth of '*p*', without having '*bel(s,p)*' as precondition, representing the underlying knowledge that this fact is a precondition to the (sincere) sender's performance of this speech act. The account of Dragoni et al. (2002) is in this regard a detailed exposition of speech act-based attribution, also focusing on abductive update of the model of others' mental states. Whether modeling mental state attribution using contexts and bridge rules is computationally feasible is a question left unanswered by Dragoni et al., but their account in any case agrees with ours in the sense that attribution is relative to the mindreader, and nesting of operators is also restricted.

2.3 van Ditmarsch & Labuschagne (2007)

Van Ditmarsch & Labuschagne (2007) use dynamic (doxastic) epistemic logic to present a detailed case study in theory of mind, restricted to agents' beliefs about others' beliefs. This is similar in spirit to our own account, a crucial difference being that our focus is on mental state attribution in regard to actions (either observed or presumed), whereas van Ditmarsch & Labuschagne (2007) focus on characterizing three types of 'believers', which have varying degrees of preference over their own beliefs and those attributed to others. Another aspect on which our account differs from that of van Ditmarsch & Labuschagne (2007) is the fact that in the latter nesting of (doxastic) operators is unrestricted, i.e. can occur 'ad infinitum'. This is not realistic if one's goal is to model bounded reasoners, in which case our approach, where attribution of beliefs/goals occurs by means of simple (non-modal) propositions, may be preferred.

2.4 Harbers et al. (2009)

This work of Harbers et al. (2009) presents a study of modeling agents with a theory of mind (ToM), comparing prototypical implementations of agents without ToM, agents with a ToM based on theory-theory (TT), and agents with a ToM based on simulation-theory (ST). Implementation is done by means of the agent programming language 2APL, using the standard version for implementation of NT and TT agents, and a version of 2APL with modules for implementation of ST agents. The implementations are evaluated from the point of view of explainable AI, and Harbers et al. find that the ToM-enabled agents (i.e. both TT-based and ST-based) are preferable over the NT agents, because, even though their observable behavior may be identical, the agents with ToM can be made to explain their actions in terms of goals or beliefs, which the NT agents cannot. The ST agents are furthermore preferred over the TT agents, mostly because of the fact that they are consid-

ered to be easier to implement and less error-prone, given that modules allow for reuse of code, and the use of agents' own means for deliberation to reason about the goals, beliefs, intentions, etc. of others. This finding supports our claim of Chapter IV (Section 3.3, specifically) that modules are suitable for instantiating the 'Possible World Boxes' in the model of mindreading by Nichols & Stich (2003), which in ST-spirit also make use of the agent's general resources for reasoning (cf. Figure IV.2); noting, though, that Nichols & Stich do not strictly segregate TT from ST, but attempt to find a middle ground.

2.5 Baral et al. (2010)

In this work, Baral et al. (2010) use answer set programming to find Kripke models of a theory in dynamic epistemic logic. The approach is illustrated by means of the Muddy Children problem, showing how with additional information (in terms of questions asked and announcements made by the father of the children) knowledge of the actual state of affairs is refined, resulting in a decrease of the possible models of the actual state. This agrees in spirit with our approach to mental state abduction; specifically, in regard to our implementation (Chapter V), where the number of answer sets decreases monotonically with additional observations (as shown for the underlying functional approach in Corollary II.2). The approach of Baral et al. can in this sense form a solid basis for an ASP implementation along the lines of our work in Chapter III, which concerns modeling by means of PDL. As stated in Section 3.2 there exist parallels between PDL and the situation calculus (Reiter, 1991), and it is therefore of interest that the work of Baral et al. (2010) draws parallels with the situation calculus as well.

2.6 Bosse et al. (2011)

This work by Bosse et al. (2011) concerns a BDI-based model of theory of mind, formalized in terms of the LEADSTO modeling language, and analyzed through simulation. Such an approach is also taken in other work of Bosse et al. (2007a,b,c), and, since the work discussed in the present section is to a large extent based on that earlier work, this discussion applies there as well. The approach taken by Bosse et al. can be summarized as consisting of the following steps: first, a case study is described and it is argued why and how theory of mind is useful for agents in that case; then, a model of theory of mind is formalized in terms of the LEADSTO language; and, lastly, simulation experiments on basis of this formalization are run, and subsequently analyzed, by means of the LEADSTO software environment. Bosse et al. (2011) discuss three such case studies, also encountered in their earlier work (Bosse et al., 2007a,b,c), and formalize those cases in terms of their own interpretation of the BDI model of practical reasoning. This formalization is done using the LEADSTO modeling language, and consists of specifications of temporal dependencies between predicates that state agents' beliefs, desires, intentions, actions, or other facts deemed relevant. Essentially, these dependencies are IF-THEN rules with a temporal interpretation, and simulations on grounds of those rules yield traces that represent the (binary) truth values of atomic propositions over time. In this regard, a downside of the LEADSTO-based approach is that it allows only for representing rather simplistic specifications of theory of mind, as it has no

means for expressing nonmonotonicity, or possibility in a modal sense; i.e., it allows only for implicatory relations between mental states or behavior. This shows through in the fact that the approach of Bosse et al. (2011) focuses on rather limited cases of attributing beliefs and desires to others, ignoring, for example, the general problem of explaining agents' observed behavior by attribution of mental states (which more often than not involves handling distinct possible explanations). Thus, and also because the premises and consequences in LEADSTO dependencies are restricted to conjunctions of literals, makes the approach of Bosse et al. suitable for modeling and simulation of scenarios in which agents' envisioned theory of mind fits within the rather restrictive constraints of the modeling tool, but not so much for more refined scenarios.

3 Reflection

In this chapter, existing work on the subjects of intention recognition and mindreading has been discussed. As pointed out before, research in those areas has been quite prolific, and as a result of this our discussion of related work is limited to those works which we consider to be particularly relevant. In summary, it appears that our work is a rather rare bird, in the sense that it focuses on explanation of the behavior of BDI-based software agents and heterogeneous agents alike, taking into account that multiple possible explanations may exist by way of nonmonotonicity and the existential modality of dynamic logic. Importantly, our approach is furthermore supported with (a specification for) implementation in terms of a state-of-the-art nonmonotonic logic programming paradigm, opening up possibilities for practical evaluation of this approach in, for example, the (serious) gaming domain. Where possible, it has been attempted in this chapter to point out in which respect related work can be used to complement an approach based on the work of this dissertation; be it in regard to the theoretical model employed for reasoning about behavior, or in regard to its implementation.

VI.3

Chapter VII

Conclusion

> " 'First of all', he said, 'if you can learn a simple trick, Scout, you'll get along better with all kinds of folks. You never really understand a person until you consider things from his point of view —' "
>
> Harper Lee, *To Kill a Mockingbird* (1960)

In the current chapter, the work presented in this dissertation is evaluated and reflected upon. Specifically, it considered whether this work achieved what it aimed for, by reviewing the research questions formulated in Section 2 of the introductory chapter to see whether these have been answered. Also, given the inevitable conclusion that there are relevant matters that we have not addressed, it is pointed out in which directions future research may follow up on the work presented in this dissertation.

1 Main Results

Recall that the overall research question that we set out to address (cf. page 4), was how the explanative attribution of mental states by BDI-based agents can be realized. To facilitate answering this question it was split up into five research 'sub-questions', each addressed in one of Chapters II–VI.[1] In the remainder of this section, these five questions are revisited and the main results are highlighted.

The first research question asked how the explanation of BDI-based agents' behavior can be formalized. In Chapter II this has been addressed by formulating an abductive framework, comprising background theories that describe observable actions and agents' mental states in relation to rule application, with the use of which observables can be explained in order to infer abducible hypotheses. It was shown how this logical abductive approach can be translated into the mental state abduction functions, which group together mentalistic rule preconditions (consisting of beliefs and goals) that have a counterpart in the abductive extensions. The elements in the output of the mental state abduction functions were shown to have both a skeptical and credulous interpretation, thus concisely capturing the essence of the logical approach. In summary, Research Question 1 has been answered on two levels: that of a classical logical approach, and that of its functional implementation. Furthermore, particular properties of the functional approach were illustrated formally, being of interest to the overall research question of how such an approach can be realized.

In Research Question 2 it was asked how the dynamics involved in the approach to answering the first question, i.e. the mental state abduction functions, can be modeled. This

[1] Noting that Research Question 5, formulated as such, is not so much a direct sub-question of the overall research question, although the consideration of related work does fall within its scope.

question arises when it is considered that the elements in the output of those functions refer to beliefs and goals an observed agent can be presumed to have had in a state *preceding* its actions. For answering this question, a propositional dynamic logic (PDL) with means for expressing attribution of mental states and observation of actions was employed, with focus on modeling the states of affairs that can be presumed to be the case if observed (and possibly missed) actions are taken into account, along with the mentalistic rule preconditions whose abduction they warrant. The converse dynamic modality was utilized here to express that particular beliefs and goals can be ascribed to the agent in possible states that precede its actions, using the existential modality as a counterpart for expressing defeasibility of the presumption that the agent *possibly* had that particular mental state. Furthermore, modeling the dynamics in mental state abduction allows for expressing the fact that an agent's mental state evolves as a result of its actions. In Chapter III this has been acknowledged by constraining models to be in accordance with the fact that ascription occurs on grounds of a *plan* an agent was executing (i.e. intention it presumedly had); it was shown that inconclusiveness on part of the observer then arises in certain cases. Summarizing, those results, it is seen that Research Question 2 has been answered by modeling the input (an observed sequence) and output (abduced mental states) of the mental state abduction functions as such, in regard to which also the modeling of unobserved actions of both the observable (but missed) and unobservable (test actions) kind has been taken into account.

