# Logics for Intelligent Agents and Multi-Agent Systems

Author: John-Jules Ch. Meyer
Reviewer: Michael J. Wooldridge

March 3, 2014

**Abstract**

This chapter presents the history of the application of logic in a quite popular paradigm in contemporary computer science and artificial intelligence, viz. the area of intelligent agents and multi-agent systems. In particular we discuss the logics that have been used to specify single agents, the so-called BDI logics, modal logics that describe the beliefs, desires and intentions of agents, after which we turn to logics that are used for specifying multi-agent systems. On the one hand these include extensions of BDI-like logics for multiple agents such as common knowledge and mutual intention, on the other hand, when there are multiple agents into play there are also issues to be dealt with that go beyond these extended individual attitudes, such as normative and strategic reasoning. We sketch also the history of this field.

## 1 Introduction

In this chapter we present the history of logic as applied in the area of intelligent agents and multi-agent systems [122, 120]. This is a quite popular field in between computer science and artificial intelligence (AI). Intelligent agents are software entities that display a certain form of intelligence and autonomy, such as reactivity, proactivity and social behaviour (the latter if there are multiple agents around in a so-called multi-agent system, sharing the same environment) [120]. Single agents are commonly described by so-called BDI logics, modal logics that describe the beliefs, desires and intentions of agents [122, 119, 121], inspired by the work of the philosopher Bratman [14, 15].

Next we turn to logics for multi-agent systems. First we will look at extensions of BDI-like attitudes for situations where there are multiple agents involved. These include notions such as common knowledge and mutual intentions. But also new notions arise when we have multiple agents around. We will look particularly at normative and strategic reasoning in multi-agent systems and the logics to describe this. But we will begin with a short introduction to modal logic which plays a very important role in most of the logics that we will encounter.

## 2 Modal logic

Modal logic is stemming from analytical philosophy to describe and analyze important philosophical notions such as knowledge and belief (epistemic / doxastic logic), time (temporal / tense logic), action (dynamic logic) and obligations, permission and prohibitions (deontic logic) [8]. Historically, modal logics were developed by philosophers in the 20th century, first only in the form of calculi but from the 50's also with a semantics, due to Kripke, Kanger and Hintikka.

The beautiful thing is that these logics all have a similar semantics, called possible world or Kripke semantics and revolve around a box operator $\Box$ and its dual diamond $\diamond$ as additions to classical (propositional or first-order) logic. In a neutral reading the box operator reads as 'necessarily' and the diamond as 'possibly', but in the various uses of modal logic the box operator gets interpretations such as 'it is known / believed that', 'always in the future', 'after the action has been performed it is necessarily the case that', it is obligatory / permitted / forbidden that'. In the propositional case where a set of $AT$ of atomic propositions is assumed, the semantics is given by a Kripke model $\langle S, R, \pi \rangle$ consisting of a set $S$ of possible worlds, a binary, so-called accessibility relation $R$ on $S$, and a truth assignment function $\pi$ yielding the truth or falsity of an atomic proposition per possible world.

The general clause for the semantics of the box operator is truth in *all* accessible worlds: for a model $M$ and a world $s$ occurring in the model:

$$M, s \models \Box\varphi \;\Leftrightarrow\; M, t \models \varphi \text{ for all } t \text{ with } R(s,t)$$

The diamond operator means truth is *some* accessible world:

$$M, s \models \diamond\varphi \;\Leftrightarrow\; M, t \models \varphi \text{ for some } t \text{ with } R(s,t)$$

A formula is valid (with respect to a class of models) if the formula is true in every model and state (of that class of models). Kripke semantics gives rise to certain validities such as

$$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$$

(called the K-axiom) or equivalently

$$(\Box\varphi \wedge \Box(\varphi \to \psi)) \to \Box\psi$$

Using the modal set-up for the various concepts mentioned above leads to logics with different properties for the box operators: for instance for knowledge, where we normally write 'K' for the box operator, we have:

- $K\varphi \to \varphi$, knowledge is true
- $K\varphi \to KK\varphi$, knowledge is known
- $\neg K\varphi \to K\neg K\varphi$, ignorance is known

The last one, called negative introspection, is controversial amongst philosophers, but rather popular among computer scientists and AI researchers. (The resulting logic is called S5.) When we turn to belief (denoted by a 'B') we see that belief enjoys the same properties of belief being believed and disbelief being believed, but belief need not be true. But it is generally held that belief (of a rational agent) should be consistent:

- $\neg B\mathtt{ff}$, belief is consistent

(The resulting logic is called KD45.) Semantically this means that the models must satisfy certain properties, such as reflexive accessibility relations for the first formula of knowledge to become valid, transitive accessibility relations for the second formula of knowledge to become valid, and euclidean accessibility relations for the third formula of knowledge to become valid. And the accessibility relation need to be serial, if the above-mentioned formula for belief is to become valid. (Seriality means that in any world of the model there is at least one successor state with respect to the accessibility relation.) In general there is a theory, called correspondence theory, that studies the relation between properties of the models (or rather frames, which are "models without truth assignment function") and validities in those models [7].

In the rest of this chapter we will see the ubiquitous use of modal logic in the field of logics for intelligent agents and multi-agent systems. But first we will consider an alternative semantics for modal logic, that sometimes is used as well.

## 2.1 Neighbourhood semantics for modal logic

While 'normal' (Kripke) models of modal logic give rise to already a number of validities that are sometimes unwanted (such as $(K\varphi \wedge K(\varphi \rightarrow \psi)) \rightarrow K\psi$, in the case of knowledge, which is sometimes referred to (part of) the logical omniscience problem), there are also so-called minimal (or Scott-Montague) models, based on the notion of neighbourhoods [104, 91, 24]. A minimal / neighbourhood model is a structure $\langle S, N, \pi \rangle$, where $S$ is a set of possible worlds and $\pi$ is a truth assignment function per world again, but $N$ is now a mapping from $S$ to sets of subsets of $S$ (these subsets are called neighbourhoods). The truth of a box and diamond operators is now given by, for model $M = \langle S, N, \pi \rangle$ and world $s$:

$$M, s \models \Box\varphi \;\Leftrightarrow\; \|\varphi\|_M \in N(s)$$

$$M, s \models \Diamond\varphi \;\Leftrightarrow\; S \backslash \|\varphi\|_M \notin N(s)$$

where $\|\varphi\|_M = \{s \in S \mid M, s \models \varphi\}$, the truth set of $\varphi$ in $M$, and $\backslash$ stands for the set-theoretic difference. So, in this semantics $\Box\varphi$ is true in $s$ if the truth set of $\varphi$ is a neighbourhood in $s$. Validity of a formula is defined as that formula being true in every minimal / neighbourhood model and every possible world in that model. This semantics gives rise to a weaker logic. In particular, it does not validate the K-axiom and when used for knowledge there is no logical omniscience (in the traditional sense) anymore. Actually what still holds is something very weak:

$$\models \varphi \leftrightarrow \psi \;\Rightarrow\; \models \Box\varphi \leftrightarrow \Box\psi$$

It is possible, though, to restore the validities for knowledge and belief as mentioned above by putting certain constraints on the models again [24].

# 3 Specifying single agent's attitudes: BDI logics

At the end of the 1980's, the philosopher Michael E. Bratman published a remarkable book, "Intention, Plans, and Practical Reason" [14], in which he lays down a theory of how people make decisions and take action. Put very succinctly, Bratman advocates the essential use of a notion of *intention* besides belief and desire in such a theory. Even more remarkably, although intended to be a theory of human decisions, it was almost immediately picked up by AI researchers to investigate its use for describing artificial agents, which was a new incarnation of the ideal of AI that originated in the 1950's as a discipline aiming at creating artifacts that are able to behave intelligently while performing complex tasks. The logician and mathematician Alan Turing was one of the founding fathers of this discipline: he wrote his famous article "Computing Machinery and Intelligence" [116] where he tries to answer the question "Can machines think?" and subsequently proposes an imitation game as a test for intelligence for a machine (later called the Turing test). The area of AI has, with ups and downs, developed into a substantial body of knowledge of how to do / program intelligent tasks, and comprises such areas as search, reasoning, planning, and learning [102].

Although the notion of an agent abounds in several areas of science and philosophy for quite some time, the concept of an artificial intelligent agent is relatively new and originates at the end of the 1980s and the work of Bratman is an important source of coining this concept. Particularly computer scientists / AI researchers such as David Israel and Martha Pollack [16], Phil Cohen and Hector Levesque [28] and Anand Rao and Michael Georgeff [97] have taken the ideas of Bratman as a starting point and have thought about how to realize artifacts that take decisions in a human way. To this end some of them devised logics to specify the behaviour of the to be constructed agents and tried to follow Bratman's ideas resulting in formalisations of (parts of) Bratman's theory.

These logics are now called BDI logics, since they (mainly) describe the attitudes of beliefs, desires and intentions of intelligent agents. Particularly the notion of an *intention* was advocated by Bratman. Very briefly, intentions are the desires that an agent chooses to commit to and he will not give up this intention unless there is a rational reason for doing so. (This provides a key link between beliefs and actions!)

An agent abandons an intention only under the following conditions:

- the intention has been achieved;
- he believes it is impossible to achieve it;
- he abandons another intention for which the current one is instrumental (so, the current intention loses its purpose).

We will treat these logics briefly in the following subsections. We will do so without giving the sometimes rather complicated semantic models. For these we refer to the original papers as well as several handbook articles [89, 86].

## 3.1 Cohen and Levesque's approach to intentions

Cohen and Levesque attempted to formalize Bratmen's theory in an influential paper "Intention is Choice with Commitment" [28]. This formalisation is based on linear time temporal logic [94], augmented with modalities for beliefs and goals, and operators dealing with actions. It is presented as a 'tiered' formalism with as atomic layer beliefs, goals, actions, and as molecular layer concepts defined in terms of primitives, such as achievement goals, persistent goals and, ultimately, intention in two varieties: $INTEND_1$ (intention_to_do) and $INTEND_2$ (intention_to_be.[1]). So given modal operators for goals and beliefs which are of the KD (cf. Section 4.2) and KD45 kind, respectively, they define achievement goals, persistent goals and intentions in the following way.

