Reasoning with Artificial Mental States
An Algebraic Approach

Nourhan Ehab[1] and Haythem O. Ismail[2,1]

[1] German University in Cairo, Egypt
Department of Computer Science and Engineering
[2] Cairo University, Egypt
Department of Engineering Mathematics
{nourhan.ehab,haythem.ismail}@guc.edu.eg

Abstract. Modelling the human mind, with its astounding complexity, has always been a long-sought goal of AI research. One of the most successful approaches to attain this goal is to ascribe human-like mental states to artificial agents. A mental state is based on a set of mental attitudes such as beliefs, desires, intentions, promises, obligations...etc. While there are several accounts in the literature for endowing artificial agents with mental attitudes, such approaches predominantly focus on investigating each attitude separately or on studying the interaction of a handful of particular attitudes, notably beliefs, desires, and intentions. Since human epistemic and practical reasoning processes are typically more complex, involving a myriad of attitudes, accounting for the interaction among generic mental attitudes is called for. To this end, we present an algebraic framework for modeling the interaction among generic mental attitudes. The framework is used to provide formal semantics for a logical language which may be used by a logic-based agent to reason with arbitrary mental attitudes.

Keywords: Agents Architectures · Mental States · Algebraic Semantics

1 Introduction

A hallmark of human intelligence is the ability to reason with a wide diversity of mental attitudes including beliefs, intentions, desires, promises, obligations...etc. which constitute our collective mental state. We are confronted everyday with situations that require us to deliberate given our current mental state, and we usually do so with ease. To demonstrate the variety of mental attitudes we deal with, even in the simplest of situations, consider the following example.

Example 1. Ted promised his best friend Marshall to go on a hunting trip with him during the weekend. Since Ted is a man of his word, he feels obliged to intend his promises and indeed intends what he is obliged to intend. If Ted intends to go to the trip, he must rent a car. At the same time, Ted has been procrastinating working on a long overdue report for weeks. He believes that if he does not work on the report this weekend, his boss will be really mad and will fire him. Ted fears being fired as he really likes his job. He regrets that he did not work on the report the previous weeks which

makes him feel obliged to work on the report during this weekend. Much to Ted's relief, he believes that he can go to the trip and dedicate some time to work on the report there if the trip location has internet connectivity. However, Ted doubts that there is internet connectivity at the trip location which makes him fear that he will not be able to work on the report after all. Since Ted is paranoid, he believes what he fears. □

In this example, Ted is reasoning with a mental state comprised of his promises, beliefs, obligations, intentions, fears, regrets, and doubts. In the modern world we often talk about machines as if they exhibit human-like mental attitudes as the aforementioned. Our daily lives typically involve numerous references to machines knowing, believing, desiring, intending, liking or disliking, understanding, owing, having duties and rights, or deserving rewards and punishments [16]. For this reason, a commonly investigated approach to achieving general AI is to ascribe mental attitudes to artificial agents as first suggested by McCarthy [15] and Newell [19]. Supporters of this line of research argue that mental-level modelling of artificial agents offers several advantages on both the theoretical and practical levels [27].

From a theoretical perspective, the abstract nature of mental models proved to be very useful in analysing and comparing different agent architectures. An example of this is Levesque's Computers as Believers paradigm which offered a uniform basis for analysing general knowledge representation schemes [14]. On the other hand, from a practical perspective, mental models offer a convenient abstraction based on well-understood attitudes while hiding low-level implementation-specific details [3]. This is a very useful feature in developing coopeartive multiagent systems as abstract explicit representations of each of the agents' mental attitudes enable more coherent interactions between them [20]. Moreover, endowing artificial agents with mental attitudes can facilitate the design of autonomous planning agents. For such agents, explicit representations of beliefs, desires, and obligations, for example, can drive the agents to take actions compatible with their beliefs to achieve their desires while respecting their obligations. Another practical realization of ascribing mental attitudes to artificial agents in a computational setting is the Agent Oriented Programming (AOP) paradigm. In AOP, the different modules of a program are viewed as agents possessing mental attitudes such as beliefs, decisions, capabilities, and obligations [26].

