Logic and **L**ogical **P**hilosophy
Volume 17 (2008), 185–207
DOI: 10.12775/LLP.2008.011

Caroline Semmling and Heinrich Wansing

FROM *BDI* AND *stit* TO *bdi-stit* LOGIC*

Abstract. Since it is desirable to be able to talk about rational agents forming attitudes toward their concrete agency, we suggest an introduction of doxastic, volitional, and intentional modalities into the multi-agent logic of *deliberatively seeing to it that*, *dstit* logic. These modalities are borrowed from the well-known *BDI* (belief-desire-intention) logic. We change the semantics of the belief and desire operators from a relational one to a monotonic neighbourhood semantic in order to handle ascriptions of conflicting but not inconsistent beliefs and desires as being satisfiable. The proposed *bdi-stit* logic is defined with respect to branching time frames, and it is shown that this logic is a generalization of a *bdi* logic based on branching time possible worlds frames (but without temporal operators) and *dstit* logic. The new *bdi-stit* logic generalizes *bdi* and *dstit* logic in the sense that for any model of *bdi* or *dstit* logic, there is an equivalent *bdi-stit* model.

Keywords: modal logic of agency, deliberative stit logic, *BDI* logic, beliefs, desires, intensions, neighbourhood semantics, branching time structures.

1. Introduction

The temporal logics *BDI* and *BDI** of beliefs, desires, and intentions, developed by Rao and Georgeff [9] are among the most prominent and widely applied formalizations of rational agents, see also [20]. In this paper, we shall introduce a modal logic of beliefs, desires, intentions, and agency. Supplementing the *BDI* vocabulary by a modal operator for agency is a very

*We dedicate this paper to the memory of Alexander Vladimirovich Kuznetsov (1926–1984).

Special Issue *To the memory of A. V. Kuznetsov*. Guest Editors: A. Muravitsky and S. Odintsov
 ISSN: 1425-3305

natural move for at least two reasons. First of all, obviously the concrete actions of rational agents (as opposed to their available action types) are often described as being guided by the agents' beliefs, desires, and intentions. Therefore it is desirable to be able to talk not only about the doxastic, volitional, and intentional attitudes of rational agents, but also about their concrete actions. Secondly, the semantics of *BDI* and *BDI** and the semantics of the perhaps most prominent multi-agent logics of seeing to it that, the stit logics developed by Belnap, Perloff, and Xu, Chellas, von Kutschera, Horty and others (see [3] and references therein), are based on certain branching time structures. Combining both approaches hence more or less suggests itself.

In Section 2 we shall briefly present the semantics of the *deliberatively seeing to it that* (*dstit*) operator, and Section 3 is devoted to a presentation of the standard semantics of the sublanguage of *BDI* comprising the alethic modalities *it is necessary that* and *it is possible that* and the belief, desire, and intention operators. This logic will be called *bdi* logic. In Section 4, we shall motivate and semantically present a combined logic of beliefs, desires, intentions, and agency, using the branching time frames from stit theory. Moreover, we shall take seriously a phenomenon frequently encountered in rational agents and in reasoning about their beliefs and desires, namely the fact that agents sometimes have conflicting (though not inconsistent) beliefs and desires. Although conflicting beliefs have been discussed in doxastic logic in the context of the logical omniscience problem and antagonistic desires have been described also in Michael Bratman's Belief-Desire-Intention theory of human practical reasoning [4], ascriptions of such attitudes are unsatisfiable in the familiar *BDI* logics. The main technical contribution of the present paper is the proof that our logic of *bdi-stit* in fact is a generalization of both *dstit* logic and *bdi* logic in the following sense: For every model of *dstit* logic and every model of *bdi* logic, there exits an equivalent *bdi-stit* model. Some concluding remarks on future work can be found in Section 5.[1]

[1] In [18] it has been suggested to combine belief, desire, intention, and *dstit* modalities, but to restrict the syntax so as to prefix intention modalities only to *dstit*-formulas. Although this restriction is well-motivated, in the present paper we nevertheless impose no syntactic restrictions on formulas in the scope of intention modalities, in order to emphasize the fact that the *bdi-stit* logic generalizes *bdi* logic.

2. Deliberative-stit logic

2.1. The *dstit* ontology

Deliberative-stit logic (*dstit* logic)[2] is a logic of agency based on so-called branching time frames. These structures reflect the idea that the past is linear and determined. In contrast to this, the future is assumed to be indetermined and hence not linear. Branching time temporal logic goes back to Arthur Prior and Richmond Thomason [14, 15].

A branching time frame is a structure of the form $\mathcal{F} = (M, \leq)$, where M is non-empty set of moments of time, and $\leq$ is a partial order on M satisfying the following property:

$$\text{if } n \leq m \text{ and } p \leq m, \text{ then } n \leq p \text{ or } p \leq n, \text{ for all } m, n, p\,.$$

For a given moment m and every pair of moments in the past of m, it is determined which one precedes the other. With respect to the future the situation is different. The partial order $\leq$ thus imposes a tree structure on M. A maximal linearly ordered subset $h \subseteq M$ is said to be a *history* in M. This means that for every two elements m, n from h, $m \leq n$ or $n \leq m$ and that it is impossible to add a further moment to h without violating linearity. In the sequel we shall denote the set of all histories of a given frame $\mathcal{F}$ by H. The set $H_{(m)} = \{\, h \,|\, m \in h\}$ then contains all histories which pass through moment m. A moment/history pair (m, h) with $m \in h$ may be called a *situation*.

Branching time frames are the structures on which Belnap, Perloff, and Xu [2, 3] build up their modal logic of agency. They extend a temporal frame $\mathcal{F}$ by a finite, non-empty set $\mathcal{A}$ of agents and by a function $C\colon \mathcal{A} \times M \to \mathcal{P}(\mathcal{P}(H))$ that assigns to every agent at each moment a family of sets of histories such that for every agent $\alpha \in \mathcal{A}$, the set $C(\alpha, m) = C^{\alpha}_{m}$ is an equivalence relation on the set $H_{(m)}$. The equivalence class $C^{\alpha}_{m}(h)$ contains the histories which are *choice-equivalent* for agent α at moment m, i.e., the histories agent α cannot distinguish at moment m by her or his actions. There exist then different equivalence classes $\{[C^{\alpha}_{m}(h)] \,|\, h \in H_{(m)}\}$ on $H_{(m)}$, which are also said to be the choice cells for α in moment m. In particular, histories which share a moment later than m are choice-equivalent at m for any agent. This is, however, only a sufficient but not a necessary condition for choice-equivalence.

[2] Note that there are several formal notions of seeing to it that, for example, the *Chellas stit*, the *achievement stit*, and the *deliberative stit*, see [3].

The resulting *stit frames* $\mathcal{F} = (M, \leq, \mathcal{A}, C)$ are the semantic structures used to interpret agentive sentences. In order to present a truth definition, we first need a language.

2.2. The syntax of *dstit* logic

The language of *dstit* logic comprises denumerably many atomic formulas $(p_1, p_2, \ldots)$, the connectives of classical propositional logic $(\neg, \wedge, \vee, \supset, \equiv)$ and the modal necessity and possibility operators $\Box$ and $\Diamond$. This vocabulary is supplemented by action modalities α_1 *dstit*:, ..., α_n *dstit*: where α_i *dstit*: $(1 \leqslant i \leqslant n)$ is read as 'agent α_i deliberatively see to it that'. We thus also assume a set of agent variables $(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

DEFINITION 1 (*dstit* syntax). The formulas of *dstit* logic are inductively defined as follows:

1. Every atomic formula $p_1, p_2 \ldots$ is a formula.
2. If φ, ψ are formulas, then so are $\neg\varphi$, $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, $(\varphi \supset \psi)$, $(\varphi \equiv \psi)$, $\Box\varphi$ and $\Diamond\varphi$.
3. If φ is a formula and α is an agent variable, then α *dstit*: φ is a formula.
4. Nothing else is a formula.

Note that Belnap, Perloff, and Xu (and other authors) use the notation $[\alpha$ *dstit*: $\varphi]$ instead of α *dstit*: φ.