The first two research questions have been answered with the underlying assumption that the rules of agents are known to the observer. This is not the case for Research Question 3 that pertains to the formalization of 'mindreading', which in our context is considered to be the first-order attribution of goals and beliefs based on general premises (i.e. observed actions, or other mental states ascribed to the agent). As a conceptual basis, two psychological models of mindreading are discussed in Chapter IV, and subsequently used to justify logical schemata termed 'mindreading patterns'. Principally, the contribution of those patterns resides in the explication of the logical *form* that can be discerned in regularities of mindreading as described by psychologists. This form is expressed using a slightly extended version of the formalism used in answering the previous research question, which is a variant of PDL. Thus, the formalization of mindreading is compatible with the approach to answering Research Question 2, and can also, by means of existing translations, be applied to other popular action formalisms. It is safe to summarize the above by stating the work of Chapter IV to indeed provide an answer to the third research question, notwithstanding the fact that its scope is limited to mindreading in virtual environments by BDI-based agents; specifically, it is not to be evaluated as a formalization of mindreading in the way it is done by humans.

In light of the overall research question of how the explanative attribution of mental states by BDI-based agents can be realized, Research Question 4 asks how mental state abduction can be implemented. After all, if it is known how a formal technique is to be implemented, one can try and put it into practice. The motivation of our research, which stems from an application domain — virtual characters for (serious) games and training environments — that requires minimizing the demand on available resources by application components, has driven us to look for state-of-the-art means to answer this question. This has resulted in selection of answer set programming (ASP) as the paradigm for presenting

an implementation that follows typical ASP methodology, taking the nonmonotonicity of mental state abduction into account by implementing this approach, as shown in Chapter V. The term 'implementation' here refers to the program that can be derived from the abductive theory in general, as put forward in answering the first research question, as opposed to referring to a specific such program. As such, we give a specification of how to implement our proposed techniques. It was formally proven that the resulting implementation is correct, in the sense that a bi-directional correspondence was shown to exist with the abductive approach of Chapter II, in regard to skeptically/cautiously and credulously/bravely inferred explanations. Given that the work reported in this chapter presents a specification for implementation of our techniques that is formally proven to yield satisfactory results, in terms of correspondence to its theoretical basis, the fourth research question is considered answered.

Last but not least, Research Question 5 concerns the position of our work in context of related approaches. This question is important, not only by itself but also in light of answering the overall research question; after all, the general topic of explanative mental state attribution extends far beyond this dissertation, and it is thus important to devote some attention to how bridges could be built between other approaches and ours. In answer of it, several existing works are reviewed that we consider to be closely related to ours, in the sense that they take a primarily logic-based approach to similar topics. Recent literature surveys have been useful in this regard as those constitute overviews of the current state of the art, and we have gratefully enlisted them for our efforts to answer the final research question. The gist of that answer, put forward in Chapter VI, is that our account of explanative mental state attribution contributes to the corpus of existing research (as far as we know it) in several respects. First and foremost, it focuses specifically on the case of BDI-based software agents, both in the sense of formalizing the explanation of such agents' behavior, as well as focusing on intended application by such agents. This agent-to-agent setting is, of course, natural in context of our underlying motivation to enable the development of more interesting virtual characters. Furthermore, we find our approach of employing PDL to formalize the dynamics involved in the abductive approach to be of interest, partly because of the light it sheds on the different interpretations of the 'possibility' involved in mental state attribution. The same holds for our use of PDL to tackle this topic on grounds of actions that stem from partially observed plans/intentions, which to our knowledge is a novel use of this framework. It is well-known that PDL is a logic suitable for reasoning about programs, but it is relevant to point out that in the agent-to-agent setting of our approach it is used to model the reasoning about a program (agent) *by a program* (observer). The same machinery is also used to formalize mindreading, something which has been done before using various formalisms, but not so specifically in context of psychological models and the realization that it involves adopting the intentional stance. Also, the possibility of combining intention recognition with mindreading, as facilitated by our choice of formalism to tackle those topics, is worth noting.

All in all, it can be concluded that this dissertation has gone a long way to answering the research questions lined out at the start, and has made interesting contributions in doing so. Inevitably, though, yet also fortunately, there are questions that remain to be answered, some of them stirred in the course of our research as reported here.

2 Future Research

As discussed in the previous section, the overall research question of this dissertation has been addressed to a significant degree; nevertheless, questions remain for future research. One of those questions concerns the selection of 'best' explanations. The work of this dissertation has focused on the selection and interpretation of explanations which are *plausible* in the light of observed actions, of which there may be several, but little has been said about which of such plausible explanations should be considered as being *the* explanation. Arguably, if one has no justification for settling on any single explanation, then our approach of narrowing down the possibilities is the best one can do. A natural basis to narrow down the possible explanations further than we have done, with respect to our setting of mental state attribution, are the beliefs the observer itself holds. It has already been shown in this dissertation how the observer's beliefs can be used as grounds for attribution in the context of modeling with PDL; future research could follow this direction further and see how those insights translate to implementation.

If one is really looking for single 'best' explanations, then in our approach one needs to resort to credulously picking any of the equally plausible possibilities. In order for such choice to be justifiable in the context of a framework like ours, it should be extended with grounds to support that choice. Computational argumentation could then, for example, come into play as a tool for choosing among competing explanations. Previous chapters have mentioned 'credulity' in the context of nonmonotonic reasoning, as well as PDL-based models. These formalized notions can be taken as starting point for tackling the question of *why* the beholder (credulously) prefers one explanation over another. In Chapter VI we have also tried to sketch how related approaches may inspire formulating grounds for selection of best explanations, in terms of probabilities — by, e.g., determining from observations of prior application scenarios the likelihood with which agents have particular goals or beliefs (Albrecht et al., 1998; Ramírez & Geffner, 2010) — or otherwise (e.g. weighted abduction (Appelt & Pollack, 1992)). Some of our own work has actually focused on determining such grounds on an a priori basis, considering the fact that an RPG can be regarded as multi-agent society described by an organizational model, so that knowledge of agents' roles, or particular norms governing the context of behavior, can be used to order explanations (Sindlar et al., 2009a). Also, we have considered agents' location or motion in their (spatial) environments in regard to objects of interest, employing the notion of object affordance (i.e. actions that can be performed on objects), possibly in combination with knowledge of agents' plans, to rank explanations (Sindlar & Meyer, 2010).

Perhaps equally pressing as the search for single 'best' explanations, is the practical evaluation of our approach in the domain that motivated it in the first place: virtual characters for (serious) games. It has been stated earlier that as grounds for our implementation we chose a state-of-the-art logic programming paradigm in hope to facilitate this type of application, but it can only be seen through practical evaluation whether that hope was idle, or not. In order to verify this, typical game scenarios can be set up, with BDI-based agents instantiating the virtual characters, and the techniques proposed in this dissertation evaluated in that context. This is not trivial, because although the type of games that we have in mind do exist (RPGs), their virtual characters are currently not BDI-based agents. It has

been argued here that BDI is, however, suitable for realizing the complex social behavior of characters in those games, a claim which as such lends itself as topic of future research. In light of practical questions such as those, it is furthermore relevant to see how our approach is to be considered in the context of agent programming languages that involve (fragments) of first-order logic with variables, given that we have focused on a propositional agent programming language. Our gut feeling is that the main challenge then becomes to sensibly restrict the binding of variables that the beholder considers, whereas the core of the approach remains the same; but, of course, such claims require evaluation.

The implementation presented in Chapter V focused solely on the abductive formalization, so that research focusing on the implementation of the PDL-based parts seems to us worthwhile. In this context, it seems important to determine whether our variation on standard PDL presents a relatively minor extension of this formalism (as our contention would be), or whether it also has implications for its computational properties. Another point for future research concerns the synergy of two topics of this dissertation, as it would be interesting to see in more detail how our work on the use of PDL for modeling the dynamics of mental state abduction combines with that on the use of PDL for mindreading. Given that in the former chapter we have taken a strongly semantic approach by focusing on different classes of models, and in the latter a more syntactic approach by focusing on the form of mindreading patterns, this is not straightforward. Finally, apart from evaluating the feasibility of our framework practically, future research may delve into more theoretical issues such as its computational complexity, of which we have steered clear.

3 A Final Reflection

As discussed in this concluding chapter, the work presented in this dissertation contributes in different respects to existing research on explanative mental state attribution. In this sense, we feel it has achieved its ambitions. It was also discussed how future research can follow up on the work of this dissertation; in that sense we hope it will fulfill its potential.

Bibliography

Afonso, N. & Prada, R. (2008). Agents that relate: Improving the social believability of non-player characters in role-playing games. In S. M. Stevens & S. J. Saldamarco (Eds.), *Proceedings of the Seventh International Conference on Entertainment Computing — ICEC*, volume 5309 of *LNCS* (pp. 34–45).

Albrecht, D., Zukerman, I., & Nicholson, A. (1998). Bayesian models for keyhole plan recognition in an adventure game. *User Modeling and User-Adapted Interaction*, 8(1-2), 5–47.

Alechina, N., Dastani, M., Khan, F., Logan, B., & Meyer, J.-J. Ch. (2010). Using theorem proving to verify properties of agent programs. In M. Dastani, K. V. Hindriks, & J.-J. Ch. Meyer (Eds.), *Specification and Verification of Multi-agent Systems* (pp. 1–33). New York, NY: Springer.

Alechina, N., Dastani, M., Logan, B., & Meyer, J.-J. Ch. (2007). A logic of agent programs. In W. Cheetham & M. Goker (Eds.), *Proceedings of the Twenty-Second National Conference on Artificial Intelligence — AAAI* (pp. 795–800). Menlo Park, CA: AAAI Press.

Aliseda-Llera, A. (1997). *Seeking Explanations: Abduction in Logic, Philosophy of Science, and Artificial Intelligence*. PhD thesis, University of Amsterdam.

Appelt, D. E. & Pollack, M. E. (1992). Weighted abduction for plan ascription. *User Modeling and User-Adapted Interaction*, 2(1-2), 1–25.

Ardissono, L. & Sestero, D. (1996). Using dynamic user models in the recognition of the plans of the user. *User Modeling and User Adapted Interaction*, (pp. 1–36).

Avrahami-Zilberbrand, D. & Kaminka, G. A. (2005). Fast and complete symbolic plan recognition. In L. P. Kaelbling & A. Saffiotti (Eds.), *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence — IJCAI* (pp. 653–658). Denver, CO: Professional Book Center.

Baeten, J. C. M. & Weijland, W. P. (1990). *Process Algebra*. Cambridge University Press.