As mentioned above the logical language of Cohen & Levesque contains layers, and starts out from a core layer with operators $BEL$ for belief, $GOAL$ for (a kind of primitive notion of) goal, along with a number of other auxiliary operators. These include the operators $LATER\,\varphi$, $DONE\,i\,\alpha$, $HAPPENS\,\alpha$, $\Box\varphi$, $BEFORE\varphi\psi$ and test $\varphi?$ with intended meanings 'sometime in the future $\varphi$ but not now', 'the action $\alpha$ has just been performed', 'the action $\alpha$ is next to be performed', 'always in the future $\varphi$', '$\varphi$ is true before $\psi$ is true', and a test on the truth of $\varphi$, respectively, where the latter means that the test is a skip if $\varphi$ is true and fails / aborts if $\varphi$ is false. These operators all have either a direct semantics or are abbreviations in the framework of Cohen and Levesque, but we refer to [28] for further details. Using these basic operators, the following 'derived' operators of achievement goal, persistent goal and two types of intentions, what I call 'intention_to_do' and 'intention_to_be', are defined:

- $A-GOAL\;i\;\varphi = GOAL\;i\;(LATER\;\varphi) \wedge BEL\;i\;\neg\varphi$
- $P{-}GOAL\,i\,\varphi = A{-}GOAL\,i\,\varphi \wedge [BEFORE(BEL\,i\,\varphi \vee BEL\,i\,\Box\neg\varphi) GOAL\,i\,(LATER\,\varphi)]$
- $INTEND_1\;i\;\alpha = P-GOAL\;i\;[DONE\;i\;(BEL\;i\;(HAPPENS\;\alpha))?;\alpha]$
- $INTEND_2\;i\;\varphi = P-GOAL\;i\;\exists a(DONE\;i\;[BEL\;i\;\exists b\;HAPPENS\;i\;b;\varphi?) \wedge \neg GOAL\;i\;\neg HAPPENS\;i\;a;\varphi?]?;a;\varphi?)$

So, the first clause says that achievement goals to have $\varphi$ are goals of having $\varphi$ at a later time but which are currently believed to be false. Persistent goals are achievement goals that before they are given up should be believed to be achieved or believed to be never possible in the

[1] I use here the same terminology as in deontic logic, where there is a distinction between ought_to_do and ought_to_be [125]. Alternatively one could call intention_to_be also intention_to_bring_about.

future. Intention_to_do an action is a persistent goal of having done this action consciously, and intention_to_be in a state where $\varphi$ holds is a persistent goal of consciously having done some action that led to $\varphi$, while the not happening of the actual action leading to $\varphi$ is not an explicit goal of the agent. The last clause is so complicated since it allows for believing some other action leading to $\varphi$ happening than actually was the case, but also preventing that this actual action was undesired by the agent.

In their framework Cohen & Levesque can prove a number of properties that corroborate their approach as a formalisation of Bratman's theory, such as Bratman's screen of admissibilty. Informally this states that prior intentions may influence later intentions, here coined as the property that if the agent intends to do an action $\beta$, and it is always believed that doing an action $\alpha$ prevents doing $\beta$ forever, then the agent should not intend doing $\alpha$ first and then $\beta$.

- (screen of admissibility)

$$INTEND_1\ i\ \beta \wedge \Box(BEL\ i\ [DONE\ i\ \alpha \rightarrow \Box\neg DONE\ i\ \beta]) \rightarrow \neg INTEND_1\ i\ \alpha;\beta$$

Although I believe the approach of Cohen & Levesque plays an important historical role in obtaining a formal theory of intentions, especially methodologically, trying to define notions of intention from more primitive ones, of course there are also some limitations and considerations of concern. Firstly, by its very methodology, it goes against the aspect of Bratman's philosophy that amounts to the irreducibility of intentions to beliefs and desires! Moreover, the logic is based on linear-time temporal logic, which does not provide the opportunity to talk about quantifying over several possible future behaviours in a syntactic way within the logic. This is remedied by the approach of Rao & Georgeff that uses branching-time temporal logic with the possibility of using path quantifiers within the logic.

## 3.2 Rao & Georgeff's BDI logic

Rao & Georgeff came up with a different formalisation of Bratman's work [97, 98]. This formalisation and the one by Cohen & Levesque have in common that intentions are a kind of special goals that are committed to, and not given up to soon, but the framework as well as the methodology is different. Rather than using linear time temporal logic like Cohen and Levesque do, Rao and Georgeff employ a branching time temporal logic (viz. CTL*, which originated in computer science to describe nondeterministic and parallel processes [26, 43]). Another difference is the method that they use. Rather than having a tiered formalism where intention is defined in terms of other more primitive notions, they introduce primitive modal (box-like) operators for the notions beliefs, goals (desires) and intentions. And then later they put constraints on the models such that there are meaningful interactions between these modalities. So this is much more in line with Bratman's irreducibility of intentions to beliefs and desires. The properties they propose are the following:

(In the following $\alpha$ is used to denote so-called *O-formulas*, which are formulas that contain no positive occurrences of the '*inevitable*' operator (or negative occurrences of '*optional*") outside the scope of the modal operators $BEL, GOAL$ and $INTEND$. A typical O-formula is *optional* $p$, where $p$ is an atomic formula. Furthermore $\varphi$ ranges over arbitrary formulas and $e$ ranges over actions.)

1. $GOAL(\alpha) \rightarrow BEL(\alpha)$
2. $INTEND(\alpha) \rightarrow GOAL(\alpha)$
3. $INTEND(does(e)) \rightarrow does(e)$
4. $INTEND(\varphi) \rightarrow BEL(INTEND(\varphi))$
5. $GOAL(\varphi) \rightarrow BEL(GOAL(\varphi))$

6. $INTEND(\varphi) \rightarrow GOAL(INTEND(\varphi))$
7. $done(e) \rightarrow BEL(done(e))$
8. $INTEND(\varphi) \rightarrow inevitable \diamond (\neg INTEND(\varphi))$

Let us now consider these properties deemed desirable by Rao & Georgeff again. The first formula describes Rao & Georgeff's notion of 'strong realism' and constitutes a kind of belief-goal compatibility: it says that the agent believes he can optionally achieve his goals. There is some controversy on this. Interestingly, but confusingly, Cohen & Levesque [28] adhere to a form of realism that renders more or less the converse formula $BELp \rightarrow GOALp$. But we should be careful and realize that Cohen & Levesque have a different logic in which one cannot express options as in the branching-time framework of Rao & Georgeff. Furthermore, it seems that in the two frameworks there is a different understanding of goals (and beliefs) due to the very difference in ontologies of time employed: Cohen & Levesque's notion of time could be called 'epistemically nondeterministic' or 'epistemically branching', while 'real' time is linear: the agents envisage several future courses of time, each of them being a linear history, while in Rao & Georgeff's approach also 'real' time is branching, representing options that are available to the agent.

The second formula is a similar one to the first. This one is called goal-intention compatibilty, and is defended by Rao & Georgeff by stating that if an optionality is intended it should also be wished for (a goal in their terms). So, Rao & Georgeff have a kind of selection filter in mind: intentions (or rather intended options) are filtered / selected goals (or rather goal (wished) options), and goal options are selected believed options. If one views it this way, it looks rather close to Cohen & Levesque's "Intention is choice (chosen / selected wishes) with commitment", or loosely, wishes that are committed to. Here the commitment acts as a filter.

The third one says that the agent really does the primitive actions that s/he intends to do. This means that if one adopts this as an axiom the agent is not allowed to do something else (first). (In our opinion this is rather strict on the agent, since it may well be that postponing executing its intention for a while is also an option.) On the other hand, as Rao & Georgeff say, the agent may also do things that are not intended since the converse does not hold. And also nothing is said about the intention to do complex actions.

The fourth, fifth and seventh express that the agent is conscious of its intentions, goals and what primitive action he has done in the sense that he believes what he intends, has as a goal and what primitive action he has just done.

The sixth one says something like that intentions are really wished for: if something is an intention then it is a goal that it is an intention.

The eighth formula states that intentions will inevitably (in every possible future) be dropped eventually, so there is no infinite deferral of the agent's intentions. This leaves open, whether the intention will be fulfilled eventually, or will be given up for other reasons. Below we will discuss several possibilities of giving up intentions according to different types of commitment an agent may have.

It is very interesting is that BDI-logical expressions can be used to characterize different types of agents. Rao & Georgeff mention the following possibilities:

1. (blindly committed agent) $INTEND(inevitable \diamond \varphi) \rightarrow$
   $inevitable(INTEND(inevitable \diamond \varphi) \mathsf{U} BEL(\varphi))$
2. (single-minded committed agent) $INTEND(inevitable \diamond \varphi) \rightarrow$
   $inevitable(INTEND(inevitable \diamond \varphi) \mathsf{U} (BEL(\varphi) \vee \neg BEL(optional \diamond \varphi)))$
3. (open minded committed agent) $INTEND(inevitable \diamond \varphi) \rightarrow$
   $inevitable(INTEND(inevitable \diamond \varphi) \mathsf{U} (BEL(\varphi) \vee \neg GOAL(optional \diamond \varphi)))$

A blindly committed agent maintains his intentions to inevitably obtaining eventually something until he actually believes that that something has been fulfilled. A single-minded committed agent is somewhat more flexible: he maintains his intention until he believes he has achieved it *or he does not believe that it can be reached (i.e. that it is still an option in some future) anymore.* Finally, the open minded committed agent is even more flexible: he can also drop his intention if it is not a goal (desire) anymore.

Rao & Georgeff are then able to obtain results under which conditions the various types of committed agents will reach their intentions. For example, for a blindly committed agent it holds that under the assumption of the axioms we have discussed earlier including the axiom that expresses no infinite deferral of intentions [2] :

$$INTEND(\varphi) \rightarrow inevitable \diamond \neg INTEND(\varphi)$$

that

$$INTEND(inevitable(\diamond\varphi)) \rightarrow inevitable(\diamond BEL(\varphi))$$

expressing that if the agent intends to eventually obtain $\varphi$ it will inevitably eventually believe that it has succeeded in achieving $\varphi$.

In his book [120] Michael Wooldridge has extended $BDI_{CTL}$ to define LORA (the Logic Of Rational Agents), by incorporating an action logic. Interestingly the way this is done resembles Cohen & Levesque's logic as to the syntax (with operators such as $HAPPENS\ \alpha$ for actions $\alpha$), but the semantics is branching-time *à la* Rao & Georgeff. In principle, LORA allows reasoning not only about individual agents, but also about communication and other interaction in a multi-agent system, so we will return to LORA, when we will look at logics for multi-agent systems.

## 3.3 KARO Logic

The KARO formalism is yet another formalism to describe the BDI-like mental attitudes of intelligent agents. In contrast with the formalisms of Cohen & Levesque and Rao & Georgeff its basis is dynamic logic [52, 53], which is a logic of action, augmented with epistemic logic (there are modalities for knowledge and belief). On this basis the other agent notions are built. The KARO framework has been developed in a number of papers (e.g. [72, 73, 58, 88]) as well as the thesis of Van Linder ([71]). Again we suppress semantical matters here.

The KARO formalism is an amalgam of dynamic logic and epistemic / doxastic logic [87], augmented with several additional (modal) operators in order to deal with the motivational aspects of agents. So, besides operators for knowledge (**K**), belief (**B**) and action ($[\alpha]$, "after performance of $\alpha$ it holds that"), there are additional operators for ability (**A**) and desires (**D**). Perhaps the *ability* operator is the most nonstandard one. It takes an action as an argument, expressing that the agent is able to perform that action. This is to be viewed as an intrinsic property of the agent. For example a robot with a gripper is able to grip. Whether the agent has also the *opportunity* to perform the action depends on the environment. In the example of the robot with gripper it depends on the enviroment whether there are things to grip. In KARO ability and opportunity are represented by different operators. We will see the opportunity operator directly below.

In KARO a number of operators are defined as abbreviations:

[2] As the reviewer of this paper observed, this would only work for non-valid / non-tautological assertions $\varphi$, since INTEND being a normal box-like operator satisfies the necessitation rule, thus causing inconsistency together with this axiom. On the other hand, a tautological or valid assertion is obviously not a true achievement goal, so exclusion of the axiom for this case is not a true restriction, conceptually speaking.