2.3. The semantics of *dstit* logic

Given a *stit* frame $\mathcal{F} = (M, \leq, \mathcal{A}, C)$, by adding a valuation function v which maps every atom to a set of situations in $\mathcal{F}$, one obtains a model $\mathcal{M} = (\mathcal{F}, v)$. In the sequel, a model $(\mathcal{F}, v)$ will be called a *dstit* model.

The satisfiability of a formula φ at a situation (m, h) in the model, $\mathcal{M}, (m, h) \models \varphi$, is defined by induction on the construction of φ. Intuitively, $v(\varphi)$ is the set of situations where φ is true. An agent variable α is interpreted in $\mathcal{M}$ by an element from set $\mathcal{A}$. For simplicity, we also use α to denote its interpretation. From now on, we shall assume that the connectives $\vee$, $\supset$, $\equiv$, and $\Diamond$ as defined as usual.

DEFINITION 2 (*dstit* semantics). Let (m, h) be a situation, and let α be an agent in $\mathcal{M} = (\mathcal{F}, v)$.

$\mathcal{M}, (m,h) \models \varphi$ iff $(m,h) \in v(\varphi)$, if φ is an atomic formula.
$\mathcal{M}, (m,h) \models \neg\varphi$ iff $\mathcal{M}, (m,h) \not\models \varphi$.
$\mathcal{M}, (m,h) \models \varphi \wedge \psi$ iff $\mathcal{M}, (m,h) \models \varphi$ and $\mathcal{M}, (m,h) \models \psi$.
$\mathcal{M}, (m,h) \models \Box\varphi$ iff $\mathcal{M}, (m,h') \models \varphi$ for all $h' \in H_{(m)}$.
$\mathcal{M}, (m,h) \models \alpha\ dstit\colon \varphi$ iff (i) $\mathcal{M}, (m,h') \models \varphi$ for all $h' \in C^{\alpha}_{m}(h)$,
(ii) there exists $h'' \in H_{(m)}$ with $\mathcal{M}, (m,h'') \not\models \varphi$.

Necessity, is thus interpreted as an *S5*-type modal operator (namely, as a universal quantifier on $H_{(m)}$). The action modalities, however, are not normal modal operators.

3. A logic of beliefs, desires, and intentions

The temporal logic of beliefs, desires, and intentions *BDI* has been developed in the 1990s by Rao and Georgeff in a series of papers, see [8, 9]. A survey and further development in book length of *BDI* logics has been presented by Michael Wooldridge [20]. We intend to define a logic whose language contains in addition to the *dstit* operator the *BDI* operators from Rao's and Georgeff's logic [9], but which does without the temporal operators of the Computational Tree Logic *CTL*. We are thus interested in the bare essentials needed to extend the syntax of *dstit* logic to obtain a language for reasoning about beliefs, desires, intentions, and concrete actions. In order to highlight that we are dealing with a fragment of the *BDI* language, we shall use the lowercase letters *bdi* to refer to the new logic.

3.1. The syntax of *bdi* logic

In addition to being able to talk about possibility and necessity, we now want to be able to express that an agent has certain beliefs, desires, and intentions. To this end, the language of alethic propositional modal logic is extended by modal operators $\alpha_i\ bel:$, $\alpha_i\ des:$ and $\alpha_i\ int:$ $(1 \leqslant i \leqslant n)$, where α_i stands for a rational agent capable of having beliefs, desires, and intentions.

Definition 3 (*bdi* syntax). The formulas of *bdi* logic are inductively defined as follows:

1. Every atomic formula $p_1, p_2, \ldots$ is a formula.
2. If φ, ψ are formulas, then so are $\neg\varphi$, $(\varphi \wedge \psi)$, and $\Box\varphi$.

3. If φ is a formula and α is an agent variable, then $\alpha\, bel\colon \varphi$, $\alpha\, des\colon \varphi$ and $\alpha\, int\colon \varphi$ are formulas.
4. Nothing else is a formula.

Note that Wooldridge, Rao and Georgeff (and other authors) use the notation $[Bel\, \alpha\ \varphi]$ or $Bel(\varphi)$ instead of $\alpha\, bel\colon \varphi$.

3.2. The semantics of *bdi* logic

Like the models of *dstit* logic, the models of *BDI* and hence of *bdi* logic are based on branching time frames. However, the models of *bdi* logic differ from the models of *dstit* logic not just in replacing the 'choice function' C by interpretation functions for belief, desire, and intention modalities, but also in the construction of the situations in which formulas are semantically evaluated.

Starting from a branching time structure $(M, \leq)$, *worlds* $w = (M_w, R_w)$ are considered where

$$M_w \subseteq M\,,$$
$$R_w \subseteq R = \{\,(m, m') \mid m \leq m' \text{ or } m' \leq m, \text{ for } m, m' \in M\} \subseteq M \times M\,,$$

$M_w \neq \emptyset$, and R_w is such that w itself is a branching time frame, i.e., a tree structure. The set of all worlds of a given frame will be denoted by W. *Situations*[3] in a frame are all pairs (w, m) with $w \in W$ and $m \in M_w$. Moreover, the notion of a path in a world w is defined. Given a situation $w_0 = (w, m_0)$, a path is an arbitrarily long sequence $(w_0, w_1, w_2, \ldots)$, where the $w_i = (w, m_i)$ are situations such that for every i, $m_i < m_{i+1}$, i.e., the moments of a path are linearly ordered. Maximal paths (fullpaths) are paths that cease to be linear upon the addition of a further situation (w, m_j) with $m_0 < m_j$.

To complete the semantic picture, a finite, non-empty set $\mathcal{A}$ of agents and suitable interpretation functions for belief, desire, and intention operators are stipulated. A frame of a *bdi* model then is a structure of the shape $\mathcal{F} = (M, \leq, W, \mathcal{A}, B, D, I)$. Every function $F \in \{B, D, I\}$ assigns to each agent a set of triples (w, m, w'), where both (w, m) and (w', m) are situations. In other words,

$$F\colon \mathcal{A} \to \mathcal{P}((W \times M) \times W).$$

[3]The situations of *dstit* logic differ from the situations in *BDI* logics and hence *bdi* logic. In the sequel, the context will resolve ambiguity.

The set $B^w_m(\alpha) = \{\, w' \mid (w, m, w') \in B(\alpha) \,\}$ is then understood as the set of all worlds compatible with that agent α believes in the situation (w, m). This conception is familiar from Hintikka's [10] seminal work on epistemic logic. Similarly, the sets $D^w_m(\alpha)$ and $I^w_m(\alpha)$ are taken to contain the worlds compatible with what agent α is desiring, respectively intending in the situation (w, m). Thus, for every agent α, each $F \in \{B, D, I\}$ is a relation between situations and worlds. In order to avoid the satisfiability of formulas $\alpha\, bel\colon \varphi$, $\alpha\, des\colon \varphi$, and $\alpha\, int\colon \varphi$, where φ is unsatisfiable, it is usually required that the relations $F \in \{B, D, I\}$ are *serial*, which is to say that every set $F^w_m(\alpha)$ is non-empty. See, for instance, [9, p. 305].[4]

A model in which formulas from Definition 3 are assigned truth values then is a pair $\mathcal{M} = (\mathcal{F}, v)$, where $\mathcal{F} = (M, \leq, W, \mathcal{A}, B, D, I)$ is a frame and v is a valuation function from atoms to sets of situations. The interpretation of a formula φ in a situation is defined as follows.

Definition 4 (*bdi* semantics). Let (w, m) be a situation, α an agent in model $\mathcal{M} = (\mathcal{F}, v)$ and let φ, ψ be formulas according to Definition 3.