Baral, C., Gelfond, G., Pontelli, E., & Son, T. C. (2010). Using answer set programming to model multi-agent scenarios involving agents' knowledge about others' knowledge. In W. van der Hoek, G. Kaminka, Y. Lespérance, M. Luck, & S. Sen (Eds.), *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems — AAMAS* (pp. 259–266). Toronto, ON: IFAAMAS.

Baral, C. & Gelfond, M. (2005). Logic programming and reasoning about actions. In *The Handbook of Temporal Reasoning in Artificial Intelligence* (pp. 301–332). Amsterdam: Elsevier.

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: The MIT Press.

Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition*, 21, 37–46.

Barwise, J. & Etchemendy, J. (1992). *The Language of First-Order Logic*. Stanford, CA: CSLI.

Bauer, M. & Paul, G. (1994). Logic-based plan recognition for intelligent help systems. In C. Backström & E. Sandewall (Eds.), *Current Trends in AI Planning* (pp. 60–73). Amsterdam: IOS Press.

Belnap, N. D. (1977). A useful four-valued logic: How a computer should think. In J. M. Dunn & G. Epstein (Eds.), *Modern Uses of Multiple-Valued Logic* (pp. 5–37). Dordrecht: Reidel.

Bessant, B. (1996). The babelism about induction and abduction. In *Proceedings of the ECAI'96 workshop on Abductive and Inductive Reasoning* (pp. 10–13).

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal Logic*. Cambridge University Press.

Bloom, P. & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77, B25–B31.

Booth, R. & Nittka, A. (2008). Reconstructing an agent's epistemic state from observations about its beliefs and non-beliefs. *Logic and Computation*, 18(5), 755–782.

Bordini, R., Dastani, M., Dix, J., & El Fallah Seghrouchni, A., Eds. (2009). *Multi-Agent Programming: Languages, Tools and Applications*. New York, NY: Springer.

Bordini, R. H., Hübner, J. F., & Wooldridge, M. (2007). *Programming Multi-Agent Systems in AgentSpeak using Jason*. Chichester: John Wiley & Sons.

Bosse, T., Memon, Z. A., & Treur, J. (2007a). Emergent stories based on autonomous characters with mindreading capabilities. In T. Y. Lin (Ed.), *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology — IAT* Los Alamitos, CA: IEEE Press.

Bosse, T., Memon, Z. A., & Treur, J. (2007b). Modelling animal behaviour based on interpretation of another animal's behaviour. In R. L. Lewis, T. A. Polk, & J. E. Laird (Eds.), *Proceedings of the Eighth International Conference on Cognitive Modeling — ICCM* (pp. 193–198). Oxford: Taylor & Francis / Psychology Press.

Bosse, T., Memon, Z. A., & Treur, J. (2007c). A two-level BDI-agent model for theory of mind and its use in social manipulation. In *Proceedings of the AISB Workshop on Mindful Environments* (pp. 335–342).

Bosse, T., Memon, Z. A., & Treur, J. (2011). A recursive BDI-agent model for theory of mind and its applications. *Applied Artificial Intelligence*, 25(1), 1–44.

Boutilier, C. (1996). Abduction to plausible causes: An event-based model of belief update. *Artificial Intelligence*, 83, 143–166.

Brafman, R. I. & Tennenholtz, M. (1994). Belief ascription and mental-level modelling. In J. Doyle, E. Sandewall, & P. Torasso (Eds.), *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning — KR* (pp. 87–98). San Francisco, CA: Morgan Kaufmann.

Bratman, M. E. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (1990). What Is Intention? In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in Communication* (pp. 15–31). Cambridge, MA: The MIT Press.

Breazeal, C. L. (2002). *Designing Sociable Robots*. Cambridge, MA: The MIT Press.

Brewka, G., Niemelä, I., & Truszczyński, M. (2008). Nonmonotonic reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Handbook of Knowledge Representation* (pp. 239–284). Amsterdam: Elsevier.

Campos, D. G. (2011). On the distinction between Peirce's abduction and Lipton's Inference to the Best Explanation. *Synthese*, 180, 419–442.

Carberry, S. (2001). Techniques for plan recognition. *User Modeling and User-Adapted Interaction*, 11(1-2), 31–48.

Charniak, E. & Goldman, R. P. (1993). A Bayesian model of plan recognition. *Artificial Intelligence*, 64(1), 53–79.

Cohen, P. R. & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42(2-3), 213–261.

Cohen, P. R., Perrault, R., & Allen, J. F. (1981). Beyond question answering. In W. Lehnert & M. Ringle (Eds.), *Strategies for Natural Language Processing* (pp. 245–274). Hillsdale, NJ: L. Erlbaum Associates.

Cox, P. T. & Pietrzykowski, T. (1986). Causes for events: Their computation and applications. In J. H. Siekmann (Ed.), *Proceedings of the Eighth International Conference on Automated Deduction — CADE*, volume 230 of *LNCS* (pp. 608–621). Heidelberg: Springer.

Dastani, M. (2008). 2APL: A practical agent programming language. *Autonomous Agents and Multi-Agent Systems*, 16, 214–248.

Dastani, M. (2009). Modular rule-based programming in 2APL. In A. Giurca, D. Gašević, & K. Taveter (Eds.), *Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches*, volume I (pp. 25–49). Hershey, PA: IGI Global.

Dautenhahn, K. & Werry, I. (2004). Towards interactive robots in autism therapy. *Pragmatics & Cognition*, 12(1), 1–35.

Dennett, D. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, 1, 568–570.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press.

van Ditmarsch, H. & Kooi, B. (2008). Semantic results for ontic and epistemic change. In G. Bonanno, W. van der Hoek, & M. Wooldridge (Eds.), *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, volume 3 of *Texts in Logic and Games* (pp. 87–117). Amsterdam University Press.

van Ditmarsch, H. & Labuschagne, W. (2007). My beliefs about your beliefs: A case study in theory of mind and epistemic logic. *Synthese*, 155(2), 191–209.

Doirado, E. & Martinho, C. (2010). I mean it! Detecting user intentions to create believable behaviour for virtual agents in games. In W. van der Hoek, G. Kaminka, Y. Lespérance, M. Luck, & S. Sen (Eds.), *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems — AAMAS* (pp. 83–90). Toronto, ON: IFAAMAS.

Dragoni, A. F., Giorgini, P., & Serafini, L. (2002). Mental states recognition from communication. *Logic and Computation*, 12(1), 119–136.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: The MIT Press.

Feil-Seifer, D. & Matarić, M. J. (2005). Defining socially assistive robotics. In *Proceedings of the IEEE Ninth International Conference on Rehabilitation Robotics — ICORR* (pp. 465–468).

Flach, P. A. (1995). *Conjectures: An Inquiry Concerning the Logic of Induction*. PhD thesis, University of Tilburg.

Flach, P. A. & Kakas, A. C. (2000). On the relation between abduction and inductive learning. In D. M. Gabbay & P. Smets (Eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, volume 4 (pp. 1–33). Norwell, MA: Kluwer Academic.

Funge, J. (2004). *Artificial Intelligence for Computer Games: An Introduction*. Natick, MA: A. K. Peters / CRC Press.

Gebser, M., Kaminski, R., Kaufmann, B., Ostrowski, M., Schaub, T., & Thiele, S. (2010). A user's guide to GRINGO, CLASP, CLINGO, and ICLINGO.

Gebser, M., Kaufmann, B., Neumann, A., & Schaub, T. (2007). CLASP: A conflict-driven answer set solver. In *Proc. of the Ninth Intl. Conf. on Logic Programming and Nonmonotonic Reasoning (LPNMR)* (pp. 260–265).

Geib, C. W. & Goldman, R. P. (2001). Plan recognition in intrusion detection systems. In J. Lala, D. Maughan, C. McCollum, & B. Witten (Eds.), *Proceedings of the DARPA Information Survivability Conference and Exposition — DISCEX* (pp. 329–342). Los Alamitos, CA: IEEE Press.

Geib, C. W. & Steedman, M. (2007). On natural language processing and plan recognition. In M. Veloso (Ed.), *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence — IJCAI* (pp. 1612–1617).

Gelfond, M. (2008). Answer sets. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Handbook of Knowledge Representation* (pp. 285–316). Amsterdam: Elsevier.

Gmytrasiewicz, P. J. & Durfee, E. H. (1995). A rigorous, operational formalization of recursive modeling. In V. R. Lesser & L. Gasser (Eds.), *Proceedings of the First International Conference on Multiagent Systems — ICMAS* (pp. 125–132). Cambridge, MA: The MIT Press.

Goultiaeva, A. & Lespérance, Y. (2007). Incremental plan recognition in an agent programming framework. *Proceedings of the AAAI Workshop on Plan, Activity and Intent Recognition — PAIR*, (pp. 52–59).

Harbers, M., van den Bosch, K., & Meyer, J.-J. Ch. (2009). Modeling agents with a theory of mind. In R. Baeza-Yates, J. Lang, S. Mitra, S. Parsons, & G. Pasi (Eds.), *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology — IAT*, volume 2 (pp. 217–224). Los Alamitos, CA: IEEE Press.

Harel, D., Kozen, D., & Tiuryn, J. (2000). *Dynamic Logic*. Cambridge, MA: The MIT Press.

Harman, G. H. (1965). Inference to the best explanation. *The Philosophical Review*, 74, 88–95.

Hartshorne, C., Weiss, P., & Burks, A. W., Eds. (1931–1958). *Collected Papers of Charles Sanders Peirce*, volume I–VIII. Cambridge, MA: Harvard University Press.

Hendriks, M. (2010). A cognitive agent model for mindreading. Master's thesis, Vrije Universiteit Amsterdam.

Hindriks, K. V. (2001). *Agent Programming Languages: Programming with Mental Models*. PhD thesis, University of Utrecht.

Hindriks, K. V. (2009). Programming rational agents in GOAL. In *Multi-Agent Programming: Languages, Tools and Applications* (pp. 119–157). New York, NY: Springer.