- (dual) $\langle\alpha\rangle\varphi = \neg[\alpha]\neg\varphi$, expressing that the agent has the opportunity to perform $\alpha$ resulting in a state where $\varphi$ holds.
- (opportunity) $\mathbf{O}\alpha = \langle\alpha\rangle\mathtt{tt}$, i.e., an agent has the opportunity to do an action iff there is a successor state w.r.t. the $R_\alpha$-relation;
- (practical possibility) $\mathbf{P}(\alpha,\varphi) = \mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$, i.e., an agent has the practical possibility to do an action with result $\varphi$ iff it is both able and has the opportunity to do that action and the result of actually doing that action leads to a state where $\varphi$ holds;
- (can) $\mathbf{Can}(\alpha,\varphi) = \mathbf{KP}(\alpha,\varphi)$, i.e., an agent can do an action with a certain result iff it knows it has the practical possibilty to do so;
- (realisability) $\Diamond\varphi = \exists a_1,\ldots,a_n\mathbf{P}(a_1;\ldots;a_n,\varphi)$[3], i.e., a state property $\varphi$ is realisable iff there is a finite sequence of atomic actions of which the agent has the practical possibility to perform it with the result $\varphi$;
- (goal) $\mathbf{G}\varphi = \neg\varphi \wedge \mathbf{D}\varphi \wedge \Diamond\varphi$, i.e., a goal is a formula that is not (yet) satisfied, but desired and realisable.[4]
- (possible intend) $\mathbf{I}(\alpha,\varphi) = \mathbf{Can}(\alpha,\varphi) \wedge \mathbf{KG}\varphi$, i.e., an agent (possibly) intends an action with a certain result iff the agent can do the action with that result and it moreover knows that this result is one of its goals.

Informally, these operators mean the following:

- The dual of the (box-type) action modality expresses that there is at least a resulting state where a formula $\varphi$ holds. It is important to note that in the context of *deterministic* actions, i.e. actions that have at most one successor state, this means that the *only* state satisfies $\varphi$, and is thus in this particular case a stronger assertion than its dual formula $[\alpha]\varphi$, which merely states that if there are any successor states they will (all) statisfy $\varphi$.
- Opportunity to do an action is modelled by having at least one successor state according to the accessibility relation associated with the action.
- Practical possibility to do an action with a certain result is modelled as having both ability and opportunity to do the action with the appropriate result. Note that $\mathbf{O}\alpha$ in the formula $\mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$ is actually redundant since it already follows from $\langle\alpha\rangle\varphi$. However, to stress the opportunity aspect it is added.
- The **Can** predicate applied to an action and formula expresses that the agent is 'conscious' of its practical possibility to do the action resulting in a state where the formula holds.
- A formula $\varphi$ is realisable if there is a 'plan' consisting of (a sequence of) atomic actions of which the agent has the practical possibility to do them with $\varphi$ as a result.
- A formula $\varphi$ is a goal in the KARO framework if it is not true yet, but desired and realisable in the above meaning, that is, there is a plan of which the agent has the practical possibility to realise it with $\varphi$ as a result.
- An agent is said to (possibly) intend an action $\alpha$ with result $\varphi$ if it 'Can' do this (knows that it has the practical possibility to do so), and, moreover, knows that $\varphi$ is a goal.

[3] We abuse our language here slightly, since strictly speaking we do not have quantification in our object language. See [88] for a proper definition.

[4] In fact, here we simplify matters slightly. In [88] we also stipulate that a goal should be explicitly selected somehow from the desires it has, which is modelled in that paper by means of an additional modal operator. Here we leave this out for simplicity's sake.

In order to manipulate both knowledge / belief and motivational matters special actions `revise`, `commit` and `uncommit` are added to the language. (We assume that we cannot nest these operators. So, e.g., $\texttt{commit}(\texttt{uncommit}\alpha)$ is not a well-formed action expression. For a proper definition of the language the reader is referred to [88].) Moreover, the formula $\mathbf{Com}(\alpha)$ is introduced to indicate that the agent is committed to $\alpha$ ("has put it on its agenda, i.e. literally, things to do").

Defining validity on the basis of the models of this logic [72, 73, 88] one obtains the following typical properties (cf. [72, 88]):

1. $\models \mathbf{A}(\alpha;\beta) \leftrightarrow \mathbf{A}\alpha \wedge [\alpha]\mathbf{A}\beta$
2. $\models \mathbf{Can}(\alpha;\beta,\varphi) \leftrightarrow \mathbf{Can}(\alpha,\mathbf{P}(\beta,\varphi))$
3. $\models \mathbf{I}(\alpha,\varphi) \rightarrow \mathbf{K}\langle\alpha\rangle\varphi$
4. $\models \mathbf{I}(\alpha,\varphi) \rightarrow \langle\texttt{commit}\alpha\rangle\mathbf{Com}(\alpha)$
5. $\models \mathbf{I}(\alpha,\varphi) \rightarrow \neg\mathbf{A}\texttt{uncommit}(\alpha)$
6. $\models \mathbf{Com}(\alpha) \rightarrow \langle\texttt{uncommit}(\alpha)\rangle\neg\mathbf{Com}(\alpha)$
7. $\models \mathbf{Com}(\alpha) \wedge \neg\mathbf{Can}(\alpha,\top) \rightarrow \mathbf{Can}(\texttt{uncommit}(\alpha),\neg\mathbf{Com}(\alpha))$
8. $\models \mathbf{Com}(\alpha) \rightarrow \mathbf{KCom}(\alpha)$
9. $\models \mathbf{Com}(\alpha_1;\alpha_2) \rightarrow \mathbf{Com}(\alpha_1) \wedge \mathbf{K}[\alpha_1]\mathbf{Com}(\alpha_2)$

The first of these properties says that if the agent is able to do the sequence $\alpha;\beta$, then this is equivalent with that the agent is able to do $\alpha$ and after doing $\alpha$ it is able to do $\beta$, which sounds very reasonable, but see the remark on this below. The second states that an agent *can* do a sequential composition of two actions with result $\varphi$ iff the agent can do the first actions resulting in a state where it has the practical possibility to do the second with $\varphi$ as result. The third states that if one possibly intends to do $\alpha$ with result $\varphi$ then one knows that there is a possibility of performing $\alpha$ resulting in a state where $\varphi$ holds. The fourth asserts that if an agent possibly intends to do $\alpha$ with some result $\varphi$, it has the opportunity to commit to $\alpha$ with result that it is committed to $\alpha$ (i.e. $\alpha$ is put into its agenda). The fifth says that if an agent intends to do $\alpha$ with a certain purpose, then it is unable to uncommit to it (so, if it is committed to $\alpha$, it has to persevere with it). This is the way persistence of commitment is represented in KARO. Note that this is much more 'concrete' (also in the sense of computability) than the persistence notions in the other approaches we have seen, where temporal operators pertaining to a possibly infinite future were employed to capture them...! In KARO we have the advantage of having dedicated actions in the action language dealing with the change of commitment that can be used to express persistence without referring to the (infinite) future, rendering the notion of persistence much 'more computable'. The sixth property says that if an agent is committed to an action and it has the opportunity to uncommit to it then indeed the commitment is removed as a result. The seventh says that whenever an agent is committed to an action that is no longer known to be practically possible, it knows that it can undo this impossible commitment. The eighth property states that commitments are known to the agent. The nineth says that if an agent is committed to a sequential composition of two actions then it is committed to the first one, and it knows that after doing the first action it will be committed to the second action.

KARO logic has as a core notion that of ability. But in the above treatment this only works well for non-failing deterministic actions. Since it is a validity in KARO that $\models \mathbf{A}(\alpha;\beta) \leftrightarrow \mathbf{A}\alpha \wedge [\alpha]\mathbf{A}\beta$, we get the undesirable result that in case there is no opportunity to do $\alpha$, the agent is able to do $\alpha;\beta$ for arbitrary $\beta$. For instance, if a lion is locked in a cage and would be able to walk out but lacks the opportunity, it is able to get out and fly away! The problem here is a kind of undesired entanglement of ability and opportunities. In [59] we extend our theory of

ability to nondeterministic actions. (Another solution is to separate results and opportunities on the one hand and abilities on the other hand rigorously by using two dynamic operators $[\alpha]_1$ and $[\alpha]_2$ for dealing with results with respect to opportunities and abilities, respectively, which we have described in [103]). Finally we mention that also related notions such as attempt and failure of actions have been studied in the literature (e.g., [80, 19]).

# 4 Logics for multi-agent systems

## 4.1 Multi-agent logics

In the previous sections we have concentrated on single agents and how to describe them. In this subsection we will look at two generalisations of single-agent logics to multi-agent logics, viz. multi-agent epistemic logic and multi-agent BDI logic.

### 4.1.1 Multi-agent epistemic logic

In a multi-agent setting one can extend a single-agent framework in several ways. To start with, with respect to the epistemic (doxastic) aspect, one can introduce epistemic (doxastic) operators for every agent, resulting in a multi-modal logic, called $\mathbf{S5}_n$. Models for this logic are inherently less simple and elegant as those for the single agent case (cf. [44, 87]). So then one has indexed operators $\mathbf{K}_i$ and $\mathbf{B}_i$ for agent $i$'s knowledge and belief, respectively. But one can go on and define knowledge operators that involve a group of agents in some way. This gives rise to the notions of common and (distributed) group knowledge.

The simplest notion is that of 'everybody knows', here denoted by the operator $\mathbf{E}_\mathbf{K}$. But one can also add an operator $\mathbf{C}_\mathbf{K}$ for 'common knowledge', which is much more powerful. Although I'll leave out the details of the semantics again, it is worth mentioning that the semantics of the common knowledge operator is given by the reflexive-transitive closure of the union of the accessibility relations of the individual agents. So it is a powerful operator that quantifies over all states reachable through the accessibility relations associated with the individual agents. This gives the power of analyzing the behavior of the agent in multi-agent systems such as communication between agents in a setting where communication channels are unreliable, like in the case of the Byzantine generals sending messages to each other about a joint attack, where it appears that under circumstances of sending messengers through enemy-controlled territory there cannot emerge common knowledge of this attack proposal without which the attack cannot safely take place! This phenomenon, known as the Coordinated Attack Problem, also has impact on more technical cases involving distributed (computer) systems, where in fact the problem originated from [49, 44, 87].

By extending the models and semantic interpretation appropriately (see, e.g., [44, 87]) we then obtain the following properties (assuming that we have $n$ agents):

- $\mathbf{E}_\mathbf{K}\varphi \leftrightarrow \mathbf{K}_1\varphi \wedge \ldots \wedge \mathbf{K}_n\varphi$
- $\mathbf{C}_\mathbf{K}(\varphi \rightarrow \psi) \rightarrow (\mathbf{C}_\mathbf{K}\varphi \rightarrow \mathbf{C}_\mathbf{K}\psi)$
- $\mathbf{C}_\mathbf{K}\varphi \rightarrow \varphi$
- $\mathbf{C}_\mathbf{K}\varphi \rightarrow \mathbf{C}_\mathbf{K}\mathbf{C}_\mathbf{K}\varphi$
- $\neg\mathbf{C}_\mathbf{K}\varphi \rightarrow \mathbf{C}_\mathbf{K}\neg\mathbf{C}_\mathbf{K}\varphi$
- $\mathbf{C}_\mathbf{K}\varphi \rightarrow \mathbf{E}_\mathbf{K}\mathbf{C}_\mathbf{K}\varphi$
- $\mathbf{C}_\mathbf{K}(\varphi \rightarrow \mathbf{E}_\mathbf{K}\varphi) \rightarrow (\varphi \rightarrow \mathbf{C}_\mathbf{K}\varphi)$

The first statement of this proposition shows that the 'everybody knows' modality is indeed what its name suggests. The next four says that common knowledge has at least the properties of knowledge: closed under implication, it is true, and enjoys the introspective properties. The sixth property says that common knowledge is known by everybody. The last is a kind of induction principle: the premise gives the condition under which one can 'upgrade' the truth of $\varphi$ to common knowledge of $\varphi$; this premise expresses that it is common knowledge that the truth of $\varphi$ is known by everybody.