$\mathcal{M}, (w, m) \models \varphi$	iff	$(w, m) \in v(\varphi)$, if φ is an atomic formula.
$\mathcal{M}, (w, m) \models \neg\varphi$	iff	$\mathcal{M}, (w, m) \not\models \varphi$.
$\mathcal{M}, (w, m) \models \varphi \wedge \psi$	iff	$\mathcal{M}, (w, m) \models \varphi$ and $\mathcal{M}, (w, m) \models \psi$.
$\mathcal{M}, (w, m) \models \Box\varphi$	iff	$\mathcal{M}, (w', m) \models \varphi$ for every situation $(w', m) \in W \times M$.
$\mathcal{M}, (w, m) \models \alpha\, bel\colon \varphi$	iff	$\mathcal{M}, (w', m) \models \varphi$ for all $w' \in B^w_m(\alpha)$.
$\mathcal{M}, (w, m) \models \alpha\, des\colon \varphi$	iff	$\mathcal{M}, (w', m) \models \varphi$ for all $w' \in D^w_m(\alpha)$.
$\mathcal{M}, (w, m) \models \alpha\, int\colon \varphi$	iff	$\mathcal{M}, (w', m) \models \varphi$ for all $w' \in I^w_m(\alpha)$.

Again, necessity is interpreted as an *S5*-type modality.

In *BDI* and *BDI**, a distinction is drawn between so-called state formulas and path formulas interpreted at paths. In the language of *bdi* logic, this distinction is superfluous, because the temporal operators which give rise to the distinction are omitted. Yet, we add the alethic modalities $\Box$ and $\Diamond$ (defined as $\neg\Box\neg$) from the language of *dstit* logic. The alethic modal operators have no counterpart in *CTL* and *CTL**. Instead of $\Box$ and $\Diamond$, in *CTL* temporal modalities are considered, for example, $F\varphi$ (*sometimes in the future it is the case that* φ) and $G\varphi$ (*always in the future it is the case*

[4] If the relations *are* serial, then formulas $(\alpha\, des\colon \varphi \wedge \alpha\, des\colon \neg\varphi)$, $(\alpha\, bel\colon \varphi \wedge \alpha\, bel\colon \neg\varphi)$, and $(\alpha\, int\colon \varphi \wedge \alpha\, int\colon \neg\varphi)$ are unsatisfiable. See the discussion on antagonistic desires and beliefs in Section 4.

that φ). We do without temporal operators in *bdi* logic to simplify matters and because our primary aim is to combine a logic of beliefs, desires and intentions with *dstit* logic. With this goal in mind, we interpret $\Box\varphi$ as true in a situation (w, m) from a *bdi* model if the formula φ is true at every world accessible from m.

This definition is compatible with the semantics of $\Box$ in *dstit* logic in the sense that also the 'historical necessity' of *dstit* logic is an $S5$-type modality. Moreover, as in *dstit* logic, the truth of a formula $\Box\varphi$ at a situation does not warrant any conclusion concerning the truth value of φ at earlier or later moments of time. The same holds true for $\Diamond\varphi$, which is true in a situation (w, m) if there exists an accessible world in which φ is true at m. This is a concise conception of possibility, which may be clearly distinguished from a temporal reading of *possibly* φ, namely *it is possible that sometimes in the future* φ *is true in some history.*

4. A generalization of *bdi* logic and *dstit* logic

Our aim in defining *bdi-stit* logic is to obtain a modal logic of agency in which one can express, for instance, that an agent desires or intends to see to it that something is the case. Also, we want to be able to express that an agent believes that a certain agent sees to it that something is the case. In this context, a lot of interesting philosophical questions arise, such as whether the worlds compatible with what an agent believes in a given situation ought to be accessible from this situation. As there may be several agents, the question turns up whether an agent can intend that another agent can bring about something. We shall not address all these questions in the present paper,[5] but in addition to adjoining the *dstit* operator to the language of *bdi* logic, we shall also *generalize bdi* logic in its own vocabulary.

The generalization of *bdi* logic pertains in the first place to the concepts of belief and desire. In Rao's and Georgeff's *BDI* the belief, desire, and intention operators are all treated in the same way, namely their semantics is explicated in terms of relations on situations. The sentence *Agent* α *believes that* φ is true in a situation (w, m) if φ is true in every situation (w', m) with $w' \in B^w_m(\alpha)$. Thus, an agent believes in a situation (w, m) that something is the case, if in every world compatible with what the agent believes in (w, m), φ is true at m. It is neither required, nor is it excluded that $w \in B^w_m(\alpha)$, but if $w \in B^w_m(\alpha)$, then α has only true beliefs in (w, m). Intuitively, this

[5] A discussion of other-agent-intending can be found in [18].

conception, or rather its notion of compatibility may well be contentious. In general, we normally believe what has been shown to be the case, but one might object that many beliefs are virtually undecidable. Agents may believe that they made optimal decisions in the past, that there exists a divine being, that money ruins character, etc. Intuitively, such believes do not preclude a world w from being compatible with what an agent believes in a situation (w, m). What if one assumes that, without further qualifications, it is neither true nor false that money ruins character? Can a world be compatible with the belief that φ, if φ is neither true nor false? We need not overemphasize this objection, because the problems raised can be evaded by considering models in which every agent/situation pair (α, s) is assigned several sets of situations. Theses sets may now be seen as 'worlds' compatible with what the agent believes in s.[6] Moreover, a similar approach has also bee suggested to partly overcome the problem of logical omniscience in epistemic logic, see [6, 7, 16].

The sentence *Agent* α *believes that* φ is then true iff there exists a set in the range of the function used to interpret the belief operator such that φ is true in every situation from this set. In particular, it is possible that in a given situation (w, m) it is true that an agent believes that φ and believes that $\neg\varphi$. This is the case iff there exist two disjoint 'worlds' compatible with what the agent believes in (w, m) such that φ is true in every situation from one set and $\neg\varphi$ is true throughout the other world.

The following example suggests that a rational agent may, in fact, simultaneously believe that φ and believe that $\neg\varphi$. Suppose that because of the desire to have children an agent decides to waive her plans of a professional career. Years later, together with her beloved children, she is listening to a friend, who is reporting about her job-related successes. In this situation our agent might well correctly be described as simultaneously believing that it was wise to decide in favour of children and believing that this decision was unwise.

The idea of being able to consistently ascribe conflicting beliefs to a rational agent is not new, see [16, 17], and must be distinguished from consistently ascribing inconsistent beliefs. In the survey by Fagin and Halpern [7] on doxastic logics for rational agents, the authors present an example which goes back to the physicist Eugene Wigner [19], who diagnosed that quantum field theory, which describes three out of the four fundamental interactions,

[6] Note that the notion of a world in *BDI* and *bdi* logic differs from the notion of a world in the *bdi-stit* logic we are about to develop.

is incompatible with the general theory of relativity which describes gravitation, the fourth interaction. A physicist may well believe the statements of quantum field theory *and* believe the statements of the general theory of relativity. However, this physicist has to avail of two 'frames of mind', in fact, this agent can be viewed as a 'society of minds'. Such considerations are quite familiar from the Philosophy of Science. The frames of mind can be associated with different, competing paradigms in Thomas Kuhn's terminology [13] or competing 'Begriffsapparaturen' in Kazimierz Ajdukiewicz's terminology [1].

An agent may enter into a similar disunion, when he has antagonistic desires. Van der Hoek and Wooldridge [11, p. 142] point out that "[i]mplemented BDI agents require that desires be *consistent* with one another, although *human* desires often fail in this respect". In other words, it may be the case that in a given situation a rational agent desires that φ and desires that $\neg\varphi$. Difficult decisions are often accompanied by antagonistic wishes. Consider an agent who is supposed to donate a kidney to his brother. The agent may be assumed both to desire that he spends his life with two healthy kidneys and to desire that he helps his brother by donating one of his organs. We thus would like to have available semantical models in which formulas are satisfiable that ascribe conflicting beliefs and antagonistic desires.

In contrast to this, the semantics should *not* be such that an agent may have paradoxical beliefs or desires, and, clearly, given the society-of-minds understanding, if an agent believes that φ and believes that $\neg\varphi$, this does not imply that the agent believes that $\varphi \wedge \neg\varphi$. The agent from one of our examples above may be aptly described as believing that her decision was wise and believing that her decision was unwise, but it would be inadequate to describe the agent as believing a contradiction. Similar things can be said with respect to an agent's desires. An agent may have the desire to donate a kidney to his brother and the desire to continue to life with two kidneys, but it would be completely inappropriate to describe the agent as desiring that he simultaneously keeps his pair of kidneys and donates one kidney to his brother.