Perhaps the most renowned approach to designing agents with mental attitudes is the BDI agent architecture [23] and its extensions to include obligations [4]. These approaches exclusively focus on investigating the interactions between beliefs, desires, obligations, and intentions. For this reason, the BDI architecture fails to provide a good mental model for the human epistemic and practical reasoning processes as they typically involve a myriad of other mental attitudes such as plans, goals, and fears (just to name a few). This makes the BDI architecture not suitable for modelling human-centered trust-worthy agents which recently attracted a lot of research interest [9]. Even if we restrict ourselves to modelling rational agents, the BDI architecture still, in our opinion, falls short. Archetypal rational behaviour, for instance, is to form intentions to avoid one's fears or to mitigate ones's doubts which can not be represented (in a straight forward way) within the BDI/BOID frameworks.

To address this shortcoming, we propose in this paper a general algebraic framework capable of representing a *first-person* perspective of artificial agents possessing any

set of mental attitudes while capturing their interactions. We follow [27] and define a *mental state* as a set of mental attitudes. To the best of our knowledge, there does not exist in the literature a framework that allows for the reasoning with an arbitrary set of attitudes like our framework does offering a more refined mental model for a human-centered logical agent. To this end, we present a logic we refer to as $Log_A\mathbf{M}$ ("Log" stands for logic, "A" for algebraic, and "$\mathbf{M}$" for mental states) with precise semantics where mental states can be represented and the reasoning with the different mental attitudes of the agent can be captured. In defining the semantics of $Log_A\mathbf{M}$, we depart from the mainstream modal approaches to representing mental attitudes and take the *algebraic* approach instead.

As a starting step, we assume in $Log_A\mathbf{M}$ that the mental state is comprised of *binary* mental That is, the mental state can include information about the agent's beliefs (for example) but will not include information about the degrees of such beliefs. Further, we define a *monotonic* consequence relation for each mental attitude in the mental state based on pure algebraic notions. We already developed an extension of $Log_A\mathbf{M}$ to accommodate non-monotonic reasoning with graded mental states. We will informally outline this extension in Section 5, but we reserve the formal presentation to a longer version of the paper. An interesting special case of this graded extension for practical reasoning with graded beliefs and motivations is presented in [6].

The rest of the paper is structured as follows. We start by motivating our choice to pursue the algebraic route in defining the semantics of $Log_A\mathbf{M}$ in Section 2. We review in Section 3 foundational concepts of Boolean algebra on which $Log_A\mathbf{M}$ is based. We also generalize the classical notion of filters in Boolean algebra into what we will refer to as *multifilters* providing a generalized algebraic treatment of reasoning with mental states. Next, in Section 4, we present the syntax and semantics of $Log_A\mathbf{M}$. In Section 5, we informally describe the graded non-monotonic extension of $Log_A\mathbf{M}$. Finally, in Section 6, we present some concluding remarks.

2 Why the Algebraic Approach?

Before we delve into the technical details of $Log_A\mathbf{M}$, it is perhaps apt to ponder the merits of choosing to pursue the algebraic route. $Log_A\mathbf{M}$ is the most recent addition to a growing family of algebraic logics [11,12,13,7]. As such, it is essential for a treatment of reasoning with mental states within the algebraic framework. Hence, independent motivations for the algebraic approach are also motivations for $Log_A\mathbf{M}$. We will briefly present the motivations for the algebraic approach in this section.

The algebraic approach is based on an ontological commitment to propositions as first-class individuals in the universe of discourse; this leads to a language with no sentences, but with some of the terms taken to denote propositions. What does this buy us? Take $Log_A\mathbf{B}$ [11] for example. As an algebraic language for reasoning about belief, it strikes a middle ground between two major approaches to doxastic logic: the dominant, modal approach [28, for example] and the (now relatively out of fashion) first-order syntactical approach [22, for instance]. This allows $Log_A\mathbf{B}$, on one hand, to avoid problems of logical omniscience, which mar the classical modal approach, while, on the the other hand, staying immune to paradoxes of self-reference plaguing