As a side remark we note that these properties, in particular the last two ones, are of the exactly the same form as those axiomatizing dynamic logic [52, 53]. This is explained by the fact that the C-operator is based on a reflexive-transitive closure of the underlying accessibility relation as it is the case with the $[\alpha^*]$ operator in dynamic logic. A further interesting link is that with *fixed point theory* dating back to Tarski [Tar55]. One can show (see e.g., [44]) that $\mathbf{C_K}\varphi$ is a greatest fixed point of the (monotone) function $\mathbf{E_K}(\varphi \wedge x)$. This implies that from $\varphi \to \mathbf{E_K}\varphi$ one can derive $\varphi \to \mathcal{C}\varphi$ ([44], page 408, bottom line, with $\psi = \varphi$), which is essentially the same as the last property shown above, stated as an assertion rather than a rule. (Note that a rule 'from $\varphi$ derive $\chi$' in modal logic with a 'reflexive $\Box$ operator' has the same meaning as a rule 'from $\Box\varphi$ derive $\chi$'.)

As to multi-agent doxastic logic one can look at similar notions of 'everybody believes' and common belief. One can introduce operators $\mathbf{E_B}$ and $\mathbf{C_B}$ for these notions. Now we obtain a similar set of properties for common belief (cf. [67, 87]):

- $\mathbf{E_B}\varphi \leftrightarrow \mathbf{B}_1\varphi \wedge \ldots \wedge \mathbf{B}_n\varphi$
- $\mathbf{C_B}(\varphi \to \psi) \to (\mathbf{C_B}\varphi \to \mathbf{C_B}\psi)$
- $\mathbf{C_B}\varphi \to \mathbf{E_B}\varphi$
- $\mathbf{C_B}\varphi \to \mathbf{C_B}\mathbf{C_B}\varphi$
- $\neg\mathbf{C_B}\varphi \to \mathbf{C_B}\neg\mathbf{C_B}\varphi$
- $\mathbf{C_B}\varphi \to \mathbf{E_B}\mathbf{C_B}\varphi$
- $\mathbf{C_B}(\varphi \to \mathbf{E_B}\varphi) \to (\mathbf{E_B}\varphi \to \mathbf{C_B}\varphi)$

Note the differences with the case for knowledge due to the fact that common belief is not based on a reflexive accessibility relation (speaking semantically). In more plain terms, common belief, like belief, need not be true.

### 4.1.2 Multi-agent BDI logic

Also with respect to the other modalities one may consider multi-agent aspects. In this subsection we focus on the notion of collective or joint intentions. We follow ideas from [41] (but we give a slightly different but equivalent presentation of definitions). We now assume that we have belief and intention operators $\mathbf{B}_i, \mathbf{I}_i$ for every agent $1 \leq i \leq n$. First we enrich the language of multi-agent doxastic with operators $\mathbf{E_I}$ (everybody intends) and $\mathbf{M_I}$ (mutual intention). (We call this a multi-agent BDI logic, although multi-agent BI logic would be a more adequate name, since we leave out the modality of desire / goal.)

Now we get similar properties for mutual intention as we had for common belief (but of course no introspective properties):

- $\mathbf{E_I}\varphi \leftrightarrow \mathbf{I}_1\varphi \wedge \ldots \wedge \mathbf{I}_n\varphi$
- $\mathbf{M_I}(\varphi \to \psi) \to (\mathbf{M_I}\varphi \to \mathbf{M_I}\psi)$
- $\mathbf{M_I}\varphi \to \mathbf{E_I}\varphi$
- $\mathbf{M_I}\varphi \to \mathbf{E_I}\mathbf{M_I}\varphi$

- $\mathbf{M_I}(\varphi \rightarrow \mathbf{E_I}\varphi) \rightarrow (\mathbf{E_I}\varphi \rightarrow \mathbf{M_I}\varphi)$

We see that E-intentions ('everybody intends') and mutual intentions are defined in a way completely analogous with E-beliefs ('everybody believes') and common beliefs, respectively. Next Dunin-Kęplicz & Verbrugge ([41]) define the notion of *collective intention* ($\mathbf{C_I}$) as follows:

- $\mathbf{C_I}\varphi \ = \ \mathbf{M_I}\varphi \wedge \mathbf{C_B}\mathbf{M_I}\varphi$

This definition states that collective intentions are those formulas that are mutually intended and of which this mutual intention is a common belief amongst all agents in the system.

We must mention here that in the literature there is also other work on BDI-like logics for multi-agent systems where we encounter such notions as joint intentions, joint goals and joint commitments, mostly coined in the setting of how to specify teamwork. Seminal work was done by Cohen & Levesque [29]. This work was a major influence on our own multi-agent version of KARO [1]. An important complication in a notion of joint goal involves that of persistence of the goal: where in the single agent case the agent pursues its goal until it believes it has achieved it or believes it can never be achieved, in the context of multiple agents, the agent that realizes this, has to inform the others of the team about it so that the group / team as a whole will believe that this is the case and may drop the goal.

Next we consider Wooldridge's LORA [120] again. As we have seen before LORA is a branching-time BDI logic combined with a (dynamic logic-like) action logic in the style of Cohen & Levesque. But from Chapter 6 onwards of [120], Wooldridge also considers multi-agent aspects: collective mental states (mutual beliefs, desires, intentions, similar to what we have seen above), communication (including speech acts as rational actions) and cooperation (with notions such as ability, team formation and plan formation). It is fair to note here that a number of these topics were pioneered by Singh [108, 109, 110, 111, 112, 113].

An interesting, rather ambitious recent development is [36]. In this paper a logic, LOA (Logic of Agent Organizations), is proposed in which a number of matters concerning agent organizations are combined. The logic is based on the branching-time temporal logic CTL*. It furthermore has operators for agent capability, ability, agent attempt, agent control and agent activity, that are subsequently lifted to group notions: (joint) capability, ability, attempt, in-control and stit (seeing to it that) [25]. With this framework the authors are able to express important MAS notions such as responsibility, initiative, delegation and supervision. For example, a supervision duty is formalized as follows. Given an organization, and group of roles $Z$ that is part of the organization and a group of agents $A$ playing the roles $U$ in the organization, the supervising duty of roles $Z$ with respect to the group of agents $V$ to realize $\varphi$ is defined as:

$$SD_{(Z,V)}\varphi \ =_{def} \ (I_Z H_{VU}\varphi \wedge \Diamond(H_{VU}\varphi \wedge X\neg\varphi)) \rightarrow I_Z\varphi$$

where $I_Z\varphi$ stands for $Z$ taking the initiative to achieve $\varphi$, $H_{VU}\varphi$ stands for agents $V$ enacting roles $U$ attempting $\varphi$, $\Diamond$ is the usual 'eventually' operator, and $X$ is the next-time operator. This definition thus states that if $Z$ initiates $V$ to attempt $\varphi$ in their roles $U$ and at some point in time this attempt fails, then the roles $Z$ become directly in charge of achieving $\varphi$.

## 4.2 Logics of norms and normative systems

### Deontic logic

Logics about norms and normative systems have their roots in the philosophical field of deontic logic, where pioneers like Von Wright [123] already tried to formalize a kind of normative reasoning.

The history of deontic logic (as a formal logic) goes back at least as far as modal logic in general, with people like Mally [82] attempting first formalizations of notions such as obligation.

But, despite interesting and laudable attempts to vindicate Mally as a serious deontic logician (e.g. [74, 75, 76] it is generally held that deontic logic started to get serious with the work of Von Wright [123]. In this paper Von Wright proposed an influential system (later to be known as OS, "Old System") that is very close to a normal modal logic (KD), which establishes the operator O (obligation) as a necessity-style operator in a Kripke-style semantics. The characteristic axiom in the system KD is the so-called D-axiom:

$$\neg O\mathtt{ff}$$

that is, obligations are consistent. To have a feeling for this axiom, we mention that it is equivalent with $\neg(Op \wedge O\neg p)$. (In fact this is the same axiom as we have encountered with belief in a previous section.) Semantically it amounts to taking models in which the accessibility relation associated with the $O$-operator is serial. The logic KD is now known as Standard Deontic Logic, and inspired many philosophers, despite or perhaps even due to various paradoxes that could be inferred from the system. [5] Over the years people have come to realise that KD is simply too simple as a deontic logic. In fact, already Von Wright realized this and came up with a "New System" NS, as early as 1964 [124], in which he tried to formalize conditional deontic logic as a 'dyadic' logic (a logic with a two-argument obligation operator $O(p/q)$, meaning that "p is obligatory under condition q"). This gave rise to a renewed study of deontic logic in the 60s and 70s. However, some problems (paradoxes) remained. To overcome these problems there were also approaches based on temporal logic [127, 42, 115, 45]. More recently temporal logic has also been employed to capture the intricacies of deadlines [21]. Meanwhile there were also attempts to reduce deontic logic to alethic modal logic (Anderson, [5]), and from the 80s also a reduction to dynamic logic was proposed [85][6], giving rise to the subfield of dynamic deontic logics. This brings to the fore another important issue in deontic logic, viz. that of ought-to-be versus ought-to-do propositions. In the former approach the deontic operator (such as obligation O) has a proposition as an argument, describing the situation that is at hand (obligatory), while the latter has an action as an argument describing that this action is obligatory. This distinction is sometimes blurred but has also received considerable attention in the deontic logic literature (cf. [84]).

Another refinement of deontic logic has to do with the distinction ideal versus actual. Standard deontic logic distinguishes these by using the deonttc operators: for instance $O\varphi$ says that in every ideal world $\varphi$ holds, while actually it may not be the case (that is to say the formula $\varphi$ does not hold). But, when trying to solve the problems mentioned above, especially pertaining to contrary-to-duties, it may be tempting to look at a more refined distinction in which we have levels of ideality: ideal versus subideal worlds. Approaches along these lines are [39, 23]. A similar line of approach is that taken by Craven and Sergot [30]. In this framework, which comprises a deontic extension of the action logic C+ of Giunchiglia et al. [47]), they ncorporate green/red states as well as green/red transitions, thus rendering a more refined treatment of deontically good and bad behavior. This language is used to describe a labelled transition system and the deontic component provides a means of specifying the deontic status

[5] It goes beyond the scope of this paper to mention all these paradoxes, but to get a feeling we mention one: in SDL it is valid that $Op \rightarrow O(p \vee q)$, which in counterintuitive if one reads this in the following instance: if it is obligatory to mail the letter, then it is obligatory to mail the letter or burn it. What is paradoxical is that in the commonsense reading of this formula it is suggested that it is left to the agent whether he will mail or burn the letter. But this is not meant: it just says that in an ideal world where the agent mails the letter it is (logically) also the case that in this world the agent mails the letter or burns it. Another, major, problem is the contrary-to-duty imperative, which deals with norms that hold when other norms are violated, such as Forrester's paradox of gentle murder [46, 90, 84]

[6] Basically this reduced, for instance, the notion of prohibition as follows: $F\alpha =_{def} [\alpha]V$, where $V$ stands for a violation atom, and $[\alpha]\varphi$ is an expression from dynamic logic, as we saw before when we treated the KARO framework. So prohibition is equated with "leading to violation".