4.1. The syntax of *bdi-stit* logic

As we intend to combine the language of *stit* logic and *bdi* logic, it should be clear how the syntax of *bdi-stit* logic is inductively defined. In addition to the obvious merger of vocabulary, however, we shall also introduce a new possibility operator, namely $\diamond$.

DEFINITION 5 (*bdi-stit* syntax). 1. Every atomic formula p_1, p_2, ... is a formula.

2. If φ, ψ are formulas and α is an agent variable, then $\neg\varphi$ $(\varphi \wedge \psi)$, $\Box\varphi$, $\diamond\!\!\!\!\cdot\,\varphi$, $\alpha\, dstit\colon \varphi$, $\alpha\, bel\colon \varphi$, $\alpha\, des\colon \varphi$ and $\alpha\, int\colon \varphi$ are formulas.

3. Nothing else is a formula.

4.2. The semantics of *bdi-stit* logic

Above, both *dstit* logic and *bdi* logic were defined model-theoretically. In merging the two logics, we have to decide on which kind of frames the *bdi-stit* models are based. A natural thought is to use either the *stit* frames or the frames of *bdi* models. In our view, the concept of histories which represent possible developments of time is not only simpler, but also its representation is more concrete to the senses than the overlapping worlds of the *bdi* semantics. Moreover, the *bdi* frames may be simplified to frames with linear worlds. For every world w in the *bdi* semantics, one may find a set of histories that represent w.

The basis for the *bdi-stit* semantics are again temporal frames $(M, \leq)$. As in the *dstit* semantics, we define the maximal linear subsets of M as histories. The set of all histories of a given temporal frame $\mathcal{F}$ is called H. A situation is a moment/history pair (m, h) with $m \in h$. The set of all situations in $\mathcal{F}$ is symbolized by S. So far the *bdi-stit* semantics is analogous to the *dstit* semantics.

In *bdi-stit* logic the frames are modified to obtain a (kind of) neighbourhood semantics (alias minimal models semantics), cf. [5, 7]. We shall use this semantics in the first place to obtain the satisfiability of formulas $(\alpha\, des\colon \varphi \wedge \alpha\, des\colon \neg\varphi)$ and $(\alpha\, bel\colon \varphi \wedge \alpha\, bel\colon \neg\varphi)$, whereas it is *not* our aim here to tackle satisfactorily the logical omniscience problem. But once we consider such a semantics, it becomes quite naturally to consider in addition to the alethic, 'historcal' possibility operator $\diamond$ also a non-historical notion of possibility, expressed by the operator $\diamond\!\!\!\!\cdot$. Therefore, for every situation s, we introduce a neighbourhood system, namely a family of sets of situations. Every such set of situations is called a *world*.[7] A world in the neighbourhood system for s is accessible from s (and may be seen to contain situations which are accessible from s). In general, there are no constraints imposed on the worlds in the neighbourhood system of a situation, but every world must

[7] Note again that the term *world* is used differently in *bdi* logic and in *bdi-stit* logic.

be non-empty.[8] The situation $s = (m, h)$ need not belong to the elements of its neighbourhood system, and an element of the system need not contain a history running through m. The neighbourhood systems for the situations of a frame are given by a function N,

$$N\colon S \to \mathcal{P}(\mathcal{P}(S) \setminus \emptyset), \qquad (m, h) \mapsto N(m, h) = N_{(m,h)}.$$

The function N thus assigns to every situation s a family of sets of situations, a neighbourhood system comprising neighbourhoods of s, alias the worlds accessible from s. To every temporal frame, we add a function N. Given a particular frame $\mathcal{F}$, in the sequel we shall denote by N both the neighbourhood systems assigning function as well as the union of all neighbourhood systems of $\mathcal{F}$. The context will disambiguate between the function N and the set N. In any case, N_s denotes the neighbourhood system of situation s.

The interpretation of the *dstit* operator is like in *dstit* logic. That is, we assume a function that identifies for every moment m choice-equivalent histories passing through m. This choice function C assigns to every agent α and every moment m an equivalence-relation on the histories containing m. The equivalence classes of this relation are the choice cells available to α at m. Since for evaluating $\alpha\, dstit\colon \varphi$ in a situation (m, h) it is irrelevant how the equivalence relation is specified for histories not choice-equivalent with h, the function C is defined as a mapping from agent/situation-pairs $(\alpha, (m, h))$ into subsets of H_m:

$$C\colon \mathcal{A} \times S \to \mathcal{P}(H), \qquad (\alpha, s) \mapsto C(\alpha, s) = C^{\alpha}_{s}\,.$$

The set $C^{\alpha}_{(m,h)}$ thus consists of the histories agent α cannot distinguish from h by her/his actions at moment m.

In order to deal with the belief, desire, and intention operators, we need interpretation functions for these operators. Whereas we want to model conflicting beliefs and antagonistic desires, we assume that it is impossible that an agent intends that φ and simultaneously intends that $\neg\varphi$. Intuitively, an important difference between desiring and believing on the one hand and intending on the other hand is that desires and beliefs are not directed towards actions in the way intentions are directed towards actions. Typically, intentions are intentions to act, but desires are not typically desires to act. Some of our actions can be explained in terms of our desires or beliefs and likewise, our intentions often can be explained in terms of our desires or beliefs. A

[8] The existence of non-empty sets as neighbourhoods (worlds) leads to the satisfiability of formulas $\alpha\, des\colon (\varphi \wedge \neg\varphi)$ and $\alpha\, bel\colon (\varphi \wedge \neg\varphi)$.

rational agent α may have antagonistic desires and, moreover, α may be a 'society of minds' and may in this sense be adequately described as believing that φ and simultaneously believing that $\neg\varphi$. But α's volitional incoherence and α's doxastic incoherence as a society of minds is not passed on to α's intentions. We may return to one of the earlier examples. Suppose an agent desires to keep his two kidneys and desires to donate one of his kidneys to his brother. We do not assume that the agent may be adequately described as intending (with respect to one desire) that he donates one of his kidneys, and simultaneously as intending (with respect to the conflicting desire) that he does not donate one of his kidneys. The agent either intends that he donates an organ or he intends that he does not donate an organ, but not both. An agent cannot simultaneously see to it that φ and see to it that $\neg\varphi$, and similarly an agent cannot simultaneously intend that φ and intend that $\neg\varphi$.

Note that an agent may intend that φ although he desires that $\neg\varphi$. An agent α may desire that he does not donate one of his kidneys, but out of sense of duty or moral considerations α may simultaneously intend that he donates one of his kidneys. Moreover, although intentions are often prestages to actions, there are also unintended actions. Suppose α believes that he will refrain from hit-and-run driving and that α causes a fatal accident. The agent leaves the place of accident, maybe because of a shock but in any case without intending to abscond after the accident. We would describe this absconding as an unintended action performed by the agent. If we imagine that the agent did not desire that he absconds, then the example shows that an agent may see to it that φ without desiring that φ.

The interpretation functions for the intention and desire operators therefore have to be chosen such that an agent in a situation may have conflicting desires, may have the intention that φ but not the desire that φ, and may see to it that φ without desiring that φ.

To this end, the intention operator is interpreted by a function I that assigns to every agent/situation-pair a set of situations, i.e., a world in a neighbourhood system, whereas the desire operator is interpreted by a function D that assigns to every agent/situation-pair (α, s) a set of so-called desire worlds, namely the worlds which are, intuitively, compatible with what the agent believes in s. If it is true in s that α desires that φ, then there exists a desire world such that φ is true in every situation from that world.

The function B, which assigns to every agent/situation-pair (α, s) the set of worlds compatible with what α believes in s, is analogous to the function D used to interpret the desire operator, because we want to account for

conflicting beliefs, as explained above. The functions B, D, and I are thus defined on the following sets from a temporal frame:

$$I\colon \mathcal{A} \times S \to N, \qquad (\alpha, s) \mapsto I(\alpha, s) = I_s^\alpha,$$
$$D\colon \mathcal{A} \times S \to \mathcal{P}(N), (\alpha, s) \mapsto D(\alpha, s) = D_s^\alpha,$$
$$B\colon \mathcal{A} \times S \to \mathcal{P}(N), (\alpha, s) \mapsto B(\alpha, s) = B_s^\alpha.$$

The set N here is the set of all neighbourhoods. The set I_s^α is thus not necessarily a neighbourhood from N_s, but there exists a situation s' such that I_s^α is a neighbourhood of s'. In general, it is neither necessary that an agent in a situation has intentions nor is it necessary that if an agent intends that φ, then φ is true in some accessible situation. The set of situations assigned by I to a pair (α, s) just must be a world accessible from some situation and it must be such that if it is true in s that α intends that φ, then φ is true in every situation from that world. Thereby an agent cannot intend that φ if φ is unsatisfiable, but the agent can intend that φ in a situation s if φ is true at every situation in an unaccessible world, a world not in N_s. If this is unwanted, one may restrict the mapping I such that for every situation s, $I_s^\alpha \in N_s$. In this case an agent can intend only what is true throughout an accessible world.