Hindriks, K. V., van Riemsdijk, M. B., Behrens, T., Korstanje, R., Kraayenbrink, N., Pasman, W., & de Rijk, L. (2011). UNREAL GOAL bots: Conceptual design of a reusable interface. In F. Dignum (Ed.), *Proceedings of the Second Workshop on Agents for Games and Simulations*, volume 6525 of *LNAI* (pp. 1–18).

van der Hoek, W., van Linder, B., & Meyer, J.-J. Ch. (1998). *An Integrated Modal Approach to Rational Agents*, (pp. 133–168). Dordrecht: Kluwer.

Hoogendoorn, M. & Soumokil, J. (2010). Evaluation of virtual agents utilizing theory of mind in a real time action game. In W. van der Hoek, G. Kaminka, Y. Lespérance, M. Luck, & S. Sen (Eds.), *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems — AAMAS* (pp. 59–66). Toronto, ON: IFAAMAS.

Howlin, P., Baron-Cohen, S., & Hadwin, J. (1999). *Teaching Children with Autism to Mind-Read*. John Wiley & Sons.

Isla, D. & Blumberg, B. (2002). New challenges for character-based AI for games. In *Proc. of the AAAI Spring Symp. on AI and Interactive Entertainment*, number SS-02-01 in AAAI Tech. Rep.

Jeffrey, R. C. (1983). *The Logic of Decision*. University Of Chicago Press.

Kadlec, R., Gemrot, J., Bída, M., Burkert, O., Havlíček, J., Zemčák, L., Píbil, R., Vansa, R., & Brom, C. (2009). Extensions and applications of Pogamut 3 platform. In Zs. Ruttkay, M. Kipp, A. Nijholt, & H. H. Vilhjálmsson (Eds.), *Proceedings of the Ninth International Conference on Intelligent Virtual Agents — IVA* (pp. 506–507). Heidelberg: Springer.

Kahneman, D. & Tversky, A. (1982). The psychology of preferences. *Scientific American*, 246(1).

Kakas, A. C., Kowalski, R. A., & Toni, F. (1998). The role of abduction in logic programming. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 5 (pp. 235–324). Oxford University Press.

Kautz, H. A. (1993). A formal theory of plan recognition and its implementation. In R. J. Brachman (Ed.), *Reasoning About Plans* (pp. 69–126). San Francisco, CA: Morgan Kaufmann.

Kiefer, P. & Schlieder, C. (2007). Exploring context-sensitivity in spatial intention recognition. In *Proceedings of the KI Workshop on Behaviour Monitoring and Interpretation — BMI* (pp. 102–116).

Kott, A. & McEneaney, W. M., Eds. (2006). *Adversarial Reasoning: Computational Approaches to Reading the Opponent's Mind*. Boca Raton, FL: Chapman & Hall / CRC.

Laird, J. (2001). It knows what you're going to do: Adding anticipation to a Quakebot. In *Proceedings of the Fifth International Conference on Autonomous Agents — AGENTS* (pp. 385–392). New York: ACM.

Laird, J. E. & van Lent, M. (2001). Human-level AI's killer application: Interactive computer games. *AAAI Magazine*, (pp. 15–25).

van Lambalgen, M. & Smid, H. (2003). Reasoning patterns in autism: rules and exceptions. In L. A. P. Miranda & J. M. Larrazabal (Eds.), *Proceedings of the International Colloquium on Cognitive Science* Dordrecht: Kluwer.

Leslie, A. (1987). Pretense and representation: The origins of "theory of mind". *Psychological Review*, 94, 412–426.

Leslie, A. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In *Mapping the Mind* (pp. 119–148). Cambridge University Press.

Lidén, L. (2002). Artificial stupidity: The art of intentional mistakes. In S. Rabin (Ed.), *AI Game Programming Wisdom* (pp. 41–48). Charles River Media.

Lifschitz, V. (2008). What is answer set programming? In *Proceedings of the Twenty-Third National Conference on Artificial Intelligence — AAAI* (pp. 1594–1597). Menlo Park, CA: AAAI Press.

Lifschitz, V. (2010). Thirteen definitions of a stable model. In N. Dershowitz, W. Reisig, & A. Blass (Eds.), *Fields of Logic and Computation: Essays Dedicated to Yuri Gurevich on the Occasion of his 70th Birthday* (pp. 488–503). Heidelberg: Springer.

van Linder, B., van der Hoek, W., & Meyer, J.-J. Ch. (1997). Seeing is believing (and so are hearing and jumping). *Logic, Language and Information*, 6, 33–61.

Lipton, P. (2004). *Inference to the Best Explanation*. International Library of Philosophy. Abingdon: Routledge, 2nd edition.

Loyall, A. B. (1997). *Believable Agents: Building Interactive Personalities*. PhD thesis, Carnegie Mellon University.

McCarthy, J. (1979/1990). Ascribing mental qualities to machines. In V. Lifschitz (Ed.), *Formalizing Common Sense: Papers by John McCarthy* (pp. 93–118). Norwood, NJ: Ablex.

Meyer, J.-J. Ch. (2000). Dynamic logic for reasoning about actions and agents. In J. Minker (Ed.), *Logic-Based Artificial Intelligence* (pp. 281–311). Boston/Dordrecht: Kluwer.

Meyer, J.-J. Ch. & van der Hoek, W. (1991). Non-monotonic reasoning by monotonic means. In J. van Eijck (Ed.), *Proceedings of the First European Conference on Logics in Artificial Intelligence (Journées Européennes sur la Logique en Intelligence Artificielle) — JELIA*, volume 478 of *LNAI* (pp. 399–411). Heidelberg: Springer.

Meyer, J.-J. Ch. & van der Hoek, W. (1995). *Epistemic Logic for AI and Computer Science*. Cambridge University Press.

Millington, I. (2006). *Artificial Intelligence for Games*. San Francisco, CA, USA: Morgan Kaufmann.

Nareyek, A. (2007). Game AI is dead. long live game AI! *IEEE Intelligent Systems*, 22(1), 9–11.

Nichols, S. & Stich, S. P. (2003). *Mindreading*. Oxford University Press.

Niemelä, I. (2010). ECAI tutorial on Answer Set Programming.

Niemelä, I., Simons, P., & Syrjänen, T. (2000). SMODELS: A system for answer set programming. In *Proceedings of the Eighth International Workshop on Non-Monotonic Reasoning — NMR*. (`http://www.tcs.hut.fi/Software/smodels/`).

Norling, E. & Sonenberg, L. (2004). Creating interactive characters with BDI agents. In Y. Pisan (Ed.), *Proceedings of the Australian Workshop on Interactive Entertainment — IE2004* (pp. 69–76).

Orkin, J. (2006). Three states and a plan. In *Presentation on the A.I. of F.E.A.R. at the Game Developers' Conference*. (`http://web.media.mit.edu/~jorkin/`).

Pereira, L. M. & Anh, H. T. (2009). Intention recognition via causal bayes networks plus plan generation. In *Proceedings of the Fourteenth Portuguese Conference on Artificial Intelligence — EPIA*, volume 5816 of *LNAI* (pp. 138–149). Heidelberg: Springer.

Pokahr, A., Braubach, L., & Lamersdorf, W. (2005). Jadex: A BDI reasoning engine. In R. Bordini, M. Dastani, J. Dix, & A. El Fallah Seghrouchni (Eds.), *Multi-Agent Programming* (pp. 149–174). New York, NY: Springer.

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a 'theory of mind'? *Behavioral and Brain Sciences*, 4, 515–526.

Pynadath, D. V. & Marsella, S. C. (2005). PsychSim: Modeling theory of mind with decision-theoretic agents. In L. P. Kaelbling & A. Saffiotti (Eds.), *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence — IJCAI* (pp. 1181–1186). Denver, CO: Professional Book Center.

Quaresma, P. & Lopes, J. G. (1995). Unified logic programming approach to the abduction of plans and intentions in information-seeking dialogues. *Logic Programming*, 54, 1–20.

Ramírez, M. & Geffner, H. (2010). Probabilistic plan recognition using off-the-shelf classical planners. In *Proceedings of the Twenty-Fourth National Conference on Artificial Intelligence — AAAI* (pp. 1121–1126). Menlo Park, CA: AAAI Press.

Rao, A. S. (1994). Means-end plan recognition — towards a theory of reactive recognition. In J. Doyle, E. Sandewall, & P. Torasso (Eds.), *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning — KR* (pp. 497–508). San Francisco, CA: Morgan Kaufmann.

Rao, A. S. (1996). AgentSpeak(L): BDI agents speak out in a logical computable language. In W. van de Velde & J. Perram (Eds.), *Agents Breaking Away*, volume 1038 of *LNCS* (pp. 42–55).

Rao, A. S. & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In J. F. Allen, R. Fikes, & E. Sandewall (Eds.), *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning — KR* (pp. 473–484). San Francisco, CA: Morgan Kaufmann.

Rao, A. S. & Georgeff, M. P. (1995). BDI agents: From theory to practice. In V. R. Lesser & L. Gasser (Eds.), *Proceedings of the First International Conference on Multiagent Systems — ICMAS* (pp. 312–319). Cambridge, MA: The MIT Press.

Rao, A. S. & Murray, G. (1994). Multi-agent mental-state recognition and its application to air-combat modelling. In *Proceedings of the Thirteenth International Distributed Artificial Intelligence workshop* (pp. 283–304).

Reiter, R. (1991). The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, (pp. 359–380).

Sadri, F. (2011). Logic-based approaches to intention recognition. In N.-Y. Chong & F. Mastrogiovanni (Eds.), *Handbook of Research on Ambient Intelligence: Trends and Perspectives* (pp. 346–375). Hershey, PA: IGI Global.

Sindlar, M. P., Dastani, M., Dignum, F., & Meyer, J.-J. Ch. (2008). Mental state abduction of BDI-based agents. In M. Baldoni, T. C. Son, M. B. van Riemsdijk, & M. Winikoff (Eds.), *Proceedings of the Sixth International Workshop on Declarative Agent Languages and Technologies — DALT*, volume 5397 of *LNCS* (pp. 161–178). Heidelberg: Springer.