the syntactical approach. $Log_A\mathbf{G}$ [7] is an algebraic logic for non-monotonic reasoning about graded beliefs. It is demonstrably useful for modelling resource-bounded reasoning; simulating inconclusive reasoning with circular, liar-like sentences; and reasoning about information arriving over a chain of sources each with a different degree of trust. As proven in [7], $Log_A\mathbf{G}$ can capture a wide array of non-monotonic reasoning formalisms such as possibilistic logic, circumscription, default logic, autoepistemic logic, and the principle of negation as failure. As such, $Log_A\mathbf{G}$ can be considered an algebraic unifying framework for non-monotonicity. $Log_A\mathbf{C}_n$ [13], which is an algebraic logic for reasoning about preference, desire, and obligation, avoids the so-called paradoxes of deontic logic [18] by, again, abandoning classical possible-worlds semantics. In [12], the algebraic approach is adopted for the representation of temporal phenomena using the language $Log_A\mathbf{S}$. In classical first-order approaches to temporal logic [17,1, for example], tersely axiomatizing temporal properties often calls for the introduction of reified fluents into the ontology. In these approaches, reference to composite fluents (conjunctions thereof, for example) either is forbidden (as, for example, in the situation calculus [17]) or results in duplicating the logical connectives for statements and fluent-denoting terms (as, for example, in [1].) In $Log_A\mathbf{S}$, reference to composite fluents is straightforward, with a single set of proposition-based logical connectives. These different motivations for the algebraic approach suggest that it is only natural to consider a language like $Log_A\mathbf{M}$ if one is to model reasoning with mental states the algebraic way to gain its several indispensable advantages.

3 Boolean Algebras and Multifilters

In this section, we start by reviewing the algebraic concepts of Boolean algebras and filters underlying classical logic, then we extend the notion of filters to accommodate a logic of mental states where a mental state is a set of mental attitudes.

Definition 1. *A Boolean algebra is a sextuple $\mathfrak{A} = \langle \mathcal{P}, +, \cdot, -, \bot, \top \rangle$ where $\mathcal{P}$ is a non-empty set with $\{\bot, \top\} \subseteq \mathcal{P}$. $\mathfrak{A}$ is closed under the two binary operators $+$ and $\cdot$ and the unary operator $-$ observing the following properties [24].*

B1.1:	$a + b = b + a$	*(Commutativity)*
B1.2:	$a \cdot b = b \cdot a$	
B2.1:	$a + (b + c) = (a + b) + c$	*(Associativity)*
B2.2:	$a \cdot (b \cdot c) = (a \cdot b) \cdot c$	
B3.1:	$a + (a \cdot b) = a$	*(Absorption)*
B3.2:	$a \cdot (a + b) = a$	
B4.1:	$a \cdot (b + c) = (a \cdot b) + (a \cdot c)$	*(Distribution)*
B4.2:	$a + (b \cdot c) = (a + b) \cdot (a + c)$	
B5.1:	$a + -a = \top$	*(Complements)*
B5.2:	$a \cdot -a = \bot$	

For the purposes of this paper, we take the elements of $\mathcal{P}$ to be propositions and the operators $+$, $\cdot$, and $-$ to be disjunction, conjunction, and negation, respectively.

The following definition of *filters* is an essential notion of Boolean algebras to represent an algebraic counterpart to logical consequence [24]. Filters are defined in pure algebraic terms, without alluding to the notion of truth, by utilizing the natural lattice order $\leq$ on the the algebra: for $p_1, p_2 \in \mathcal{P}$, $p_1 \leq p_2 =_{def} p1 \cdot p2 = p1$. Henceforth, $\mathfrak{A}$ is a Boolean algebra $\langle \mathcal{P}, +, \cdot, -, \bot, \top \rangle$.

Definition 2. *A filter of $\mathfrak{A}$ is a subset F of $\mathcal{P}$ where*

1. $\top \in F$;
2. *If $a, b \in F$, then $a \cdot b \in F$; and*
3. *If $a \in F$ and $a \leq b$, then $b \in F$.*

The filter generated by $\mathcal{Q} \subseteq \mathcal{P}$ is the smallest filter $F(\mathcal{Q})$ of which $\mathcal{Q}$ is a subset.