(permitted/acceptable/legal/green) of states and transitions. It features the so called green-green-green (ggg) constraint: a green transition in a green state always leads to a green state. Or, equivalently, any transition from a green state to a red state must itself be red!

Recently there are also approaches to deontic logic using stit theory (seeing to it that, [25, 66]). Deontic stit logics are also logics that deal with actions but contrary to dynamic logic these actions are not named (reified) in the logical language ([6, 62, 18]). Since the 1990s also defeasible / non-monotonic approaches to deontic logic have arisen [117, 92]. Another recent development is that of input/output logic [81] that also takes its origin in the study of conditional norms. In this approach conditional norms are not treated as bearing truth-values as in most deontic logics. Technically in input/output logic conditional norms are viewed as sets of pairs and a normative code as a set of these. Thus, in this approach formally norms themselves have no truth value anymore, but descriptions of normative situations, called normative propositions, have.

### Other logics

A range of other logical formalisms for reasoning about normative aspects of multi-agent systems have also been proposed.

To begin with we mention here combinations of BDI logics with logics about norms (such as deontic logic), such as the BOID and B-DOING frameworks [20, 37]. Also a merge of KARO and deontic logic has been proposed [40, 38]. Of these, the BOID framework is the most well-known. It highlights the interplay between the agent's 'internal' (BDI) versus 'external' motivations (norms / obligations), which enables one to distinguish between several agent types. For example, a benevolent agent will give priority to norms, while an egocentric agent will not. The system is not a logic proper, but rather a rule-based system: the system contains rules (not unlike rules in default logic [99]) which determine extensions of formulas pertaining to beliefs, obligations, intentions and desires to be true. The order of the rules applied to create these extensions depends on the agent type.

An important notion in multi-agent system (MAS) is that of an *institution*. An institution is any structure or mechanism of social order and cooperation governing the behavior of a set of individuals within a given community ([118]). Apart from several computational mechanisms that have been devised for controlling MAS, there have also been proposed dedicated logics to deal with institutional aspects, such as the 'counts-as' conditional, expressing that a 'brute' fact counts as an 'institutional' fact in a certain context. Here the terminology of Searle [105, 106] is used: brute facts pertain to the real world, while institutional facts pertain to the institutional world. Examples are [50]:

- (constitutive) "In system $S$, conveyances transporting people or goods count as vehicles"
- (classificatory) "Always, bikes count as conveyances transporting people or goods"
- (proper classificatory) "In system $S$, bikes count as vehicles"

One of the first truly logical approaches is that by Jones & Sergot [65], which presents a study of the counts-as conditional (in a minimal modal / neighbourhood semantic setting). Their work was later improved by Grossi et al. [51, 50], where the three different interpretations of 'counts-as' mentioned above are disentangled, viz. constitutive, classificatory and proper classificatory. which are all following a different (modal) logic.

Given a modal logic of contexts [87, 78] with modal operators $[c]$ (within context $c$, quantifying over all the possible worlds lying within $c$) and 'universal' context operator $[u]$ (quantifying over all possible worlds, an S5-modality), Grossi formalizes the three counts-as notions $\gamma_1 \Rightarrow^i_c \gamma_2$, with $c$ a context denoting a set of possible worlds (actually this context is given by a formula that we also will denote as $c$: the context is then the set of all possible worlds satisfying formula $c$), as follows:

**constitutive counts-as** for $\gamma_1 \to \gamma_2 \in \Gamma$:

$$\gamma_1 \Rightarrow^1_{c,\Gamma} \gamma_2 \;\; =_{def} \; [c]\Gamma \wedge [\neg c]\neg\Gamma \wedge \neg[u](\gamma_1 \to \gamma_2)$$

**classificatory counts-as**

$$\gamma_1 \Rightarrow^2_c \gamma_2 \;\; =_{def} \; [c](\gamma_1 \to \gamma_2)$$

**proper classificatory counts-as**

$$\gamma_1 \Rightarrow^3_c \gamma_2 \;\; =_{def} \; [c](\gamma_1 \to \gamma_2) \wedge \neg[u](\gamma_1 \to \gamma_2)$$

Here the context $c, \Gamma$ denotes the set of possible worlds within c that satisfy the set of formulas $\Gamma$. So the simplest notion of counts-as is the classificatory counts as, meaning that within the context $c$ $\gamma_1$ just implies $\gamma_2$. Proper classificatory counts-as is classificatory counts-as together with the requirement that the the implication $\gamma_1$ implying $\gamma_2$ should not hold universally. Constitutive counts as w.r.t. the context $c$ together with the formulas $\Gamma$ says that, within the context $c$, $\Gamma$ holds, while outside context $c$ $\Gamma$ does not hold, and moreover, the implication $\gamma_1$ implies $\gamma_2$ is not universally true.

## 4.3 Logics for Strategic Reasoning

In the context of Multi-Agent Systems there has arisen a completely new branch of logics, which has to do with strategies in game-theoretic sense. One of the first of these was Pauly's Coalition Logic [93]. This is basically a modal logic with Scott-Montague (neighbourhood semantics.) [104, 91, 24]. Thus, Coalition Logic employs a modal language with as box operator $[C]\varphi$, where C is a subset of agents, a *coalition*. The reading of this operator is that the coalition $C$ can force the formula $\varphi$ to be true. The interpretation is by employing neighbourhoods, which here take the form of so-called effectivity functions $E : S \to (\mathcal{P}(\mathcal{A}) \to \mathcal{P}(\mathcal{P}(\mathcal{S})))$, where $S$ is the set of possible worlds and $A$ is the set of agents. Intuitively, $E(s)(C)$ is the collection of sets $X \subseteq S$ such that $C$ can force the world to be in some state of $X$ (where $X$ represents a proposition). $[C]\varphi$ is now immediately interpreted by:

$$M, s \models [C]\varphi \;\; \Leftrightarrow \;\; \|\varphi\|_M \in E(s)(C)$$

As Coalition Logic is a form of minimal modal logic, it satisfies:

$$\models \varphi \leftrightarrow \psi \;\; \Rightarrow \;\; \models [C]\varphi \leftrightarrow [C]\psi$$

By putting constraints on the effectivity functions one obtains a number of further validities, for example,

- $\models \varphi \to \psi \;\; \Rightarrow \;\; \models [C]\varphi \to [C]\psi$ iff $E$ is outcome monotonic, i.e. for all $s$, $C$, $E(s)(C)$ is closed under supersets.
- $\models \neg[C]\mathtt{ff}$ iff $\emptyset \notin E(s)(C)$ for every $s \in S$.

In fact, Pauly [93] considers an important class of effectivity functions that he calls *playable*, which is characterized by the following set-theoretic conditions:

- $\emptyset \notin E(s)(C)$ for every $s \in S, \; C \subseteq A$
- $S \in E(s)(C)$ for every $s \in S, \; C \subseteq A$
- if $X \subseteq Y \subseteq S$ and $X \in E(s)(C)$ then $Y \in E(s)(C)$, for all $s \in S, \; C \subseteq A$ (outcome monotonicity)
- if $S \backslash X \notin E(s)(\emptyset)$ then $X \in E(s)(A)$, for every $X \subseteq S, \; s \in S$ ($A$-maximality)

- if $C_1 \cap C_2 = \emptyset,\ X_1 \in E(s)(C_1)$ and $X_2 \in E(s)C_2)$ then $X_1 \cap X_2 \in (E(s)(C_1 \cup C_2)$, for all $C_1, C_2 \subseteq A$ and $X_1, X_2 \subseteq S$ (superadditivity)

Informally playability amounts to that there is always some world in $S$ any coalition can force the world in and that it cannot force the world to be in no state at all; if a coalition can force the world to be in some state in $X$ then it can also force the world to be in some state in a (possibly) bigger set $Y$; if the empty coalition (sometimes referred to as Nature) cannot force the world to be in a state outside $X$ (so Nature doesn't forbid to be in a state of $X$) then the grand coalition of all agents $A$ must be able to force the world in a state of $X$; and finally if disjoint coalitions can force the world to be in states of $X_1$ and $X_2$, respectively, then the union of those coalitions must be able to force the world to be in a state of its intersection $X_1 \cap X_2$. We will return to the concept of playable effectivity functions below.

A much more expressive logic to reason about the strategies of agents is Alternating-Time Temporal Logic (ATL), first proposed by Alur, Henzinger and Kupferman [3]. They introduce ATL as a third variety of temporal logic that contrary to linear-time and branching-time temporal logics that are "natural specification languages for closed systems", are that for open systems. ATL offers "selective quantification over those paths that are possible outcomes of games, such as the game in which the system and the environment alternate moves". The crucial modality is the path quantifier denoted $\langle\langle A\rangle\rangle$, where $A$ is a set of agents, which ranges over all computations that the agents in $A$ can force the game into, irrespective of how the agents not in $A$ proceed. The existential path quantifier from CTL corresponds to $\langle\langle AGT\rangle\rangle$, where $AGT$ is the set of all agents, while the universal path quantifier from CTL corresponds to $\langle\langle\emptyset\rangle\rangle$. Ågotnes describes ATL as "a propositional logic in which statements about what coalitions can achieve by strategic cooperation can be expressed" [126]. One can show that ATL embeds Pauly's Coalition Logic [48]. (The operator $[C]\varphi$ from coalition logic corresponds to $\langle\langle C\rangle\rangle X\varphi$ in ATL, where $X$ is the nexttime operator.)

Some validities in ATL, based on *playable* effectivity functions, are ([48]):

- $\neg\langle\langle C\rangle\rangle X\mathtt{ff}$
- $\langle\langle C\rangle\rangle X\mathtt{tt}$
- $\neg\langle\langle\emptyset\rangle\rangle\neg\varphi \rightarrow \langle\langle AGT\rangle\rangle X\varphi$
- $\langle\langle C_1\rangle\rangle X\varphi \wedge \langle\langle C_2\rangle\rangle X\psi \rightarrow \langle\langle C_1 \cup C_2\rangle\rangle X(\varphi \wedge \psi)$ for disjoint $C_1$ and $C_2$
- $\langle\langle C\rangle\rangle\Box\varphi \leftrightarrow \varphi \wedge \langle\langle C\rangle\rangle X\langle\langle C\rangle\rangle\Box\varphi$ where $\Box$ stands for the always operator
- $\models \varphi \rightarrow \psi \ \Rightarrow\ \models \langle\langle C\rangle\rangle X\varphi \rightarrow \langle\langle C\rangle\rangle X\psi$
- $\models \varphi \ \Rightarrow\ \models \langle\langle C\rangle\rangle\Box\varphi$

Perhaps the reader can appreciate that the operator $\langle\langle C\rangle\rangle$ is a complicated one: it is not just a box or diamond operator in a modal logic. In fact, in [22] Broersen et al. show that within the STIT (seeing to it that [25, 66]) framework one can decompose the $\langle\langle C\rangle\rangle$ operator from ATL into a number of more primitive operators within STIT. Without going into details, we think this is an important result in order to understand ATL and its pivotal operator better. Moreover, Broersen [17] proposes CTL.STIT, a join of CTL and a variant of STIT logic, that subsumes ATL and adds expressivity in order to extend ATL to be able to express important multi-agent system verification properties (having to do with reasoning about norms, strategic games, knowledge games and the like).