Sindlar, M. P., Dastani, M., Dignum, F., & Meyer, J.-J. Ch. (2009a). Explaining and predicting the behavior of BDI-based agents in role-playing games. In M. Baldoni, J. Bentahar, M. B. van Riemsdijk, & J. Lloyd (Eds.), *Proceedings of the Seventh International Workshop on Declarative Agent Languages and Technologies — DALT*, volume 5948 of *LNCS* (pp. 174–191). Heidelberg: Springer.

Sindlar, M. P., Dastani, M., & Meyer, J.-J. Ch. (2009b). BDI-based development of virtual characters with a theory of mind. In Zs. Ruttkay, M. Kipp, A. Nijholt, & H. H. Vilhjálmsson (Eds.), *Proceedings of the Ninth International Conference on Intelligent Virtual Agents — IVA*, volume 5773 of *LNCS* (pp. 34–41). Heidelberg: Springer.

Sindlar, M. P., Dastani, M., & Meyer, J.-J. Ch. (2010a). A logical account of theory of mind. In V. Goranko & W. Jamroga (Eds.), *Proceedings of the Third Workshop on Logical Aspects of Multi-Agent Systems — LAMAS*.

Sindlar, M. P., Dastani, M., & Meyer, J.-J. Ch. (2010b). Mental state ascription using dynamic logic. In H. Coelho, R. Studer, & M. Wooldridge (Eds.), *Proceedings of the Nineteenth European Conference on Artificial Intelligence — ECAI*, volume 215 of *Frontiers in Artificial Intelligence and Applications* (pp. 561–566). Amsterdam: IOS Press.

Sindlar, M. P., Dastani, M., & Meyer, J.-J. Ch. (2011). Programming mental state abduction. In K. Tumer, P. Yolum, L. Sonenberg, & P. Stone (Eds.), *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems — AAMAS* (pp. 301–308). Taipei: IFAAMAS.

Sindlar, M. P. & Meyer, J.-J. Ch. (2010). Affordance-based intention recognition in virtual spatial environments. In N. Desai, A. Liu, & M. Winikoff (Eds.), *Proceedings of the Thirteenth International Conference on Principles and Practice of Multi-Agent Systems — PRIMA*, volume 7057 of *LNCS* (pp. *to appear*). Heidelberg: Springer.

Singh, M. P., Rao, A. S., & Georgeff, M. P. (1999). Formal methods in DAI: Logic-based representation and reasoning. In *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence* (pp. 331–376). Cambridge, MA: The MIT Press.

Stanford Encyclopedia of Philosophy (2009). Paraconsistent logic. Online text available at `http://plato.stanford.edu/entries/logic-paraconsistent/`.

Stanford Encyclopedia of Philosophy (2010). Intentionality. Online text available at `http://plato.stanford.edu/entries/intentionality/`.

Steunebrink, B. R. (2010). *The Logical Structure of Emotions*. PhD thesis, University of Utrecht.

Turner, H. (1997). Representing actions in logic programs and default theories: A situation calculus approach. *Logic Programming*, 31(1–3), 245–298.

Wahl, S. & Spada, H. (2000). Children's reasoning about intentions, beliefs and behaviour. *Cognitive Science Quarterly*, 1, 5–34.

Wimmer, H. & Perner, J. (1983). Belief about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–128.

Wooldridge, M. & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2), 115–152.

Zhang, D. & Foo, N. (2005). Frame problem in dynamic logic. *Applied Non-Classical Logics*, 15(2), 215–239.

In 's aanschouwers oog — Samenvatting

Dit proefschrift betreft het verklaren van geobserveerd gedrag. Deze activiteit wordt door mensen veelvuldig beoefend en gaat dan veelal gepaard met het veronderstellen van een 'geestesleven' bij de ander, als het verklaarde gedrag dat van een menselijk persoon is. Een synonieme benaming voor een dergelijk geestesleven is *mentale toestand*, wat gezien kan worden als het 'interne' geheel van overtuigingen, wensen, verlangens, emoties, enz., die een persoon erop nahoudt. Eén model wat beschrijft hoe mensen praktisch redeneren — d.w.z. hun gedrag bepalen als uitkomst van een proces waarbij met elementen van de mentale toestand wordt geredeneerd — is het BDI ('Belief-Desire-Intention') model. Dit model van menselijk praktisch redeneren is het onderwerp geweest van *formalisatie*, wat betekent dat met behulp van exacte, wiskundige, methoden (om precies te zijn, bepaalde varianten van *logica*) gepoogd is de regelmatigheden die dit model beschrijft vast te leggen.

Als veronderstelde wetmatigheden van een systeem zijn geformaliseerd, dus in een strikt formaat zijn vastgelegd waarbij geen onduidelijkheid of 'vaagheid' is toegestaan, dan wordt het doorgaans ook makkelijker ze met een computersysteem te verwerken. Een dergelijke verwerking is het maken van kunstmatige intelligentie, wat vrij gedefiniëerd kan worden als "computers die een 'typisch menselijke' eigenschap uitoefenen". In het geval van de formalisaties van het BDI model heeft dit geleid tot het ontwerp en de bouw van zogeheten *autonome software-agenten*.[1] Dit type computersoftware wordt beschreven, of zelfs direct geprogrammeerd, in termen van een mentale toestand en is in staat zelfstandig zijn gedrag te bepalen volgens de beginselen van het BDI model. In het geval dat zo een software-agent direct in 'mentale termen' wordt geprogrammeerd kan deze agent de beschikking hebben over *informatie* over zijn leefomgeving (de 'Beliefs', in verwijzing naar 'BDI'), alsmede over *doelen* welke toestanden in zijn leefomgeving beschrijven die hij probeert te verwezenlijken ('Goals') en *plannen* die gekoppeld zijn aan gedragsregels die bepalen onder welke omstandigheden een plan geschikt is om een doel te bereiken. Als de agent een bepaald plan geschikt acht om een doel wat hij heeft te verwezenlijken, kan de agent dit plan aannemen en wordt het in de vakliteratuur vaak een *intentie* genoemd ('Intention').

Het leeuwendeel van dit proefschrift betreft het verklaren van gedrag van BDI-gebaseerde software-agenten, welke direct geprogrammeerd zijn in termen van hun mentale toestand. Tot dit doeleinde worden verschillende formele (logische) redeneertechnieken gebruikt, om uit te drukken welke mentale toestand(en) de geobserveerde agent *mogelijk* zou kunnen hebben. Eén van deze technieken is niet-monotone logica, waarbij het concept 'mogelijkheid' wordt uitgedrukt als een uitbreiding op zaken die voor 'zeker waar' worden gehouden met zaken die voor 'mogelijk waar' worden gehouden. De andere gebruikte techniek is modale (dynamische) logica, waarbij datzelfde concept wordt uitgedrukt m.b.v. een modaliteit die een bepaald feit kwalificeert als zijnde 'waar, in een bepaalde voor mogelijk

[1] Hier heeft 'agent' de betekenis van 'actor', niet van 'politieagent'.

gehouden toestand'. Een beoogde toepassing van de geopperde technieken is het maken van meer intelligente *virtuele karakters*[2] voor computertoepassingen zoals spellen, of virtuele trainingsomgevingen. Gezien het feit dat het BDI model rekening houdt met een begrensd redeneervermogen van de agent, alsmede het feit dat de basisconcepten van dit model — de Belief-, Desire-, en Intention-eenheden van de mentale toestand — zich bij uitstek lenen voor beschrijving van complex sociaal gedrag, is het aannemelijk te maken dat deze software-benadering geschikt is voor het ontwerpen en/of programmeren van virtuele karakters die dergelijk gedrag vertonen (zoals bijvoorbeeld het geval is in zogenaamde *rollenspellen*, ofwel 'role-playing games'). Als dergelijke karakters in staat zijn om hun eigen gedrag te bepalen als gevolg van redeneren over hun directe omgeving, maar in het bijzonder de veronderstelde mentale toestand van anderen in hun omgeving, zo luidt onze redenatie, dan zal dit hun kunstmatige intelligentie ten goede komen.

De rode draad die door dit proefschrift loopt is dat het uiteindelijk aan de *aanschouwer* — degene die gedrag observeert en verklaart — is om te bepalen welke mentale toestand aan de geobserveerde agent wordt toegeschreven. Dit wordt bedoeld als wij zeggen dat deze verklaringen "in 's aanschouwers oog" zijn.[3] Als er geredeneerd wordt over het gedrag van formeel beschreven software-agenten dan kunnen (in sommige gevallen) hun mogelijke mentale toestanden precies bepaald worden. Bij de in voorgaande paragraaf beschreven technieken gebeurt deze bepaling door de aanschouwer kennis te geven van de gedragsregels die de agent tot zijn beschikking heeft. Er zijn ook situaties denkbaar waarin het lastig is de mogelijke mentale toestanden van de geobserveerde agent precies te bepalen. Dit is zo als men geen weet heeft van de interne werking van de agent, of als de geobserveerde agent zodanig van aard is dat inspectie van diens geestesleven niet mogelijk is. Dat laatste is bijvoorbeeld het geval bij redenatie over menselijke spelers in een virtuele omgeving. Een aanschouwer kan dan terugvallen op verbanden die *verondersteld* worden te bestaan tussen observeerbare feiten (zoals uitgevoerde acties) en niet-observeerbare feiten (zoals elementen van een mentale toestand die wordt toegeschreven aan de agent). Het laatste deel van dit proefschrift is gericht op het bestuderen van de logische vorm die dergelijke verbanden kunnen hebben, aan de hand van bestaande psychologische modellen. Deze aanpak resulteert in een beschrijving van deze verbanden met hetzelfde modaal-logische 'vocabulaire' dat eerder werd gebruikt door de aanschouwer om uitspraken te doen over de mentale toestand van geobserveerde software-agenten, en culmineert in een gedetailleerde toepassing op het redeneren van een aanschouwer over gedrag wat wordt geobserveerd in een virtuele tegenhanger van de *false-belief task* uit de psychologie.