As to the desire worlds of an agent α in a situation s and the worlds compatible with what α believes in s, we also do not require that they belong to N_s, the set of worlds accessible from s. The necessity operator $\Box$ is interpreted as in *dstit* logic, and formulas $\diamondsuit\varphi$ are defined to be true at a situation if there is an accessible world throughout which φ is true.

A *bdi-stit* model, used to interpret the formulas from Definition 5, thus consists of a frame $\mathcal{F} = (M, \leq, \mathcal{A}, N, C, B, D, I)$ with the just introduced functions together with a valuation v. Satisfiability of a formula in a *bdi-stit* model is then defined as follows.

DEFINITION 6 (*bdi-stit* semantics). Let $s = (m, h)$ be a situation, let α be an agent in model $\mathcal{M} = (\mathcal{F}, v)$, and let φ, ψ be formulas according to Definition 5. Then:

$\mathcal{M}, s \models \varphi$ iff $s \in v(\varphi)$, if φ is an atomic formula.
$\mathcal{M}, s \models \neg\varphi$ iff $\mathcal{M}, s \not\models \varphi$.
$\mathcal{M}, s \models \varphi \wedge \psi$ iff $\mathcal{M}, s \models \varphi$ and $\mathcal{M}, s \models \psi$.
$\mathcal{M}, s \models \Box\varphi$ iff $\mathcal{M}, (m, h') \models \varphi$ for all $h' \in H_{(m)}$.
$\mathcal{M}, s \models \diamondsuit\varphi$ iff there exists $U \in N_s$ with $U \subseteq \{ s' \mid \mathcal{M}, s' \models \varphi \}$.[9]

[9] In standard neighbourhood semantics the condition would be $\{ s' \mid \mathcal{M}, s' \models \varphi \} \in N_s$.

$$\mathcal{M}, s \models \alpha\, dstit\colon \varphi \text{ iff } \begin{array}{ll} \text{(i) } \{\,(m,h') \mid h' \in C_s^\alpha\,\} \subseteq \\ \qquad \{\,(m,h') \mid \mathcal{M}, (m,h') \models \varphi\,\}, \\ \text{(ii) } \mathcal{M}, s \models \neg\Box\varphi. \end{array}$$

$$\mathcal{M}, s \models \alpha\, int\colon \varphi \quad \text{iff } I_s^\alpha \subseteq \{\, s' \mid \mathcal{M}, s' \models \varphi\,\}$$

$$\mathcal{M}, s \models \alpha\, des\colon \varphi \quad \text{iff there exists } U \in D_s^\alpha \text{ with } U \subseteq \{\, s' \mid \mathcal{M}, s' \models \varphi\,\}.$$

$$\mathcal{M}, s \models \alpha\, bel\colon \varphi \quad \text{iff there exists } U \in B_s^\alpha \text{ with } U \subseteq \{\, s' \mid \mathcal{M}, s' \models \varphi\,\}.$$

Note that it is not possible that a neighbourhood U satisfying the truth condition of the belief, desire or $\diamondsuit$-operator is empty, because neighbourhoods (alias worlds) must not be empty. Obviously, the notion of historical necessity expressed by $\Box$ and the notion of possibility expressed by $\diamondsuit$ can coincide in a *bdi-stit* model, if every set of an arbitrary neighbourhood system $N_{(m,h)}$ includes every moment/history pair at this moment m.

Remark 7. *Let $\mathcal{M}_{ep} = (M, \leq, \mathcal{A}, N, C, B, D, I, v)$ be a bdi-stit model. If*

$$N\colon (m,h) \mapsto \{\, U \subseteq S \mid \forall\, h' \in H_{(m)} : (m,h') \in U\,\}, \qquad (*)$$

then it holds for any formula φ according to Definition 5 that

$$\mathcal{M}_{ep}, (m,h) \models \Box\varphi \quad \text{iff} \quad \mathcal{M}_{ep}, (m,h) \models \diamondsuit\varphi\,.$$

Instead of stipulating (*) it is alternatively possible to define a neighbourhood system in the following way. Every neighbourhood system $N_{(m,h)}$ of a situation (m,h) contains the neighbourhood $U = \{\,(m,h') \mid h' \in H_{(m)}\,\}$ and for every neighbourhood $V \in N_{(m,h)}$ it holds that $U \subseteq V$.

Now, after we have motivated and semantically defined *bdi-stit* logic, our goal is to show that *bdi-stit* logic in fact is a generalization of *dstit* logic and *bdi* logic. We first consider the logic of agency and make the rather obvious observation that for every model of *dstit* logic there exists a model of *bdi-stit* logic satisfying the same formulas in the language of *dstit* logic.

Remark 8. *Let φ be a formula according to Definition 1. Then for every dstit model $\mathcal{M}_d = (M, \leq, \mathcal{A}, C, v)$ there exists a model $\mathcal{M}_{ep} = (M, \leq, \mathcal{A}, N, C, B, D, I, v)$ such that in every situation (m,h) the following holds:*

$$\mathcal{M}_d, (m,h) \models \varphi \quad \text{iff} \quad \mathcal{M}_{ep}, (m,h) \models \varphi\,.$$

Proof. The model $\mathcal{M}_{ep}$ takes over from a given arbitrary *dstit* model $\mathcal{M}_d$ the components M, $\leq$, $\mathcal{A}$, and C and the valuation function v. As to the definition of N, I, D, and B, it may be noted that the operators $\diamondsuit$, $\alpha\, int\colon$, $\alpha\, des\colon$, and $\alpha\, bel\colon$ do not occur in φ. Therefore the choice of N, I, D, and B is irrelevant. ⊣

A similar but less obvious observation can be made concerning *bdi* logic. The problem we encounter here is that there are different constructions of situations in *bdi* logic and in *bdi-stit* logic. We obtain the following Theorem.

THEOREM 9. *Let φ be a formula according to Definition 3. Then for every bdi model $\mathcal{M}_e = (M, \leq, W, \mathcal{A}, B, D, I, v)$ there exists a model $\mathcal{M}_{ep} = (M', \sqsubseteq, \mathcal{A}', N', C', B', D', I', v')$, such that for every situation (w, m) in $\mathcal{M}_e$ there exists a situation (m, h) in $\mathcal{M}_{ep}$ with:*

$$\mathcal{M}_e, (w, m) \models \varphi \quad \textit{iff} \quad \mathcal{M}_{ep}, (m, h) \models \varphi\,.$$

The proof of this theorem is relegated to Appendix A, where we define a 1-1 mapping from the set of situations of the *bdi* model into the set of situations of the *bdi-stit* model such that in the assigned set the same formulas are satisfiable.

5. Summary and prospects

What have we achieved in this paper?