The just presented definition of a filter is only suitable for modelling reasoning with a single set of propositions $\mathcal{Q}$. With the purpose of modelling the reasoning with mental states defined as a tuple of sets of propositions where each set represents a separate mental attitude of the agent, we extend the notion of filters to what we refer to as *multifilters*. For this reason, we extend classical filters that rely on the natural order $\leq$ on the Boolean algebra to what will refer to as *multifilters* that rely on an order on tuples of propositions where each proposition belongs to a mental attitude. (Recall that $\leq$ is the classical lattice order.)

Definition 3. *Let k be a positive integer. A k partial-order on $\mathfrak{A}$ is a partial order $\preceq_k$ on $\mathcal{P}^k$ such that, $(a_1, \ldots, a_k) \preceq_k (b_1, \ldots, b_k)$ and $b_i = \bot$, for some $1 \leq i \leq k$, only if $a_j = \bot$, for some $1 \leq j \leq k$. Further, we say that $\preceq_k$ is* classical in i *just in case, (i) if $(a_1, ..., a_k) \preceq_k (b_1, ..., b_k)$ then $a_i \leq b_i$ and (ii) if $a \leq b$ then $(\{\top\}^{i-1} \times \{a\} \times \{\top\}^{k-i}) \times (\mathcal{P}^{i-1} \times \{b\} \times \mathcal{P}^{k-i}) \subseteq \preceq_k$.*

In the sequel, we will drop the subscript k in $\preceq_k$ whenever there is no resulting ambiguity. We now define multifilters based on a k partial-order $\preceq$.

Definition 4. *Let $\preceq$ be a k partial order on $\mathfrak{A}$ and $S \subseteq \{1, ..., k\}$. A $\preceq$-multifilter of $\mathfrak{A}$ with respect to S is a tuple $\mathfrak{F}_{\preceq}(S) = \langle F_1, F_2, ..., F_k \rangle$ of subsets of $\mathcal{P}$ such that*

1. *$\top \in F_i$, for all i such that $1 \leq i \leq k$;*
2. *for all i, if $i \in S$, $a \in F_i$, and $b \in F_i$, then $a.b \in F_i$; and*
3. *if $(a_1, ..., a_k) \preceq (b_1, ..., b_k)$ and $(a_1, ..., a_k) \in \bigtimes_{i=1}^{k} F_i$, then $(b_1, ..., b_k) \in \bigtimes_{i=1}^{k} F_i$.*

We can observe at this point that the three conditions on multifilters are just generalizations of the three conditions on filters. The second condition though need not apply to all the sets $F_1, ..., F_k$. The index set S specifies the sets which are closed under the meet operation "$\cdot$" and hence observe the second condition. This is necessary as some mental attitudes need not be closed under "$\cdot$".

We next define how multifilters can be generated by a tuple of sets of propositions. The intuition is that each set of propositions represents a mental attitude and the tuple of sets represents the collective mental state. In this way, multifilters generalize filters to accomodate reasoning with multiple mental attitudes where some attitudes need not be closed under "$\cdot$".

Definition 5. *Let $\mathcal{Q}_1, ..., \mathcal{Q}_k \subseteq \mathcal{P}$, $\preceq$ be a k partial order on $\mathfrak{A}$, and $S \subseteq \{1, ..., k\}$. The $\preceq$-multifilter generated by $\langle \mathcal{Q}_1, ..., \mathcal{Q}_k \rangle$ with respect to S, denoted $\mathfrak{F}_{\preceq}(\langle \mathcal{Q}_1, ..., \mathcal{Q}_k \rangle, S)$, is a $\preceq$-multifilter $\langle \mathcal{Q}'_1, ..., \mathcal{Q}'_k \rangle$ with respect to S where $\mathcal{Q}'_i$ is the smallest set containing $\mathcal{Q}_i$, for all $1 \leq i \leq k$.*

Having defined multifilters, we are now ready to present an important result. The following theorem states that, under certain conditions, multifilters can be reduced to classical filters applied to the different sets of propositions representing the mental state.