Het in voorgaande paragrafen beschreven werk is theoretisch van aard, in de zin dat het is gericht op formalisatie (zoals omschreven in de eerste paragraaf) van gedragsverklaring door toeschrijving van een mentale toestand. Om *implementatie* (het 'in praktijk brengen') van deze theorie te bemogelijken is het slotdeel van dit proefschrift gericht op een vertaling van onze aanpak naar een daarvoor geschikt geachte taal voor logisch programmeren. Tot slot worden verbanden gelegd met bestaand gerelateerd werk, ter inbedding van ons eigen werk in een breder kader en om kruisbestuiving te bevorderen.

---

[2]Hier heeft 'karakter' de betekenis van 'personage', niet van 'letterteken'.

[3]Onze vertaling van '*in the eye of the beholder*', vaak onderdeel van "beauty is in the eye of the beholder".

SIKS Dissertation Series

1998

Johan van den Akker, CWI (01). *DEGAS — An Active, Temporal Database of Autonomous Objects*

Floris Wiesman, UM (02). *Information Retrieval by Graphically Browsing Meta-Information*

Ans Steuten, TUD (03). *A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective*

Dennis Breuker, UM (04). *Memory versus Search in Games*

E.W. Oskamp, RUL (05). *Computerondersteuning bij Straftoemeting*

1999

Mark Sloof, VU (01). *Physiology of Quality Change Modelling: Automated modelling of Quality Change of Agricultural Products*

Rob Potharst, EUR (02). *Classification using Decision Trees and Neural Nets*

Don Beal, UM (03). *The Nature of Minimax Search*

Jacques Penders, UM (04). *The Practical Art of Moving Physical Objects*

Aldo de Moor, KUB (05). *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems*

Niek J.E. Wijngaards, VU (06). *Re-design of compositional systems*

David Spelt, UT (07). *Verification support for object database design*

Jacques H.J. Lenting, UM (08). *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation*

2000

Frank Niessink, VU (01). *Perspectives on Improving Software Maintenance*

Koen Holtman, TUE (02). *Prototyping of CMS Storage Management*

Carolien M.T. Metselaar, UVA (03). *Sociaal-organisatorische gevolgen van kennistechnologie: een procesbenadering en actorperspectief*

Geert de Haan, VU (04). *ETAG, A Formal Model of Competence Knowledge for User Interface Design*

Ruud van der Pol, UM (05). *Knowledge-based Query Formulation in Information Retrieval*

Rogier van Eijk, UU (06). *Programming Languages for Agent Communication*

Niels Peek, UU (07). *Decision-theoretic Planning of Clinical Patient Management*

Veerle Coup, EUR (08). *Sensitivity Analyis of Decision-Theoretic Networks*

Florian Waas, CWI (09). *Principles of Probabilistic Query Optimization*

Niels Nes, CWI (10). *Image Database Management System Design Considerations, Algorithms and Architecture*

Jonas Karlsson, CWI (11). *Scalable Distributed Data Structures for Database Management*

2001

Silja Renooij, UU (01). *Qualitative Approaches to Quantifying Probabilistic Networks*

Koen Hindriks, UU (02). *Agent Programming Languages: Programming with Mental Models*

Maarten van Someren, UvA (03). *Learning as problem solving*

Evgueni Smirnov, UM (04). *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets*

Jacco van Ossenbruggen, VU (05). *Processing Structured Hypermedia: A Matter of Style*

Martijn van Welie, VU (06). *Task-based User Interface Design*

Bastiaan Schonhage, VU (07). *Diva: Architectural Perspectives on Information Visualization*

Pascal van Eck, VU (08). *A Compositional Semantic Structure for Multi-Agent Systems Dynamics*

Pieter Jan 't Hoen, RUL (09). *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes*

Maarten Sierhuis, UvA (10). *Modeling and Simulating Work Practice. BRAHMS: A Multiagent Modeling and Simulation Language for Work Practice Analysis and Design*

Tom van Engers, VUA (11). *Knowledge Management: The Role of Mental Models in Business Systems Design*

2002

Nico Lassing, VU (01). *Architecture-Level Modifiability Analysis*

Roelof van Zwol, UT (02). *Modelling and searching web-based document collections*

Henk Ernst Blok, UT (03). *Database Optimization Aspects for Information Retrieval*

Juan Roberto Castelo Valdueza, UU (04). *The Discrete Acyclic Digraph Markov Model in Data Mining*

Radu Serban, VU (05). *The Private Cyberspace: Modeling Electronic Environments inhabited by Privacy-concerned Agents*

Laurens Mommers, UL (06). *Applied legal epistemology: Building a knowledge-based ontology of the legal domain*

Peter Boncz, CWI (07). *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications*

Jaap Gordijn, VU (08). *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas*

Willem-Jan van den Heuvel, KUB (09). *Integrating Modern Business Applications with Objectified Legacy Systems*

Brian Sheppard, UM (10). *Towards Perfect Play of Scrabble*

Wouter C.A. Wijngaards, VU (11). *Agent Based Modelling of Dynamics: Biological and Organisational Applications*

Albrecht Schmidt, UVA (12). *Processing XML in Database Systems*

Hongjing Wu, TUE (13). *A Reference Architecture for Adaptive Hypermedia Applications*

Wieke de Vries, UU (14). *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems*

Rik Eshuis, UT (15). *Semantics and Verification of UML Activity Diagrams for Workflow Modelling*

Pieter van Langen, VU (16). *The Anatomy of Design: Foundations, Models and Applications*

Stefan Manegold, UVA (17). *Understanding, Modeling, and Improving Main-Memory Database Performance*

2003

Heiner Stuckenschmidt, VU (01). *Ontology-Based Information Sharing In Weakly Structured Environments*

Jan Broersen, VU (02). *Modal Action Logics for Reasoning About Reactive Systems*

Martijn Schuemie, TUD (03). *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy*

Milan Petkovic, UT (04). *Content-Based Video Retrieval Supported by Database Technology*

Jos Lehmann, UVA (05). *Causation in Artificial Intelligence and Law - A modelling approach*

Boris van Schooten, UT (06). *Development and specification of virtual environments*

Machiel Jansen, UvA (07). *Formal Explorations of Knowledge Intensive Tasks*

Yongping Ran, UM (08). *Repair Based Scheduling*

Rens Kortmann, UM (09). *The resolution of visually guided behaviour*

Andreas Lincke, UvT (10). *Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture*

Simon Keizer, UT (11). *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*

Roeland Ordelman, UT (12). *Dutch speech recognition in multimedia information retrieval*

Jeroen Donkers, UM (13). *Nosce Hostem - Searching with Opponent Models*

Stijn Hoppenbrouwers, KUN (14). *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations*

Mathijs de Weerdt, TUD (15). *Plan Merging in Multi-Agent Systems*

Menzo Windhouwer, CWI (16). *Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses*

David Jansen, UT (17). *Extensions of Statecharts with Probability, Time, and Stochastic Timing*

Levente Kocsis, UM (18). *Learning Search Decisions*

2004

Virginia Dignum, UU (01). *A Model for Organizational Interaction: Based on Agents, Founded in Logic*

Lai Xu, UvT (02). *Monitoring Multi-party Contracts for E-business*

Perry Groot, VU (03). *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving*

Chris van Aart, UVA (04). *Organizational Principles for Multi-Agent Architectures*

Viara Popova, EUR (05). *Knowledge discovery and monotonicity*

Bart-Jan Hommes, TUD (06). *The Evaluation of Business Process Modeling Techniques*

Elise Boltjes, UM (07). *Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes*

Joop Verbeek, UM (08). *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise*

Martin Caminada, VU (09). *For the Sake of the Argument: explorations into argument-based reasoning*

Suzanne Kabel, UVA (10). *Knowledge-rich indexing of learning-objects*

Michel Klein, VU (11). *Change Management for Distributed Ontologies*

The Duy Bui, UT (12). *Creating emotions and facial expressions for embodied agents*

Wojciech Jamroga, UT (13). *Using Multiple Models of Reality: On Agents who Know how to Play*

Paul Harrenstein, UU (14). *Logic in Conflict. Logical Explorations in Strategic Equilibrium*

Arno Knobbe, UU (15). *Multi-Relational Data Mining*

Federico Divina, VU (16). *Hybrid Genetic Relational Search for Inductive Learning*

Mark Winands, UM (17). *Informed Search in Complex Games*

Vania Bessa Machado, UvA (18). *Supporting the Construction of Qualitative Knowledge Models*

Thijs Westerveld, UT (19). *Using generative probabilistic models for multimedia retrieval*

Madelon Evers, Nyenrode (20). *Learning from Design: Facilitating Multidisciplinary Design Teams*

2005

Floor Verdenius, UVA (01). *Methodological Aspects of Designing Induction-Based Applications*

Erik van der Werf, UM (02). *AI techniques for the game of Go*

Franc Grootjen, RUN (03). *A Pragmatic Approach to the Conceptualisation of Language*

Nirvana Meratnia, UT (04). *Towards Database Support for Moving Object Data*

Gabriel Infante-Lopez, UVA (05). *Two-Level Probabilistic Grammars for Natural Language Parsing*

Pieter Spronck, UM (06). *Adaptive Game AI*

Flavius Frasincar, TUE (07). *Hypermedia Presentation Generation for Semantic Web Information Systems*

Richard Vdovjak, TUE (08). *A Model-driven Approach for Building Distributed Ontology-based Web Applications*

Jeen Broekstra, VU (09). *Storage, Querying and Inferencing for Semantic Web Languages*

Anders Bouwer, UVA (10). *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments*

Elth Ogston, VU (11). *Agent Based Matchmaking and Clustering — A Decentralized Approach to Search*

Csaba Boer, EUR (12). *Distributed Simulation in Industry*

Fred Hamburg, UL (13). *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen*

Borys Omelayenko, VU (14). *Web-Service configuration on the Semantic Web: Exploring How Semantics Meets Pragmatics*

Tibor Bosse, VU (15). *Analysis of the Dynamics of Cognitive Processes*

Joris Graaumans, UU (16). *Usability of XML Query Languages*

Boris Shishkov, TUD (17). *Software Specification Based on Re-usable Business Components*

Danielle Sent, UU (18). *Test-selection strategies for probabilistic networks*

Michel van Dartel, UM (19). *Situated Representation*

Cristina Coteanu, UL (20). *Cyber Consumer Law, State of the Art and Perspectives*