- We have defined *bdi-stit* logic, a logic of beliefs, desires, intentions, agency, and alethic modalities. This logic brings together the most prominent logic of agency and a fragment of the most influential logics for multi-agent systems.
- We have shown that *bdi-stit* logic is a generalization of both *dstit* logic and *bdi* logic. This is a non-trivial observation in the case of *bdi* logic.
- Due to the generalization of the standard truth conditions for belief ascriptions, we evaded part of the logical omniscience the agents display in Rao's and Georgeff's [9] *BDI* framework. There are residual forms of logical omniscience, as an agent still cannot distinguish in her beliefs between logically equivalent formulas, and the following formula, for instance, is valid (true in every situation of every *bdi-stit* model): $\alpha\, bel : (\varphi \wedge (\varphi \supset \psi)) \supset \alpha\, bel : \psi$. On the other hand, $(\alpha\, bel : \varphi \wedge (\varphi \supset \psi)) \supset \alpha\, bel : \psi$ and $(\alpha\, bel : \varphi \wedge \alpha\, bel : (\varphi \supset \psi)) \supset \alpha\, bel : \psi$ are refutable.
- Although *BDI* theorists like Bratman perceive and discuss conflicting desires, the formal theories *BDI* and *BDI** do not admit modeling such desires, if at the same time ascriptions of inconsistent desires are unsatisfiable. Treating beliefs and desires alike in our neighbourhood semantics, not only the formula $\alpha\, bel : \varphi \wedge \alpha\, bel : \neg\varphi$ but also $\alpha\, des : \varphi \wedge \alpha\, des : \neg\varphi$ is satisfiable.

- In contrast to our treatment of beliefs and desires, the intention operators are interpreted not in a neighbourhood semantics but in a relational semantics. This implies that agents cannot have antagonistic intentions. Moreover, the formula $(\alpha\, int\colon \varphi \wedge \alpha\, int\colon (\varphi \supset \psi)) \supset \alpha\, int\colon \psi$ is valid and if $\varphi \supset \psi$ is universally valid, so is $\alpha\, int\colon \varphi \supset \alpha\, int\colon \psi$.

There are some obvious lines of future research. One task consists in developing a sound and complete proof system for *bdi-stit* logic. Is the logic decidable and if so, is decidability preserved under addition of axiom schemata or inference rules describing certain interaction between the modalities involved? One example of possible interest would be axiom schemata saying that certain agents desire everything they intend.

A. Proof of Theorem 9

Theorem 9. *Let φ be a formula according to Definition 3. Then for every bdi model $\mathcal{M}_e = (M, \leq, W, \mathcal{A}, B, D, I, v)$ there exists a model $\mathcal{M}_{ep} = (M', \sqsubseteq, \mathcal{A}', N', C', B', D', I', v')$ such that for every situation (w, m) in $\mathcal{M}_e$ there is a situation (m, h) in $\mathcal{M}_{ep}$ with:*

$$\mathcal{M}_e, (w, m) \models \varphi \quad \textit{iff} \quad \mathcal{M}_{ep}, (m, h) \models \varphi\,.$$

Proof. The proof consists of two parts. At first we construct a *bdi* model $\mathcal{M}_h$ in such a way that the set of moments of every world is a maximal linear subset of the set of all moments of $\mathcal{M}_h$ nevertheless satisfying the same sets of formulas as the model $\mathcal{M}_e$. For such a model $\mathcal{M}_h$ the way of mapping a situation of this model to a set of situations belonging to a *bdi-stit* model $\mathcal{M}_{ep}$ is evident and then it is also easy to see that $\mathcal{M}_{ep}$ satisfies the same formulas as $\mathcal{M}_h$ and $\mathcal{M}_e$.

Lemma 10. *Let φ be a formula according to Definition 3. Then for every bdi model $\mathcal{M}_e = (M, \leq, W, \mathcal{A}, B, D, I, v)$ there is a bdi model $\mathcal{M}_h = (M'', \preceq, W'', \mathcal{A}'', B'', D'', I'', v'')$, such that:*

$$\mathcal{M}_e, (w, m) \models \varphi \quad \textit{iff} \quad \mathcal{M}_h, (w'', m'') \models \varphi\,, \tag{$\dagger$}$$

where for all $w'' = (T, R) \in W''$ it holds that T is a totally ordered maximal subset of M''.

Proof. We introduce a well-order $\subseteq$ on the set of moments M of model $\mathcal{M}_e$ with $M = \{m_i\}_{i \in I}$ and identify m_0 as the least element according to this

order.[10] In the following we expand each fullpath containing m_0 of every world to a maximal linear subset of M, such that two maximal linear sets associated with two fullpaths of different worlds are different. For getting this we must enlarge set M.

We generate a set $G_w^{m_0}$ for each world $w \in W$, such that an element $h \in G_w^{m_0}$ is construed as a maximal linear subset of M containing one fullpath of w, which comprises m_0, and such that to each fullpath of w there is exactly one element $h \in G_w^{m_0}$. We enumerate all elements of the union of $G_w^{m_0}$ for all $w \in W$,[10] $\{h_r\}_{r \in I^{m_0}}$, such that for all h_r there is exactly one world $w \in W$ with $h_r \in G_w^{m_0}$. However, it is possible that for $r' \neq r$ the sets $h_{r'}$, h_r are identical, for example, if $h_r \in G_w^{m_0}$ and $h_{r'} \in G_{w'}^{m_0}$ for different worlds w, w' which contain the same fullpath.

Without loss of generality we consider $(I^{m_0}, \trianglelefteq)$ as well-ordered. By h_0 we denote the maximal set given with to the least element of I^{m_0}. For h_0 we define the following sets:

$$\leq_{m_0}^0 = \{\, (m', m_0) \,|\, m' \leq m_0,\ m' \in M \,\} \cup \{\, (m_0, m'') \,|\, m_0 \leq m'',\ m'' \in M \,\},$$

$$M_{m_0}^0 = \{m_0\} \text{ and } H_{m_0}^0 = \{h_0\}.$$

By transfinite induction we can take as given the sets $M_{m_0}^{r'}$, $H_{m_0}^{r'}$ and $\leq_{m_0}^{r'}$ for all $r', r'' \in I^{m_0}$ with $r'', r' \triangleleft r$ and such that for $r' \neq r''$ it holds that $h_{r'} \neq h_{r''}$.

Being maximal sets it follows for any r, $r' \in I^{m_0}$, if $h_r \subseteq h_{r'}$, then $h_{r'} = h_r$. Hence it is sufficient to consider two cases:

1. If it applies to all $r' \triangleleft r$, that $h_{r'} \neq h_r$. We stipulate $\leq_{m_0}^r = \emptyset$, $M_{m_0}^r = \emptyset$ and $H_{m_0}^r = \{h_r\}$.

2. If there is exactly one $r' \in I^{m_0}$ with $r' \triangleleft r$ and $h_{r'} = h_r$. In this case we introduce two new moments n and n', which do not belong to M or any set $M_{m_0}^{r'}$ with $r' \triangleleft r$, and we put:

$$\leq_{m_0}^{r'} = \{\, (m', n') \,|\, m_0 \leq m',\ m' \in M \cup M_{m_0}^{r'} \,\} \cup \{(n', n')\},$$

$$M_{m_0}^{r'} = M_{m_0}^{r'} \cup \{n'\},\ H_{m_0}^{r'} = \{h_{r'} \cup \{n'\}\},$$

$$\leq_{m_0}^{r} = \{\, (m', n) \,|\, m_0 \leq m',\ m' \in M \,\} \cup \{(n, n)\},$$

$$M_{m_0}^{r} = \{n\},\ H_{m_0}^{r} = \{h_r \cup \{n\}\},$$

$$G_w^{m_0} = G_w^{m_0} \backslash \{h_r\} \cup \{h_r \cup \{n\}\},\ G_{w'}^{m_0} = G_{w'}^{m_0} \backslash \{h_{r'}\} \cup \{h_{r'} \cup \{n'\}\}.$$

[10] We multiply use the well-ordering theorem resp. the axiom of choice in this proof.

Finally we define the following sets:

$$H_{m_0} = \bigcup_{r \in I^{m_0}} H^r_{m_0}, \quad M_{m_0} = \bigcup_{r \in I^{m_0}} M^r_{m_0}, \quad \leq_{m_0} = \bigcup_{r \in I^{m_0}} \leq^r_{m_0},$$

where H_{m_0} is the set of all maximal, totally ordered, and pairwise different subsets of M_{m_0}, which contain m_0 and one fullpath of any world w. However, the set $G^{m_0}_w$ is a subset of H_{m_0}, which includes for every fullpath p of w exactly one element h, with $m_0 \in p$ and $p \subseteq h$.