Finally we mention that there are also epistemic variants of ATL, called ATEL [61, 126], i.e. ATL augmented with knowledge operators, which enables expressing properties such as "group $A$ can cooperate to bring about $\varphi$ iff it is common knowledge in $A$ that $\psi$". More generally, one can express properties about resource-bounded reasoners, cryptography and security and agent communication [61]. An example of the last category is $K_a\varphi \rightarrow \langle\langle a\rangle\rangle \diamond K_aK_b\varphi$, expressing

“freedom of speech": an agent can always tell the truth to other agents (and know this!). (Here $\diamond$ stands for the dual of $\Box$, as usual.) That ATEL is, besides an expressive logic, also a complicated one, was proven by some controversy about the proper set-up and interpretation of the logic. It appears that, if one is not careful, agents may not have incomplete information when choosing actions, which runs counter to the very conception of epistemic logic. A further elaboration of this issue goes beyond the scope of this chapter (cf. [63, 64]).

# 5 Related work

In this section I briefly indicate logic-related work in the field of intelligent agents and multi-agent systems that is not (only) based on logic proper, but also employs other techniques, such as programming techniques and model-checking algorithms.

## 5.1 Model-checking MAS

One of the most important applications of employing logic to describe multi-agent systems is that one can try to verify MAS programs. Although there is also some work on using theorem proving techniques for this (e.g., [2]), the most popular is employing model-checking techniques [27]. For instance, MOCHA [4] supports Alternating-Time Temporal Logic. But also model checkers exist that are even more targeted at model-checking MAS. For example, the model checker MCMAS supports specifications in a wide range of agent-relevant logics: the branching-time temporal logic CTL, epistemic logic (including operators of common and distributed knowledge), Alternating-Time Temporal Logic, and deontic modalities [77, 79]. Interestingly, this checker works with *interpreted systems* [44] interpretations of all the modalities involved: “ATL and CTL modalities are interpreted on the induced temporal relation given by the protocols and transition functions, the epistemic modalities are defined on the equivalence relations built on the equality of local states, and the deontic modalities are interpreted on “green states", i.e., subsets of local states representing states of locally correct behavior for the agent in question" [79].

## 5.2 BDI programming languages

After the formalization of (BDI-like) agent theories, researchers started to think of how to realize BDI -based agents, and came up with several architectures [119, 121]. To obtain a more systematic way of programming these kinds of agents, Yoav Shoham in 1993 introduced even a new programming paradigm, viz. *Agent-Oriented Programming* and a first agent-oriented programming language, viz. AGENT0. Later many other of these languages emerged. To mention a few: AgentSpeak [96], Jason [13], JADEX [95], 3APL [55], 2APL [31], GOAL [9, 54]. But there are many more nowadays, see [11, 10, 12]. Most of these are basically dedicated languages for programming agents in terms of BDI-like concepts such as beliefs, desires, goals, intentions, plans, ... Although the languages mostly also have imperative programming-style elements, there is logic(al reasoning) involved such as reasoning about beliefs, desires and goals represented in appropriate data bases, typically querying these bases to see whether they make certain conditions true to fire some rules that spawn plans (which amounts to the reactive planning concept). But normally the reasoning about BDI aspects is much less involved than full-blown BDI logic; normally there is no nesting of the operators, and the semantics is generally non-modal. But naturally researchers have tried to make connections between this rather ’flat’ BDI reasoning and full-blown BDI logics, typically using the latter as specification and verification logics for programs written in agent programming languages [60, 33, 57, 32, 56].

### 5.3 Situation Calculus and the GOLOG family of programming languages

The Situation Calculus is a ('first-order-like'[7] logical formalism to reason about action and change. It was devised by John McCarthy [83] in 1963. Since then it has gained enormous popularity, especially among AI researchers in North America. The area of representation of action and change has been dominated by the so-called frame problem. Although there are several aspects of this problem, one of the most important issues is that of how to specify succinctly what changes and what does *not* change while doing an action. Typically when one also wants to derive what doesn't change, so-called *frame axioms* are needed, and many of them, in the order of the number $n$ of fluents (changeable predicates) involved in the scenario at hand times the number $m$ of actions involved ([101], p. 26). Since this is deemed unmanageable in practice, this was considered a huge problem. Reiter [68, 100, 101] solved the problem by the use of so-called Basic Action Theories, consisting of successor state axioms (one per fluent) and precondition axioms (one per action), reducing the number of axioms to the order of $n+m$. Of course, this demands that the scenario is exactly known: there are no hidden fluents or actions in the scenario. (So the general frame problem remains unsolved, and is inherently unsolvable. It pertains to the question of what fluents and what actions to represent in a model of the scenario.) Based on the Situation Calculus, Reiter and Levesque [69] proposed an (imperative 'ALGOL'-like) programming language GOLOG, which is used to overcome the problem with the high complexity of first principle planning, in which for every goal a plan needs to be made. This is so complex since in general there are many different possible combinations of actions of which a combination must be found that leads to the goal. Typically Reiter et al. use a regression technique for this: starting out from the goal and looking backward to see what situation may lead to the goal in subsequently 0, 1, 2 , ..., $k$ steps (viz. actions). GOLOG eases the problem by providing a user-defined program to give some structure to the search (this is sometimes called *sketchy planning*). Since GOLOG's semantics is given in a logical form (using the situation calculus), a GOLOG program together with a basic action theory can be sent to a theorem prover that obtains a (constructive) proof of the question how the goal can be reached. More precisely, it determines what situation, expressed as a sequence of actions, i.e., a plan, performed in the initial state, satisfies the goal. Since the original version of GOLOG there has been proposed a number of refinements of the language, each dealing with particular features such as CONGOLOG [34] and IndiGOLOG [35].

## 6 Conclusion

In this chapter we have reviewed some of the history of logic as applied to an important area of artificial intelligence, viz. the area of intelligent agents and multi-agent systems. We have seen logics to describe the attitudes of single agents, notably BDI logics for describing the beliefs, desires and intentions of agents. As to multi-agent systems we have considered logics of norms and normative systems, such as deontic logic as well as hybrids of BDI-logics and deontic logic. We also have seen logics to describe counts-as relations in institutions. We have looked at logics for strategic reasoning such as Coalition Logic and AT(E)L. Finally we have briefly sketched some other developments related to the logics for agent-based systems.

[7] In fact the situation calculus is a first-order language with as special sorts: situations, actions and fluents[70]. An unorthodox aspect of it is that *situations*, as well as actions and fluents, are 'reified' in the languages, i.e. represented as first-order objects which are elements that are semantical in nature and typically used in the metalevel description of models in traditional first-order logic. This has as a consequence that they can be quantified over in the language, giving situation calculus a second-order feel [101]

**Acknowledgement**. This chapter benefitted substantially from the remarks and suggestions made by the reviewer.

# References

[1] H.M. Aldewereld, W. van der Hoek & J.-J.Ch. Meyer, Rational Teams: Logical Aspects of Multi-Agent Systems. *Fundamenta Informaticae* 63 (2-3), 2004, pp. 159 Ð 183.

[2] N. Alechina, M. Dastani, F. Kahn, B. Logan & J.-J. Ch. Meyer, Using Theorem Proving to Verify Properties of Agent Programs, in: Specification and Verification of Multi-Agent Systems (M. Dastani, K.V. Hindriks & J.-J. Ch. Meyer, eds.), Springer, New York/Dordrecht/Heidelberg/London, 2010, pp. 1–33.

[3] R. Alur, T.A. Henzinger & O. Kupferman, Alternating-Time Temporal Logic, *J. ACM* 49(5), 2002, pp. 672-713.

[4] R. Alur, T. Henzinger, F. Mang, S. Qadeer, S. Rajamani & S. Tasiran. MOCHA: Modularity in model checking, in: *Proc. 10th International Conference on Computer Aided Verification (CAVÕ98)*, LNCS 1427. Springer, Berlin, 1998, pp. 521Ð525 .

[5] A.R. Anderson, A Reduction of Deontic Logic to Alethic Modal Logic, *Mind* 67, 1958, pp. 100–103.

[6] P. Bartha, Conditional Obligation, Deontic Paradoxes, and the Logic of Agency, *Annals of Mathematics and Artificial Intelligence* 9(1–2), 1993, pp. 1–23.

[7] J. van Benthem, Correspondence Theory, in: *Handbook of Philosophical Logic, 2nd edition* (D. Gabbay &F. Guenthner, eds.), number 3, Kluwer, 2001, pp. 325–408

[8] P. Blackburn, J.F.A.K. van Benthem, & F. Wolter (eds.), *Handbook of Modal Logic*, Elsevier, Amsterdam, 2007.

[9] F.S. de Boer, K.V. Hindriks, W. van der Hoek & J.-J. Ch. Meyer, Agent Programming with Declarative Goals, *J. of Applied Logic* 5, 2007, pp. 277-302.

[10] R.H. Bordini, L. Braubach, M.M. Dastani, A.E.F. Seghrouchni, J.J. Gomez-Sanz, J. Leite, G. O'Hare, A. Pokahr & A. Ricci, (2006). A Survey of Programming Languages and Platforms for Multi-Agent Systems, *Informatica* 30, 2008, pp. 33-44.

[11] R.H. Bordini, M.M. Dastani, J. Dix & A.E.F. Seghrouchni, eds., *Multi-Agent Programming: Languages, Platforms and Applications*, Springer, New York, 2005.

[12] R.H. Bordini, M.M. Dastani, J. Dix & A.E.F. Seghrouchni, eds., *Multi-Agent Programming (Languages, Tools and Applications*, Springer, Dordrecht / Heidelberg, 2009

[13] R.H. Bordini, J.F. Hubner, & M. Wooldridge. *Programming Multi-Agent Systems in AgentSpeak Using Jason*. John Wiley & Sons, Chichester, UK, 2007.

[14] M.E. Bratman, *Intentions, Plans, and Practical Reason*, Harvard University Press, Massachusetts, 1987.

[15] M.E. Bratman. What is intention? In P. R. Cohen, J. Morgan & M. E. Pollack (eds.), *Intentions in Communication*, chapter 2, pages 15Ð31. MIT Press, Cambridge, MA, 1990.

[16] M.E. Bratman, D.J. Israel, M.E. Pollack: Plans and resource-bounded practical reasoning. *Computational Intelligence* 4, 1988, pp. 349-355.