Wijnand Derks, UT (21). *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics*

2006

Samuil Angelov, TUE (01). *Foundations of B2B Electronic Contracting*

Cristina Chisalita, VU (02). *Contextual issues in the design and use of information technology in organizations*

Noor Christoph, UVA (03). *The role of metacognitive skills in learning to solve problems*

Marta Sabou, VU (04). *Building Web Service Ontologies*

Cees Pierik, UU (05). *Validation Techniques for Object-Oriented Proof Outlines*

Ziv Baida, VU (06). *Software-aided Service Bundling — Intelligent Methods & Tools for Graphical Service Modeling*

Marko Smiljanic, UT (07). *XML schema matching — balancing efficiency and effectiveness by means of clustering*

Eelco Herder, UT (08). *Forward, Back and Home Again — Analyzing User Behavior on the Web*

Mohamed Wahdan, UM (09). *Automatic Formulation of the Auditor's Opinion*

Ronny Siebes, VU (10). *Semantic Routing in Peer-to-Peer Systems*

Joeri van Ruth, UT (11). *Flattening Queries over Nested Data Types*

Bert Bongers, VU (12). *Interactivation – Towards an e-cology of people, our technological environment, and the arts*

Henk-Jan Lebbink, UU (13). *Dialogue and Decision Games for Information Exchanging Agents*

Johan Hoorn, VU (14). *Software Requirements: Update, Upgrade, Redesign. Towards a Theory of Requirements Change*

Rainer Malik, UU (15). *CONAN: Text Mining in the Biomedical Domain*

Carsten Riggelsen, UU (16). *Approximation Methods for Efficient Learning of Bayesian Networks*

Stacey Nagata, UU (17). *User Assistance for Multitasking with Interruptions on a Mobile Device*

Valentin Zhizhkun, UVA (18). *Graph Transformation for Natural Language Processing*

Birna van Riemsdijk, UU (19). *Cognitive Agent Programming: A Semantic Approach*

Marina Velikova, UvT (20). *Monotone models for prediction in data mining*

Bas van Gils, RUN (21). *Aptness on the Web*

Paul de Vrieze, RUN (22). *Fundaments of Adaptive Personalisation*

Ion Juvina, UU (23). *Development of a Cognitive Model for Navigating on the Web*

Laura Hollink, VU (24). *Semantic Annotation for Retrieval of Visual Resources*

Madalina Drugan, UU (25). *Conditional log-likelihood MDL and Evolutionary MCMC*

Vojkan Mihajlovic, UT (26). *Score Region Algebra: A Flexible Framework for Structured Information Retrieval*

Stefano Bocconi, CWI (27). *Vox Populi: generating video documentaries from semantically annotated media repositories*

Borkur Sigurbjornsson, UVA (28). *Focused Information Access using XML Element Retrieval*

2007

Kees Leune, UvT (01). *Access Control and Service-Oriented Architectures*

Wouter Teepe, RUG (02). *Reconciling Information Exchange and Confidentiality: A Formal Approach*

Peter Mika, VU (03). *Social Networks and the Semantic Web*

Jurriaan van Diggelen, UU (04). *Achieving Semantic Interoperability in Multi-agent Systems: A Dialogue-based Approach*

Bart Schermer, UL (05). *Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance*

Gilad Mishne, UVA (06). *Applied Text Analytics for Blogs*

Natasa Jovanovic, UT (07). *To Who It May Concern — Addressee Identification in Face-to-Face Meetings*

Mark Hoogendoorn, VU (08). *Modeling of Change in Multi-Agent Organizations*

David Mobach, VU (09). *Agent-Based Mediated Service Negotiation*

Huib Aldewereld, UU (10). *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols*

Natalia Stash, TUE (11). *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System*

Marcel van Gerven, RUN (12). *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty*

Rutger Rienks, UT (13). *Meetings in Smart Environments; Implications of Progressing Technology*

Niek Bergboer, UM (14). *Context-Based Image Analysis*

Joyca Lacroix, UM (15). *NIM: a Situated Computational Memory Model*

Davide Grossi, UU (16). *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems*

Theodore Charitos, UU (17). *Reasoning with Dynamic Networks in Practice*

Bart Orriens, UvT (18). *On the development and management of adaptive business collaborations*

David Levy, UM (19). *Intimate relationships with artificial partners*

Slinger Jansen, UU (20). *Customer Configuration Updating in a Software Supply Network*

Karianne Vermaas, UU (21). *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005*

Zlatko Zlatev, UT (22). *Goal-oriented design of value and process models from patterns*

Peter Barna, TUE (23). *Specification of Application Logic in Web Information Systems*

Georgina Ramírez Camps, CWI (24). *Structural Features in XML Retrieval*

Joost Schalken, VU (25). *Empirical Investigations in Software Process Improvement*

2008

Katalin Boer-Sorbán, EUR (01). *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach*

Alexei Sharpanskykh, VU (02). *On Computer-Aided Methods for Modeling and Analysis of Organizations*

Vera Hollink, UVA (03). *Optimizing Hierarchical Menus: A Usage-Based Approach*

Ander de Keijzer, UT (04). *Management of Uncertain Data — towards unattended integration*

Bela Mutschler, UT (05). *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective*

Arjen Hommersom, RUN (06). *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective*

Peter van Rosmalen, OU (07). *Supporting the tutor in the design and support of adaptive e-learning*

Janneke Bolt, UU (08). *Bayesian Networks: Aspects of Approximate Inference*

Christof van Nimwegen, UU (09). *The paradox of the guided user: assistance can be counter-effective*

Wauter Bosma, UT (10). *Discourse oriented summarization*

Vera Kartseva, VU (11). *Designing Controls for Network Organizations: A Value-Based Approach*

Jozsef Farkas, RUN (12). *A Semiotically Oriented Cognitive Model of Knowledge Representation*

Caterina Carraciolo, UVA (13). *Topic Driven Access to Scientific Handbooks*

Arthur van Bunningen, UT (14). *Context-Aware Querying: Better Answers with Less Effort*

Martijn van Otterlo, UT (15). *The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains*

Henriette van Vugt, VU (16). *Embodied agents from a user's perspective*

Martin Op 't Land, TUD (17). *Applying Architecture and Ontology to the Splitting and Allying of Enterprises*

Guido de Croon, UM (18). *Adaptive Active Vision*

Henning Rode, UT (19). *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search*

Rex Arendsen, UVA (20). *Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven*

Krisztian Balog, UVA (21). *People Search in the Enterprise*

Henk Koning, UU (22). *Communication of IT-Architecture*

Stefan Visscher, UU (23). *Bayesian network models for the management of ventilator-associated pneumonia*

Zharko Aleksovski, VU (24). *Using background knowledge in ontology matching*

Geert Jonker, UU (25). *Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency*

Marijn Huijbregts, UT (26). *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*

Hubert Vogten, OU (27). *Design and Implementation Strategies for IMS Learning Design*

Ildiko Flesch, RUN (28). *On the Use of Independence Relations in Bayesian Networks*

Dennis Reidsma, UT (29). *Annotations and Subjective Machines — Of Annotators, Embodied Agents, Users, and Other Humans*

Wouter van Atteveldt, VU (30). *Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content*

Loes Braun, UM (31). *Pro-Active Medical Information Retrieval*

Trung H. Bui, UT (32). *Toward Affective Dialogue*

Management using Partially Observable Markov Decision Processes

Frank Terpstra, UVA (33). *Scientific Workflow Design: theoretical and practical issues*

Jeroen de Knijf, UU (34). *Studies in Frequent Tree Mining*

Ben Torben Nielsen, UvT (35). *Dendritic morphologies: function shapes structure*

2009

Rasa Jurgelenaite, RUN (01). *Symmetric Causal Independence Models*

Willem Robert van Hage, VU (02). *Evaluating Ontology-Alignment Techniques*

Hans Stol, UvT (03). *A Framework for Evidence-based Policy Making Using IT*

Josephine Nabukenya, RUN (04). *Improving the Quality of Organisational Policy Making using Collaboration Engineering*

Sietse Overbeek, RUN (05). *Bridging Supply and Demand for Knowledge Intensive Tasks — Based on Knowledge, Cognition, and Quality*

Muhammad Subianto, UU (06). *Understanding Classification*

Ronald Poppe, UT (07). *Discriminative Vision-Based Recovery and Recognition of Human Motion*

Volker Nannen, VU (08). *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*

Benjamin Kanagwa, RUN (09). *Design, Discovery and Construction of Service-oriented Systems*

Jan Wielemaker, UVA (10). *Logic programming for knowledge-intensive interactive applications*

Alexander Boer, UVA (11). *Legal Theory, Sources of Law & the Semantic Web*

Peter Massuthe, TUE, Humboldt-Universität zu Berlin (12). *Operating Guidelines for Services*

Steven de Jong, UM (13). *Fairness in Multi-Agent Systems*

Maksym Korotkiy, VU (14). *From Ontology-enabled Services to Service-enabled Ontologies: Making Ontologies Work in e-Science with ONTO-SOA*

Rinke Hoekstra, UVA (15). *Ontology Representation — Design Patterns and Ontologies that Make Sense*

Fritz Reul, UvT (16). *New Architectures in Computer Chess*

Laurens van der Maaten, UvT (17). *Feature Extraction from Visual Data*

Fabian Groffen, CWI (18). *Armada, An Evolving Database System*

Valentin Robu, CWI (19). *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*

Bob van der Vecht, UU (20). *Adjustable Autonomy: Controling Influences on Decision Making*

Stijn Vanderlooy, UM (21). *Ranking and Reliable Classification*

Pavel Serdyukov, UT (22). *Search For Expertise: Going beyond direct evidence*

Peter Hofgesang, VU (23). *Modelling Web Usage in a Changing Environment*

Annerieke Heuvelink, VUA (24). *Cognitive Models for Training Simulations*

Alex van Ballegooij, CWI (25). *RAM: Array Database Management through Relational Mapping*

Fernando Koch, UU (26). *An Agent-Based Model for the Development of Intelligent Mobile Services*