Theorem 1. *Let $\mathcal{Q}_1, ..., \mathcal{Q}_k \subseteq \mathcal{P}$, $S \subseteq \{1, ..., k\}$, and $\preceq$ be a k partial order on $\mathfrak{A}$ which is classical in i for some $i \in S$. If $\mathfrak{F}_{\preceq}(\langle \mathcal{Q}_1, ..., \mathcal{Q}_k \rangle, S) = \langle \mathcal{Q}'_1, ..., \mathcal{Q}'_k \rangle$, then $\mathcal{Q}'_i = F(\mathcal{Q}_i)$.*

4 $Log_A\mathbf{M}$ Languages

In this section, we present the syntax and semantics of $Log_A\mathbf{M}$ in addition to defining a logical consequence relation for each mental attitude in the mental state. Utilizing the multifilters presented in Section 3, we show that our logical consequence relations have the distinctive properties of classical Tarskian logical consequence. The proofs of the theorems presented in this section are omitted for space limitations but can be found in [8].

4.1 $Log_A\mathbf{M}$ Syntax

$Log_A\mathbf{M}$ consists of terms constructed algebraically from function symbols. There are no sentences; instead, we use terms of a distinguished syntactic type to denote propositions. Propositions are included as first-class individuals in the $Log_A\mathbf{M}$ ontology and are structured in a Boolean algebra. Though non-standard, the inclusion of propositions in the ontology has been suggested by several authors [5,2,21,25].

A $Log_A\mathbf{M}$ language is a many-sorted language composed of a set of terms partitioned into two base sorts: σ_P is a set of terms denoting propositions and σ_I is a set of terms denoting anything else. A $Log_A\mathbf{M}$ *alphabet* Ω includes a non-empty, countable set of constant and function symbols each having a syntactic sort from the set $\sigma = \{\sigma_P, \sigma_I\} \cup \{\tau_1 \longrightarrow \tau_2 \mid \tau_1 \in \{\sigma_P, \sigma_I\} \text{ and } \tau_2 \in \sigma\}$ of syntactic sorts. Intuitively, $\tau_1 \longrightarrow \tau_2$ is the syntactic sort of function symbols that take a single argument of sort σ_P or σ_I and produce a functional term of sort τ_2. Given the restriction of the first argument of function symbols to base sorts, $Log_A\mathbf{M}$ is, in a sense, a first-order language.

An alphabet Ω includes a countably infinite set of variables of the two base sorts; a set of syncategorematic symbols including the comma, various matching pairs of brackets and parentheses, the symbol $\forall$, and a set of logical symbols defined as the union of the following sets:

- $\{\neg\} \subseteq \sigma_P \longrightarrow \sigma_P$.
- $\{\wedge, \vee\} \subseteq \sigma_P \longrightarrow \sigma_P \longrightarrow \sigma_P$
- $\{\mathbf{A}_i\}_{i=1}^{k} \subseteq \sigma_P \longrightarrow \sigma_P$.

The symbols $\neg, \wedge, \vee$ denote negation, conjunction, and disjunction respectively. $\mathbf{A}_i(t)$ denotes that the agent has the attitude $\mathbf{A}_i$ towards the propositional term t. Terms involving $\Rightarrow$ (material implication), $\Leftrightarrow$ (logical equivalence), and $\exists$ are abbreviations defined in the standard way. In the following, we define $Log_A\mathbf{M}$ languages.