[17] J. Broersen, CTL.STIT: Enhancing ATL to Express Important Multi-Agent Systems Verification Properties, in: in: *Proc. of the 9th Int. Conf. on Autonomous Agents and Multi-agent Systems (AAMAS 2010)* (van der Hoek, Kaminka, Lespérance, Luck & Sen, eds.), Toronto, Canada, 2010, pp. 683–690.

[18] J.M. Broersen, Deontic epistemic stit logic distinguishing modes of mens rea, *Journal of Applied Logic* 9(2), 2011, pp. 127–152.

[19] J.M. Broersen, Modelling Attempt and Action Failure in Probabilistic stit Logic, in: *Proc. 22nd Int. J. Conf. on Artif. Intell. (IJCAI 2011)*, 2011, pp. 792–797.

[20] J.M. Broersen, M.M. Dastani, J. Hulstijn & L. van der T|orre, Goal Generation in the BOID Architecture, *Cognitive Science Quarterly Journal* 2(3–4), 2002, pp. 428–447.

[21] J. Broersen, F. Dignum, V. Dignum & J.-J. Ch. Meyer, Designing a Deontic Logic of Deadlines, in: *Proc. Deontic Logic in Computer Science (DEON 2004)* (A. Lomuscio & D. Nute, eds.), LNAI 3065, Springer, Berlin, 2004, pp. 43–56.

[22] J. Broersen, A. Herzig & N. Troquard, Embedding Alternating-Time Temporal Logic in Strategic STIT Logic of Agency, *J. of Logic and Computation* 16(5), 2006, pp. 559–578

[23] J. Carmo & A.J.I. Jones, Deontic Database Constraints, Violation and Recovery, *Studia Logica* 57(1), 1996, pp. 139-165.

[24] B.F. Chellas, *Modal Logic, An Introduction*, Cambridge University Press, Cambridge, UK, 1980.

[25] B.F. Chellas, On bringing it about, *Journal of Philosophical Logic* 24, 1995, pp. 563–571.

[26] E.M. Clarke & E.A. Emerson, Design and Synthesis of Synchronization Skeletons Using Branching-Time Temporal Logic, in: *Proc. Workshop on Logic of Programs*, LNCS 131, Springer, Berlin, 1981, pp. 52–71.

[27] E.M. Clarke, O. Grumberg & D.A. Peled, Model Checking, The MIT Press, Cambridge, Massachusetts, 1999.

[28] P.R. Cohen & H.J. Levesque, Intention is Choice with Commitment, *Artificial Intelligence* 42(3), 1990, pp. 213–261.

[29] P. Cohen & H. Levesque, Teamwork, *Nous* 24(4), 1991, pp. 487–512.

[30] R. Craven & M.J. Sergot, Agent Strands in the Action Language nC+, *J. of Applied Logic* 6, 2008, pp. 172-191.

[31] M.M. Dastani, M.M., 2APL: A Practical Agent Programming Language. International Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS), 16(3), 2008, pp. 214-248.

[32] M. Dastani, K.V. Hindriks & J.-J. Ch. Meyer (eds.), *Specification and Verification of Multi-Agent Systems*, Springer, New York / Dordrecht / Heidelberg / London, 2010.

[33] M. Dastani, B. van Riemsdijk & J.-J. Ch. Meyer, A Grounded Specification Language for Agent Programs, in: *Proc. 6th Int. J. Conf. On Autonomous Agents and Multi-Agent Systems (AAMASÕ07)* (M. Huhns, O. Shehory, E.H. Durfee & M. Yokoo, eds.), Honolulu, Hawai'i, USA, 2007, pp. 578-585.

[34] G. De Giacomo, Y. Lespérance, and H. J. Levesque, ConGolog, a concurrent programming language based on the situation calculus, *Artificial Intelligence Journal* 121(1Ð2):109Ð169, 2000.

[35] G. De Giacomo, Y. Lespérance, H. J. Levesque, and S. Sardina, IndiGolog: A high-level programming language for embedded reasoning agents. In: R. H. Bordini, M. Dastani, J. Dix & A. E. Fallah-Seghrouchni, eds, Multi-Agent Programming: Languages, Platforms and Applications, chapter 2, Springer, 2009, pp. 31Ð72.

[36] M.V. Dignum & F.P.M. Dignum, A logic of agent organizations. *Logic Journal of the IGPL* 20, 2012, pp. 283–316.

[37] F.P.M. Dignum, D. Kinny & E. Sonenberg (2001). Motivational Attitudes of Agents: On Desires Obligations and Norms, in: *Proc. 2nd int. workshop of central and eastern europe on multi-agent systems (CEEMAS'01)* (B. Dunin-Keplicz & E. Nawarecki, eds.), Krakow, Poland, 2001, pp. 61-70.

[38] F.P.M. Dignum & B. van Linder, Modeling Social Agents: Towards deliberate communication, in: *Agent-Based Defeasible Control in Dynamic Environments, Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 7*, (J.-J.Ch. Meyer & J. Treur, eds.), Kluwer, Dordrecht, 2002, pp. 357-380.

[39] F. Dignum, J.-J. Ch. Meyer & R.J. Wieringa, A Dynamic Logic for Reasoning about Sub-Ideal States, in: *Proc. ECAI'94 Workshop "Artificial Normative Reasoning"* (J. Breuker, ed.), Amsterdam, 1994, pp. 79-92.

[40] F. Dignum, J.-J. Ch. Meyer, R.J. Wieringa & R. Kuiper, A Modal Approach to Intentions, Commitments and Obligations: Intention plus Commitment Yields Obligation, in: *Deontic Logic, Agency and Normative Systems (Proc. DEON'96)* (M.A. Brown & J. Carmo, eds.), Workshops in Computing, Springer, Berlin, 1996, pp. 80–97.

[41] B. Dunin-Kęplicz & R. Verbrugge, Collective Intentions, *Fundamenta Informaticae* 51(3) (2002), pp. 271–295.

[42] J.A, van Eck, A System of Temporally Relative Modal and Deontic Predicate Logic and Its Philosophical Applications, *Logique et Analyse* 100, 1982, pp. 249–381.

[43] E.A. Emerson. Temporal and modal logic, in J. van Leeuwen, ed., *Handbook of Theoretical Computer Science, volume B: Formal Models and Semantics*, chapter 14, Elsevier Science, Amsterdam, 1990, pp. 996–1072.

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

[45] J. Fiadeiro & T. Maibaum, Temporal Reasoning over Deontic Specifications, *J. of Logic and Computation* 1(3), 1991, pp. 357–395.

[46] J.W. Forrester, Gentle Murder, or the Adverbial Samaritan, *Journal of Philosophy* 81, pp. 193–196.

[47] E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain & H. Turner, Nonmonotonic Causal Theories, *Artificial Intelligence* 153, 2004, pp. 49-104.

[48] V. Goranko, Coalition Games and Alternating Temporal Logics, in: *Proc. of the 8th Conference on Theoretical Aspects of Rationality and Knowledge (TARK VIII)* (J. van Benthem, ed.), Siena, Italy, Morgan Kaufmann, 2001, pp. 259-272.

[49] J. Gray, Notes on Database Operating Systems, Operating Systems, 1978, pp. 393-481.

[50] D. Grossi, Designing Invisible Handcuffs (Formal Investigations in Institutions and Organizations for Multi-Agent Systems), PhD Thesis, Utrecht University, Utrecht, 2007.

[51] D. Grossi, J.-J. Ch. Meyer & F. Dignum. Classificatory Aspects of Counts-as: An Analysis in Modal Logic, *Journal of Logic and Computation*16(5), 2006, pp. 613Ð643.

[52] D. Harel, Dynamic Logic, in: D. Gabbay & F. Guenthner (eds.), *Handbook of Philosophical Logic, Vol. II*, Reidel, Dordrecht/Boston, 1984, pp. 497–604.

[53] D. Harel, D. Kozen & J. Tiuryn, *Dynamic Logic*, MIT Press, Cambridge, Massachusetts, 2000.

[54] K.V. Hindriks: Modules as Policy-Based Intentions: Modular Agent Programming in GOAL, *Proc. PROMAS 2007* (M. Dastani, A. El Fallah Seghrouchni, A. Ricci & M. Winikoff, eds.), LNAI 4908, Springer, Berlin / Heidelberg, 2008, pp. 156-171.

[55] K.V. Hindriks, F.S. de Boer, W. van der Hoek & J.-J. Ch. Meyer, Agent Programming in 3APL, *Int. J. of Autonomous Agents and Multi-Agent Systems* 2(4), 1999, pp.357–401.

[56] K.V. Hindriks, W. van der Hoek & J.-J. Ch. Meyer, GOAL Agents Instantiate Intention Logic, in: *Logic Programs, Norms and Action (Sergot Festschrift)* (A. Artikis, R. Craven, N.K. Çiçekli, B. Sadighi & K. Stathis, eds), LNAI 7360, Springer, Heidelberg, 2012, pp. 196-219.

[57] K.V. Hindriks & J.-J. Ch. Meyer, Agent Logics as Program Logics: Grounding KARO, in: *Proc. 29th Annual German Conference on AI, KI 2006* (C. Freksa, M. Kohlhase, & K. Schill, eds.)), Bremen, Germany, June 14-17, 2006, Proceedings, LNAI 4314, Springer, 2007, pp. 404–418

[58] W. van der Hoek, B. van Linder & J.-J. Ch. Meyer, An Integrated Modal Approach to Rational Agents, in: M. Wooldridge & A. Rao (eds.), *Foundations of Rational Agency*, Applied Logic Series 14, Kluwer, Dordrecht, 1998, pp. 133–168.

[59] W. van der Hoek, B. van Linder & J.-J. Ch. Meyer, On Agents That Have the Ability to Choose, *Studia Logica* 65, 2000, pp. 79–119.

[60] W. van der Hoek & M. Wooldridge, Towards a Logic of Rational Agency, *Logic J. of the IGPL* 11(2), 2003, pp. 133–157.

[61] W. van der Hoek & M. Wooldridge, Cooperation, Knowledge, and Time: Alternating-Time Temporal Epistemic Logic and Its Applications, *Studia Logica* 75, 2003, pp. 125–157.

[62] J.F. Horty, *Agency and Deontic Logic*, Oxford University Press, 2001.

[63] W. Jamroga, Some Remarks on Alternating Temporal Epistemic Logic, in: *Proc. FAMAS 03 - Formal Approaches to Multi-Agnet Systems*, Warsaw, Poland, 2003, pp. 133–140.

[64] W. Jamroga & W. van der Hoek, Agents That Know How to Play, *Fundamenta Informaticae* 63, 2004, pp. 185–219.

[65] A. Jones & M. Sergot, A Formal Characterzation of Institutionalised Power, *J. of the IGPL* 3, 1996, pp. 427–443.

[66] M. Kracht, J.-J. Ch. Meyer & K. Segerberg, The Logic of Action, in: *The Stanford Encyclopedia of Philosophy (2009 Edition)*, Edward N. Zalta (ed.), URL=http://plato.stanford.edu/entries/logic-action/, first published March 31, 2009, 29 p.

[67] S. Kraus & D. Lehmann, Knowledge, Belief and Time, in: L. Kott (ed.), *Proceedings of the 13th Int. Colloquium on Automata, Languages and Programming*, Rennes, LNCS 226, Springer, Berlin, 1986.