The sets H_{m_i}, M_{m_i}, and $\leq_{m_i}$ for all other $i \in I$ can be obtained by transfinite induction. If we assume that $H_{m_{i'}}$, $M_{m_{i'}}$, and $\leq_{m_{i'}}$ are appropriately chosen, we can start the algorithm for H_{m_i} in the same way as for H_{m_0}. But in this process we do not build up the sets $G^w_{m_i}$ for all $w \in W$ over M, but over $M_i = \bigcup_{i' \subset i} M_{m_{i'}}$. In the end we get for all $m \in M$ the sets H_m, M_m, G^w_m, and $\leq_m$ with the desired properties.

Then we take for the *bdi* model $\mathcal{M}_h = (M'', \preceq, W'', \mathcal{A}'', B'', D'', I'', v'')$ the following sets for all formulas φ and each agent $\alpha \in \mathcal{A}'' = \mathcal{A}$:[11]

$$\begin{aligned}
M'' &= \bigcup_{m \in M} M_m, \\
\preceq &= \mathrm{trcl}\big\{(n, n') \,|\, n, n' \in M'', \text{ and there is a } m \in M \text{ with } (n, n') \in \leq_m \big\}, \\
W'' &= \big\{ (h, \preceq_{|h}) \,|\, h \in H_m,\, m \in M \big\}, \\
v''(\varphi) &= \big\{ ((h, \preceq_{|h}), m) \,|\, m \in M,\, w \in W,\, h \in G^w_m,\, (w, m) \in v(\varphi) \big\}, \\
B''(\alpha) &= \big\{ ((h, \preceq_{|h}), m, (h', \preceq_{|h'})) \,|\, m \in M,\, h \in G^w_m,\, h' \in G^{w'}_m,\, (w, m, w') \in B(\alpha) \big\}, \\
D''(\alpha) &= \big\{ ((h, \preceq_{|h}), m, (h', \preceq_{|h'})) \,|\, m \in M,\, h \in G^w_m,\, h' \in G^{w'}_m,\, (w, m, w') \in D(\alpha) \big\}, \\
I''(\alpha) &= \big\{ ((h, \preceq_{|h}), m, (h', \preceq_{|h'})) \,|\, m \in M,\, h \in G^w_m,\, h' \in G^{w'}_m,\, (w, m, w') \in I(\alpha) \big\}.
\end{aligned}$$

Thus every world $w'' \in W''$ is a maximal linear subset of M'' and it holds that for two arbitrary nonequal worlds their sets of moments are different. It remains to show that the situations of $\mathcal{M}_e$ can be mapped to $W'' \times M''$, such that for all situations (w, m) in $\mathcal{M}_e$ there is a situation (w'', m'') in $\mathcal{M}_h$ and the following applies to any formula φ:

$$\mathcal{M}_e, (w, m) \models \varphi \quad \text{iff} \quad \mathcal{M}_h, (w'', m'') \models \varphi\,. \tag{$\dagger$}$$

For all situations (w, m) in M_e it obviously holds that for any h, $h' \in G^m_w$ and any formula φ:

$$M_h, ((h, \preceq_{|h}), m) \models \varphi \quad \text{iff} \quad M_h, ((h', \preceq_{|h'}), m) \models \varphi\,.$$

[11] By $\mathrm{trcl}\, M$ we denote the transitive closure of M, and by $\preceq_{|h}$ we mark the restriction of $\preceq$ to h.

Therefore we assign to every (w, m) in $\mathcal{M}_e$ a situation (w'', m'') in $\mathcal{M}_h$ with $m'' = m$ and $w'' = (h, \preceq_{|_h})$ for an arbitrary $h \in G^m_w$. Then we can show by induction on formulas that (†) holds for this allocation. Let φ be an atomic formula.

$$\begin{aligned}
\mathcal{M}_e, (w, m) \models \varphi \quad & \text{iff} \quad (w, m) \in v(\varphi) \\
& \text{iff} \quad ((h', \preceq_{|_{h'}}), m) \in v''(\varphi) \text{ for all } h' \in G^m_w \\
& \text{iff} \quad \mathcal{M}_h, ((h, \preceq_{|_h}), m) \models \varphi \\
& \text{iff} \quad \mathcal{M}_h, (w'', m) \models \varphi.
\end{aligned}$$

In the case of $\neg\varphi$ or $\varphi \wedge \psi$ the induction step is trivial. Let φ be a formula of the form $\Box\psi$.

$$\begin{aligned}
\mathcal{M}_e, (w, m) \models \Box\psi \quad & \text{iff} \quad \mathcal{M}_e, w' \models \psi \text{ for all situations } (w', m) \in S \\
& \text{iff} \quad \mathcal{M}_h, ((h', \preceq_{|_{h'}}), m) \models \psi \text{ for all } h' \in H_m = \bigcup_{w' \in W} G^m_{w'} \\
& \text{iff} \quad \mathcal{M}_h, ((h', \preceq_{|_{h'}}), m) \models \psi \text{ for all } ((h', \preceq_{|_{h'}}), m) \in S \\
& \text{iff} \quad \mathcal{M}_h, (w'', m) \models \Box\psi.
\end{aligned}$$

Let φ be a formula of the form $\alpha\, bel : \psi$.

$$\begin{aligned}
\mathcal{M}_e, (w, m) \models \alpha\, bel : \psi \quad & \text{iff} \quad \mathcal{M}_e, (w', m) \models \psi \text{ for all } (w, m, w') \in B(\alpha) \\
& \text{iff} \quad \mathcal{M}_h, ((h', \preceq_{|_{h'}}), m) \models \psi \text{ for all } h' \in G^m_{w'} \\
& \qquad \text{with } ((h'', \preceq_{|_{h''}}), m, (h', \preceq_{|_{h'}})) \in B''(\alpha) \\
& \qquad \text{and } h'' \in G^m_w \\
& \text{iff} \quad \mathcal{M}_h, ((h'', \preceq_{|_{h''}}), m) \models \alpha\, bel : \psi \text{ for all } h'' \in G^m_w \\
& \text{iff} \quad \mathcal{M}_h, (w'', m) \models \alpha\ bel : \psi.
\end{aligned}$$

In a similar manner the claim can be shown for $\varphi = \alpha\, des : \psi$ and $\varphi = \alpha\, int : \psi$. Thus the equivalence (†) holds. ⊣

Now we want to prove Theorem 9. For an arbitrary *bdi* model $\mathcal{M}_e = (M, \leq, W, \mathcal{A}, B, D, I, v)$ there is a *bdi-stit* model $\mathcal{M}_{ep} = (M', \sqsubseteq, \mathcal{A}', N', C', B', D', I', v')$, such that for every situation (w, m) of $\mathcal{M}_e$ there exists a situation (m', h') in $\mathcal{M}_{ep}$ and every formula φ, which can be interpreted in both kinds of models, is satisfied in situation (w, m) if and only if it is satisfied in (m', h'). Using Lemma 10 it is possible to take the set of moments of every world in $\mathcal{M}_e$ without loss of generality as a maximal linear subset of M, where the sets of moments are pairwise different. Then the model $\mathcal{M}_{ep} = (M', \sqsubseteq, \mathcal{A}', N', C', B', D', I', v')$ is constructed as follows:

$$\mathcal{A}' := A,\ M' := M,\ C' := \emptyset,\ \sqsubseteq := \leq.$$

The set of all histories of $\mathcal{M}_{ep}$ arises from construction. It is evident that there is a one-to-one mapping from the set of worlds of $\mathcal{M}_e$ onto the set of histories of $\mathcal{M}_{ep}$, so we can apply this to map the situations also one-to-one by allocating a situation $((h, R_h), m)$ in M_e with the situation (m, h) in $\mathcal{M}_{ep}$. In the last step of constructing $\mathcal{M}_{ep}$ we stipulate the functions N', B', D', I' and the valuation function v' for an arbitrary situation (m, h) and an arbitrary formula φ by setting:

$$\begin{aligned}
N'(m,h) &:= \mathcal{P}(\mathcal{P}(\{\,(m,h) \mid (m,h) \text{ situation in } \mathcal{M}_{\text{ep}}\,\})\backslash\emptyset),\\
B'(\alpha,(m,h)) &:= \{\{\,(m,h') \mid ((h,\leq_{|h}),m,(h',\leq_{|h'})) \in B(\alpha)\,\}\},\\
D'(\alpha,(m,h)) &:= \{\{\,(m,h') \mid ((h,\leq_{|h}),m,(h',\leq_{|h'})) \in D(\alpha)\,\}\},\\
I'(\alpha,(m,h)) &:= \{\,(m,h') \mid ((h,\leq_{|h}),m,(h',\leq_{|h'})) \in I(\alpha)\,\},\\
v'(\varphi) &:= \{\,(m,h) \mid ((h,\leq_{|h}),m) \in v(\varphi)\,\}.
\end{aligned}$$

Note that none of the sets $B^h_m(\alpha)$, $D^h_m(\alpha)$ and $I^h_m(\alpha)$ is empty. Consequently, $B'(\alpha,(m,h))$, $D'(\alpha,(m,h))$, and $I'(\alpha,(m,h))$ fail to be empty, too.