Christian Glahn, OU (27). *Contextual Support of Social Engagement and Reflection on the Web*

Sander Evers, UT (28). *Sensor Data Management with Probabilistic Models*

Stanislav Pokraev, UT (29). *Model-Driven Semantic Integration of Service-Oriented Applications*

Marcin Zukowski, CWI (30). *Balancing vectorized query execution with bandwidth-optimized storage*

Sofiya Katrenko, UVA (31). *A Closer Look at Learning Relations from Text*

Rik Farenhorst and Remco de Boer, VU (32). *Architectural Knowledge Management: Supporting Architects and Auditors*

Khiet Truong, UT (33). *How Does Real Affect Affect Affect Recognition In Speech?*

Inge van de Weerd, UU (34). *Advancing in Software Product Management: An Incremental Method Engineering Approach*

Wouter Koelewijn, UL (35). *Privacy en Politiegegevens: Over geautomatiseerde normatieve informatie-uitwisseling*

Marco Kalz, OUN (36). *Placement Support for Learners in Learning Networks*

Hendrik Drachsler, OUN (37). *Navigation Support for Learners in Informal Learning Networks*

Riina Vuorikari, OU (38). *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*

Christian Stahl, TUE, Humboldt-Universität zu

Berlin (39). *Service Substitution — A Behavioral Approach Based on Petri Nets*

Stephan Raaijmakers, UvT (40). *Multinomial Language Learning: Investigations into the Geometry of Language*

Igor Berezhnyy, UvT (41). *Digital Analysis of Paintings*

Toine Bogers, UvT (42). *Recommender Systems for Social Bookmarking*

Virginia Nunes Leal Franqueira, UT (43). *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*

Roberto Santana Tapia, UT (44). *Assessing Business-IT Alignment in Networked Organizations*

Jilles Vreeken, UU (45). *Making Pattern Mining Useful*

Loredana Afanasiev, UvA (46). *Querying XML: Benchmarks and Recursion*

2010

Matthijs van Leeuwen, UU (01). *Patterns that Matter*

Ingo Wassink, UT (02). *Work flows in Life Science*

Joost Geurts, CWI (03). *A Document Engineering Model and Processing Framework for Multimedia Documents*

Olga Kulyk, UT (04). *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*

Claudia Hauff, UT (05). *Predicting the Effectiveness of Queries and Retrieval Systems*

Sander Bakkes, UvT (06). *Rapid Adaptation of Video Game AI*

Wim Fikkert, UT (07). *Gesture Interaction at a Distance*

Krzysztof Siewicz, UL (08). *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*

Hugo Kielman, UL (09). *Politiële gegevensverwerking en Privacy, Naar een effectieve waarborging*

Rebecca Ong, UL (10). *Mobile Communication and Protection of Children*

Adriaan Ter Mors, TUD (11). *The world according to MARP: Multi-Agent Route Planning*

Susan van den Braak, UU (12). *Sensemaking software for crime analysis*

Gianluigi Folino, RUN (13). *High Performance Data Mining using Bio-inspired techniques*

Sander van Splunter, VU (14). *Automated Web Service Reconfiguration*

Lianne Bodenstaff, UT (15). *Managing Dependency Relations in Inter-Organizational Models*

Sicco Verwer, TUD (16). *Efficient Identification of Timed Automata: Theory and Practice*

Spyros Kotoulas, VU (17). *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*

Charlotte Gerritsen, VU (18). *Caught in the Act: Investigating Crime by Agent-Based Simulation*

Henriette Cramer, UvA (19). *People's Responses to Autonomous and Adaptive Systems*

Ivo Swartjes, UT (20). *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*

Harold van Heerde, UT (21). *Privacy-aware data management by means of data degradation*

Michiel Hildebrand, CWI (22). *End-user Support for Access to Heterogeneous Linked Data*

Bas Steunebrink, UU (23). *The Logical Structure of Emotions*

Dmytro Tykhonov, (24). *Designing Generic and Efficient Negotiation Strategies*

Zulfiqar Ali Memon, VU (25). *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*

Ying Zhang, CWI (26). *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*

Marten Voulon, UL (27). *Automatisch contracteren*

Arne Koopman, UU (28). *Characteristic Relational Patterns*

Stratos Idreos, CWI (29). *Database Cracking: Towards Auto-tuning Database Kernels*

Marieke van Erp, UvT (30). *Accessing Natural History: Discoveries in data cleaning, structuring, and retrieval*

Victor de Boer, UVA (31). *Ontology Enrichment from Heterogeneous Sources on the Web*

Marcel Hiel, UvT (32). *An Adaptive Service Oriented Architecture: Automatically Solving Interoperability Problems*

Robin Aly, UT (33). *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*

Teduh Dirgahayu, UT (34). *Interaction Design in Service Compositions*

Dolf Trieschnigg, UT (35). *Proof of Concept: Concept-based Biomedical Information Retrieval*

José Janssen, OU (36). *Paving the Way for Lifelong Learning: Facilitating competence development through a learning path specification*

Niels Lohmann, TUE (37). *Correctness of services and their composition*

Dirk Fahland, TUE (38). *From Scenarios to Components*

Ghazanfar Farooq Siddiqui, VU (39). *Integrative modeling of emotions in virtual agents*

Mark van Assem, VU (40). *Converting and Integrating Vocabularies for the Semantic Web*

Guillaume Chaslot, UM (41). *Monte-Carlo Tree Search*

Sybren de Kinderen, VU (42). *Needs-driven service bundling in a multi-supplier setting — The computational e3-service approach*

Peter van Kranenburg, UU (43). *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*

Pieter Bellekens, TUE (44). *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*

Vasilios Andrikopoulos, UvT (45). *A theory and model for the evolution of software services*

Vincent Pijpers, VU (46). *e3alignment: Exploring Inter-Organizational Business-ICT Alignment*

Chen Li, UT (47). *Mining Process Model Variants: Challenges, Techniques, Examples*

Milan Lovric, EUR (48). *Behavioral Finance and Agent-Based Artificial Markets*

Jahn-Takeshi Saito, UM (49). *Solving difficult game positions*

Bouke Huurnink, UVA (50). *Search in Audiovisual Broadcast Archives*

Alia Khairia Amin, CWI (51). *Understanding and supporting information seeking tasks in multiple sources*

Peter-Paul van Maanen, VU (52). *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*

Edgar Meij, UVA (53). *Combining Concepts and Language Models for Information Access*

2011

Botond Cseke, RUN (01). *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*

Nick Tinnemeier, UU (02). *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*

Jan Martijn van der Werf, TUE (03). *Compositional Design and Verification of Component-Based Information Systems*

Hado van Hasselt, UU (04). *Insights in Reinforcement Learning — Formal analysis and empirical evaluation of temporal-difference learning algorithms*

Base van der Raadt, VU (05). *Enterprise Architecture Coming of Age — Increasing the Performance of an Emerging Discipline*

Yiwen Wang, TUE (06). *Semantically-Enhanced Recommendations in Cultural Heritage*

Yujia Cao, UT (07). *Multimodal Information Presentation for High Load Human Computer Interaction*

Nieske Vergunst, UU (08). *BDI-based Generation of Robust Task-Oriented Dialogues*

Tim de Jong, OU (09). *Contextualised Mobile Media for Learning*

Bart Bogaert, UvT (10). *Cloud Content Contention*

Dhaval Vyas, UT (11). *Designing for Awareness: An Experience-focused HCI Perspective*

Carmen Bratosin, TUE (12). *Grid Architecture for Distributed Process Mining*

Xiaoyu Mao, UvT (13). *Airport under Control: Multi-agent Scheduling for Airport Ground Handling*

Milan Lovric, EUR (14). *Behavioral Finance and Agent-Based Artificial Markets*

Marijn Koolen, UVA (15). *The Meaning of Structure: The Value of Link Evidence for Information Retrieval*

Maarten Schadd, UM (16). *Selective Search in Games of Different Complexity*

Jiyin He, UVA (17). *Exploring Topic Structure: Coherence, Diversity and Relatedness*

Mark Ponsen, UM (18). *Strategic Decision-Making in Complex Games*

Ellen Rusman, OU (19). *The Mind's Eye on Personal Profiles*

Qing Gu, VU (20). *Guiding service-oriented software engineering — A view-based approach*

Linda Terlouw, TUD (21). *Modularization and Specification of Service-Oriented Systems*

Junte Zhang, UVA (22). *System Evaluation of Archival Description and Access*

Wouter Weerkamp, UVA (23). *Finding People and their Utterances in Social Media*

Herwin van Welbergen, UT (24). *Behavior Generation for Interpersonal Coordination with Virtual Humans. On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*

Syed Waqar ul Qounain Jaffry, VU (25). *Analysis and Validation of Models for Trust Dynamics*

Matthijs Aart Pontier, VU (26). *Virtual Agents for Human Communication*

Aniel Bhulai, VU (27). *Dynamic website optimization through autonomous management of design patterns*

Rianne Kaptein, UVA (28). *Effective Focused Retrieval by Exploiting Query Context and Document Structure*

Faisal Kamiran, TUE (29). *Discrimination-aware Classification*

Egon van den Broek, UT (30). *Affective Signal Processing (ASP): Unraveling the mystery of emotions*

Ludo Waltman, EUR (31). *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*

Nees Jan van Eck, EUR (32). *Methodological Advances in Bibliometric Mapping of Science*

Tom van der Weide, UU (33). *Arguing to Motivate Decisions*

Paolo Turrini, UU (34). *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*

Maaike Harbers, UU (35). *Explaining Agent Behavior in Virtual Training*

Erik van der Spek, UU (36). *Experiments in Serious Game Design: A Cognitive Approach*

Adriana Burlutiu, RUN (37). *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*

Nyree Lemmens, UM (38). *Bee-inspired Distributed Optimization*

Joost Westra, UU (39). *Organizing Adaptation using Agents in Serious Games*

Viktor Clerc, VU (40). *Architectural Knowledge Management in Global Software Development*

Luan Ibraimi, UT (41). *Cryptographically Enforced Distributed Data Access Control*

Michal Sindlar, UU (42). *In the Eye of the Beholder: Explaining Behavior through Mental State Attribution*