[68] H. Levesque, F. Pirri & R. Reiter, Foundations for the Situation Calculus, Linköping Electronic Articles in Computer and Information Science 3(18), Linköping University Electronic Press, Linköping, 1998.

[69] H.J. Levesque, R. Reiter, Y. Lespérance, F. Lin & R.B. Scherl, GOLOG: A Logic Programming Language for Dynamic Domains, *J. of Logic Programming* 31, 1997, pp. 59–84.

[70] F. Lin, Situation Calculus, Chapter 16 of: Handbook of Knowledge Representation (F. van Harmelen, V. Lifschitz & B. Porter, eds.), Elsevier, Amsterdam / London / New York, 2008, pp. 649–.669.

[71] B. van Linder, Modal Logics for Rational agents, PhD. Thesis, Utrecht University, 1996.

[72] B. van Linder, W. van der Hoek & J.-J. Ch. Meyer, Actions that Make You Change Your Mind: Belief Revision in an Agent-Oriented Setting, in: *Knowledge and Belief in Philosophy and Artificial Intelligence* (A. Laux & H. Wansing, eds.), Akademie Verlag, Berlin, 1995, pp. 103–146.

[73] B. van Linder, W. van der Hoek & J.-J. Ch. Meyer, Seeing is Believing (And So Are Hearing and Jumping), *Journal of Logic, Language and Information* 6, 1997, pp. 33–61.

[74] G.J.C. Lokhorst. Mally's deontic logic. In Edward N. Zalta, ed., *Stanford Encyclopedia of Philosophy*, The Metaphysics Research Lab at the Center for the Study of Language and Information, Stanford University, Stanford, CA, 2002.

[75] G.J.C. Lokhorst. Where did Mally go wrong? In: *Proc. DEON 2010* (G. Governatori & G. Sartor, eds.), Lecture Notes in Artificial Intelligence (LNAI) 6181, Springer-Verlag, Berlin / Heidelberg, 2010, pp. 247-258.

[76] G.J.C. Lokhorst. An intuitionistic reformulation of Mally's deontic logic, *Journal of Philosophical Logic*, published online, 2012.

[77] A. Lomuscio & F. Raimondi, MCMAS: A model checker for multi-agent systems, in: *Proc. TACAS 2006*, LNCS 3920, Springer, Berlin, 2006, pp. 450Ð454

[78] A. Lomuscio & M. Sergot, Deontic Interpreted Systems, *Studia Logica* 75(1), 2003, pp. 63–92.

[79] A. Lomuscio, H. Qu & F. Raimondi. MCMAS: A model checker for the verification of multi-agent systems, in: *Proc. CAV09*, LNCS 5643, Springer, Berlin, 2009, pp. 682Ð688.

[80] E. Lorini & A. Herzig, A logic of intention and attempt, *Synthese KRA* 163(1), 2008, pp. 45–77.

[81] D. Makinson & L. van der Torre, 2000. Input/output logics, *J. Philosophical Logic* 29, 2000, pp. 383–408.

[82] E. Mally, *Grundgesetze des Sollens, Elemente der Logik des Willens*, Leuschner & Lubensky, Graz, 1926.

[83] J. McCarthy. Situations, actions and causal laws. Technical Report, Stanford University,1963; later published in: M. Minsky, ed., Semantic Information Processing, MIT Press, Cambridge, MA, 1968, pp. 410Ð417. .

[84] P. McNamara, Deontic Logic, in: *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2010/entries/logic-deontic/>.

[85] J.-J. Ch. Meyer, A Different Approach to Deontic Logic: Deontic Logic Viewed As a Variant of Dynamic Logic, *Notre Dame J. of Formal Logic* 29 (1), 1988, pp. 109-136.

[86] J.-J. Ch. Meyer, J. Broersen & A. Herzig, BDI Logics, in: *Handbook of Epistemic Logic* (H. van Ditmarsch, J. Halpern, W. van der Hoek & B. Kooi eds.), College Publications, 2014, to appear.

[87] J.-J. Ch. Meyer & W. van der Hoek, *Epistemic Logic for AI and Computer Science*, Cambridge Tracts in Theoretical Computer Science 41, Cambridge University Press, 1995.

[88] J.-J. Ch. Meyer, W. van der Hoek & B. van Linder, A Logical Approach to the Dynamics of Commitments, *Artificial Intelligence* 113, 1999, 1–40.

[89] J.-J. Ch. Meyer & F. Veltman, Intelligent Agents and Common Sense Reasoning, Chapter 18 of: P. Blackburn, J.F.A.K. van Benthem, & F. Wolter (eds.), *Handbook of Modal Logic*, Elsevier, Amsterdam, 2007, pp. 991–1029.

[90] J.-J. Ch. Meyer, R.J. Wieringa & F.P.M. Dignum, The Role of Deontic Logic in the Specification of Information Systems, in: *Logics for Databases and Information Systems* (J. Chomicki & G. Saake, eds.), Kluwer, Boston/Dordrecht, 1998, pp. 71-115.

[91] R. Montague, Universal Grammar, *Theoria* 36, 1970, pp. 373–398.

[92] Donald Nute (ed.), *Defeasible Deontic Logic: Essays in Nonmonotonic Normative Reasoning*. Kluwer Academic Publishers, Dordrecht, Holland, 1997.

[93] M. Pauly, Logic for Social Software, ILLC Dissertations Series, Amsterdam, 2001.

[94] A. Pnueli, The Temporal Logic of Programs, in: *Proc. 18th Symp. on Foundations of Computer Science*, IEEE Computer Society Press, 1977, pp. 46–57.

[95] A. Pokahr, L. Braubach & W. Lamersdorf, JADEX: Implementing a BDI-infrastructure for JADE agents, *EXP - in search of innovation* (Special Issue on JADE), 3(3), 2003, pp. 76Ð85.

[96] A.S. Rao, AgentSpeak(L): BDI Agents Speak Out in a Logical Computable Language, in: *Agents Breaking Away* (W. van der Velde & J. Perram, eds.), LNAI 1038, Springer, Berlin, 1996, pp. 42–55.

[97] A.S. Rao & M.P. Georgeff, Modeling rational agents within a BDI-architecture, in *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)* (J. Allen, R. Fikes & E. Sandewall, eds.), Morgan Kaufmann, 1991, pp. 473–484.

[98] A.S. Rao & M.P. Georgeff, Decision Procedures for BDI Logics, *J. of Logic and Computation* 8(3), 1998, pp. 293–344.

[99] R. Reiter, A Logic for Default Reasoning, *Artificial Intelligence*13, 1980, pp. 81–132.

[100] R. Reiter. The frame problem in the situation calculus: a simple solution (sometimes) and a completeness result for goal regression. In: V. Lifschitz, ed., {em Artificial Intelligence and Mathematical Theory of Computation. Papers in Honor of John McCarthy, Academic Press, San Diego, CA, 1991, pp. 418Ð420.

[101] R. Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, The MIT Press, Cambridge, Massachusetts, 2001.

[102] S. Russell & P. Norvig, *Artificial Intelligence: A Modern Approach (3rd edition)*, Prentice Hall, Upper Saddle River /Boston, 2009.

[103] W. van der Hoek, J.-J. Ch. Meyer & J.W. van Schagen, Formalizing Potential of Agents: The KARO Framework Revisited, in: *Formalizing the Dynamics of Information* (M. Faller, S. Kaufmann & M. Pauly, eds.), CSLI Publications, (CSLI Lect. Notes 91), Stanford, 2000, pp. 51-67.

[104] D. Scott, Advice in modal logic, in: *Philosophical Problems in Logic* (K. Lambert, ed.). Reidel, Dordrecht, 1970.

[105] J. Searle, *Speech Acts, An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge, UK, 1969.

[106] J. Searle, *The Construction of Social Reality*, Free Press, 1995.

[107] Y. Shoham, Agent-Oriented Programming, *Artificial Intelligence* 60(1), 1993, pp. 51–92.

[108] M.P. Singh, Group Intentions, in: *Proc. 10th Int. Workshop on Distributed Artificial Intelligence (IWDAI-90)*, 1990.

[109] M.P. Singh, Group Ability and Structure, in: *Proc. of 2nd European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW-90)* (Y. Demazeau & J.-P. M uller, eds.), Elsevier, Amsterdam, 1991, pp. 127–146.

[110] M.P. Singh, Towards a Formal Theory of Communication for Multi-Agent Systems, in: *Proc. 12th Int. J. Conf. on Artif. Intell. (IJCAI-91)*, Sydney, Australia, 1991, pp. 69–74.

[111] M.P. Singh, A Semantics for Speech Acts, *Annals of Mathematics and Artificial Intelligence* 8(I-II), 1993, pp. 47–71.

[112] M.P. Singh, *Multi-Agent Systems: A Theoretical Framework for Intentions, Know-How, and Communications*, Springer, Heidelberg, 1994.

[113] M.P. Singh, The Intentions of Teams: Team Structure, Endodeixis, and Exodeixis, in: *Proc. 13th Eur. Conf. on Artif. Intell. (ECAI'98)* (H. Prade, ed.), Wiley, Chichester, 1998, pp. 303–307.

[114] A.Tarski, A Lattice-theoretical Fixpoint Theorem and Its Applications, *Pacific Journal of Mathematics* 5:2, 1955, pp. 285Ð309.

[115] R.H. Thomason, Deontic Logic As Founded on Tense Logic, in: *New Studies ion Deontic Logic* (R. Hilpinen, ed.), Reidel, Dordrecht, 1981, pp. 165–176.

[116] A.M. Turing, Computing Machinery and Intelligence, *Mind* 59, 1950, pp. 433–460.

[117] L. van der Torre, Reasoning about Obligations: Defeasibility in Preference-Based Deontic Logic, PhD Thesis, Erasmus University Rotterdam, Rotterdam, 1997.

[118] http://en.wikipedia.org/wiki/Institution, February 18th, 2013

[119] M.J. Wooldridge, Intelligent Agents, in: *Multiagent Systems* (G. Weiss, ed.), The MIT Press, Cambridge, MA, 1999, pp. 27–77.

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

[121] M. Wooldridge, *An Introduction to MultiAgent Systems* (2nd edition), John Wiley & Sons, Chichester, UK, 2009.

[122] M.J. Wooldridge & N.R. Jennings (eds.), *Intelligent Agents*, Springer, Berlin, 1995.

[123] G.H. von Wright, Deontic Logic, *Mind* 60, 1951, pp. 1-15.

[124] G.H. von Wright, A New System of Deontic Logic, *Danish Yearbook of Philosophy* 1, 1964, pp. 173-182.

[125] G.H. von Wright, Problems and Prospects of Deontic Logic: A Survey, in: E. Agazzi (ed.), Modern Logic - A Survey: Historical, Philosophical and Mathematical Aspects of Modern Logic and Its Applications, Reidel, Dordrecht / Boston, 1980, pp. 399-423.

[126] T. Ågotnes, Action and Knowledge in Alternating-Time Temporal Logic, *Synthese / KRA* 149, 2006, pp. 375–407.

[127] L. Åqvist & J. Hoepelman, Some Theorems about a Tree System of Deontic Tense Logic, in: *New Studies ion Deontic Logic* (R. Hilpinen, ed.), Reidel, Dordrecht, 1981, pp. 187-221.