The verification of

$$\mathcal{M}_{ep},(m,h) \models \varphi \quad \text{iff} \quad \mathcal{M}_h,(w,m) \models \varphi \tag{‡}$$

with $w = (h, \leq_{|h})$ is also by induction. Let φ be an atomic formula.

$$\begin{aligned}
\mathcal{M}_{ep},(m,h) \models \varphi \quad &\text{iff} \quad (m,h) \in v'(\varphi)\\
&\text{iff} \quad (w,m) \in v(\varphi), \text{ hence } w = (h,\leq_{|h})\\
&\text{iff} \quad \mathcal{M}_e,(w,m) \models \varphi.
\end{aligned}$$

Let φ be a negation or conjunction, then the induction step is again trivial. If $\varphi = \Box\psi$, then the following equivalences hold:

$$\begin{aligned}
\mathcal{M}_{ep},(m,h) \models \Box\psi \quad &\text{iff} \quad \mathcal{M}_{ep},\,(m,h') \models \psi \text{ for all } h' \in H_{(m)}\\
&\text{iff} \quad \mathcal{M}_e,(w',m) \models \psi \text{ with } w' = (h',\leq_{|h'})\\
&\qquad \text{for all } h' \in H_{(m)}\\
&\text{iff} \quad \mathcal{M}_e,(w',m) \models \psi \text{ for all } w' \in W\\
&\text{iff} \quad \mathcal{M}_e,(w,m) \models \Box\psi.
\end{aligned}$$

Let $\varphi = \alpha\, bel : \psi$, the same applies to formulas of form $\alpha\, des : \psi$, then:

$$\begin{aligned}
\mathcal{M}_{ep},(m,h) \models \alpha\, bel : \psi \quad &\text{iff} \quad \text{there is a } U \in B'^{\alpha}_{(m,h)} \text{ with}\\
&\qquad U \subseteq \{\,s' \mid \mathcal{M}_{ep},\, s' \models \varphi\,\}
\end{aligned}$$

$$
\begin{aligned}
&\text{iff } \mathcal{M}_{ep}, (m, h') \models \psi \text{ for all } (m, h') \in U \text{ with} \\
&\quad U = \{\, (m, h') \mid ((h, \leq_{|h}), m, (h', \leq_{|h'})) \in B(\alpha) \,\} \in B'^{\alpha}_{(m,h)} \\
&\text{iff } \mathcal{M}_{e}, (w', m) \models \psi \text{ for all } w' = (h', \leq_{|h'}) \text{ and} \\
&\quad w' \in B^{w}_{m}(\alpha), \text{ where } w = (h, \leq_{|h}) \\
&\text{iff } \mathcal{M}_{e}, (w, m) \models \alpha \; bel : \psi \,.
\end{aligned}
$$

Since in a *bdi-stit* model the intention operator differs in kind of the interpretation from the belief and the desire operator, the equivalence (‡) has to be shown separately.

$$
\begin{aligned}
\mathcal{M}_{ep}, (m, h) \models \alpha \; int : \psi \quad &\text{iff} \quad I'^{\alpha}_{(m,h)} \subseteq \{\, (m', h') \mid \mathcal{M}_{ep},\, (m', h') \models \psi \,\} \\
&\text{iff} \quad \{\, (m, h') \mid ((h, \leq_{|h}), m, (h', \leq_{|h'})) \in I(\alpha) \,\} \\
&\qquad \text{is a subset of } \{\, (m', h') \mid \mathcal{M}_{ep},\, (m', h') \models \psi \,\} \\
&\text{iff} \quad \mathcal{M}_{ep},\, (m, h') \models \psi \text{ for all } h' \text{ with} \\
&\qquad (h', \leq_{|h'}) \in I_{m}^{(h, \leq_{|h})}(\alpha) \\
&\text{iff} \quad \mathcal{M}_{e},\, (w', m) \models \psi \text{ for all } w' \in I^{w}_{m}(\alpha) \text{ with} \\
&\qquad w = (h, \leq_{|h}) \\
&\text{iff} \quad \mathcal{M}_{e},\, (w, m) \models \alpha \; int : \psi. \qquad \dashv
\end{aligned}
$$

References

[1] Ajdukiewicz, K., "Das Weltbild und die Begriffsapparatur", *Erkenntnis* 4 (1934), 259–287.

[2] Belnap, N. D., and M. Perloff, "Seeing to it that: a canonical form for agentives", *Theoria* 54 (1988), 175–199.

[3] Belnap, N.D., M. Perloff, and M. Xu, *Facing the Future: Agents and Choices in our Indeterminist World*, Oxford UP, New York, 2001.

[4] Bratman, M. E., *Intentions, Plans and Practical Reason*, Harvard University Press, Cambridge MA, 1987.

[5] Chellas, B., *Modal Logic: An Introduction*, Cambridge University Press, Cambridge, 1980.

[6] Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi, *Reasoning about Knowledge*, MIT Press, Cambridge MA, 1995.

[7] Fagin, R., and J. Y. Halpern, "Belief, awareness and limited reasoning", *Artificial Intelligence* 34 (1988), 39–76.

[8] Georgeff, M. P., and A. S. Rao, "Modeling rational agents within a BDI-architecture", *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, Morgan Kaufmann Publishers, San Mateo, 1991.

[9] Georgeff, M. P., and A. S. Rao, "Decision procedures for BDI logics", *Journal of Logic and Computation* 8 (1998), 293–342.

[10] Hintikka, J., *Knowledge and Belief. An Introduction to the Logic of the Two Notions*, Cornell University Press, Ithaca NY, 1962, Kings College Publications, London, 2005.

[11] van der Hoek, W., and M. Wooldridge, "Towards a logic of reational agency", *Logic Journal of the IGPL* 11 (2003), 135–159.

[12] Horty, J. F., and N. D. Belnap, "The Deliberative Stit: A study of action, omission, ability and obligation", *Journal of Philosophical Logic* 24 (1995), 583–644.

[13] Kuhn, T. S., *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, 1962.

[14] Prior, A., *Past, Present, and Future*, Oxford University Press, Oxford, 1967.

[15] Thomason, R., "Indeterminist time and truth-value gaps", *Theoria* 36 (1970), 264–281.

[16] Vardi, M., "On epistemic logic and logical omniscience", pages 293–305 in J. Y. Halpern (ed.), *Theoretical Aspects of Reasoning about Knowledge. Proceedings of the 1986 Conference*, Morgan Kaufmann Publishers, Los Altos, 1986.

[17] Wansing, H., "A general possible worlds framework for reasoning about knowledge and belief, *Studia Logica* 49 (1990), 523–539, and 50 (1991), 359.

[18] Wansing, H., "Tableaux for multi-agent deliberative-stit logic", pages 503–520 in G. Governatori, I. Hodkinson and Y. Venema (eds.), *Advances in Modal Logic*, vol. 6, College Publications, London, 2006.

[19] Wigner, E. P., "The unreasonable effectiveness of mathematics in the natural sciences", *Communications in Pure and Applied Mathematic* 13 (1960), 1–14.

[20] Wooldridge, M., *Reasoning about Rational Agents*, MIT Press, Cambridge MA, 2000.

Caroline Semmling, Heinrich Wansing
Institute of Philosophy
Dresden University of Technology
01062 Dresden, Germany
Caroline.Semmling@gmx.de
Heinrich.Wansing@tu-dresden.de