Definition 6. *A $Log_A\mathbf{M}$ language $\mathcal{L}$ is the smallest set of terms formed according to the following rules, where t and t_i ($i \in \mathbb{N}$) are terms in $\mathcal{L}$.*

- *All variables and constants in the alphabet Ω are in $\mathcal{L}$.*
- *$f(t_1, \ldots, t_m) \in \mathcal{L}$, where $f \in \Omega$ is of type $\tau_1 \longrightarrow \ldots \longrightarrow \tau_m \longrightarrow \tau$ ($m > 0$) and t_i is of type τ_i.*
- *$\neg t \in \mathcal{L}$, where $t \in \sigma_P$.*
- *$(t_1 \otimes t_2) \in \mathcal{L}$, where $\otimes \in \{\wedge, \vee\}$ and $t_1, t_2 \in \sigma_P$.*
- *$\forall x(t) \in \mathcal{L}$, where x is a variable in Ω and $t \in \sigma_P$.*
- *$\mathbf{A}_i(t) \in \mathcal{L}$, where $t \in \sigma_P$.*

We are now ready to define $Log_A\mathbf{M}$ theories based on the previously defined $Log_A\mathbf{M}$ languages.

Definition 7. *A $Log_A\mathbf{M}$ theory $\mathbb{T}$ is a triple $\langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ where:*

- *$\mathbb{A} = (\mathbb{A}_1, ..., \mathbb{A}_k)$ is a k-tuple where $\mathbb{A}_1, ..., \mathbb{A}_k \subseteq \sigma_P$; and*
- *$\mathbb{R}$ is a set of bridge rules each of the form $A_1, ..., A_k \longmapsto A'_1, ..., A'_k$ where $A_1, ..., A_k, A'_1, ..., A'_k \subseteq \sigma_P$.*
- *$\mathbb{S} \subseteq \{1, ..., k\}$.*

The tuple $\mathbb{A}$ represents the mental state of the agent. Each set $\mathbb{A}_1, ..., \mathbb{A}_k$ represents a separate mental attitude of the agent. If a propositional term $t \in \mathbb{A}_i$, then the agent has the attitude $\mathbb{A}_i$ towards t. It is worth pointing out here the utility of the terms of the form $\mathbf{A}_i(t)$. The membership of $\mathbf{A}_i(t)$ in $\mathbb{A}_j$ for all $1 \leq i, j \leq k$ means that the agent has the attitude $\mathbb{A}_j$ towards $\mathbf{A}_i(t)$. For example, if $\mathbf{A}_3(\phi) \in \mathbb{A}_2$, $\mathbf{A}_3(\phi)$ represents that the agent intends ϕ and $\mathbb{A}_2$ represents the agent's beliefs, then this means that the agent believes that it intends ϕ. This is very useful in representing higher-order motivations first suggested in [10]. The bridge rules serve to "bridge" propositions across the different mental attitudes. A bridge rule $A_1, ..., A_k \longmapsto A'_1, ..., A'_k$ means that if A_i is a subset of the current i-th mental attitude, then A'_i should be added to the current i-th mental attitude. The bridge rules facilitate the representation of the interaction between the mental attitudes. The set $\mathbb{S}$ specifies the sets in $\mathbb{A}$ whose denotations are closed under the meet operation and, hence, observe the second condition in the definition of multifilters. This facilitates having some attitudes in the mental state that are not closed under meet/conjunction. An example of such attitudes are desires. If one desires to go to the beach and desires to work on the report, one might not desire to go to the beach and work on the report. We now go back to Example 1 showing a corresponding encoding of it as a $Log_A\mathbf{M}$ theory.

Example 2. Let "r" denote working on the report, "t" denote going to the trip, "m" denote the boss's getting mad, "f" denote Ted getting fired, "c" denote Ted renting a car, "l" denote Ted liking his job, and "i" denote internet connectivity at the trip location. A possible $Log_A\mathbf{M}$ theory representing Example 1 is
$\mathbb{T} = \langle (\mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3, \mathbb{A}_4, \mathbb{A}_5, \mathbb{A}_6), \mathbb{R}, \mathbb{S} \rangle$ where:

- $\mathbb{A}_1$ represents Ted's promises (P). $\mathbb{A}_1 = \{t\}$.
- $\mathbb{A}_2$ represents Ted's beliefs (B). $\mathbb{A}_2 = \{\neg r \Rightarrow m, m \Rightarrow f, l, i \Rightarrow r \wedge t\}$.
- $\mathbb{A}_3$ represents Ted's intentions (I). $\mathbb{A}_3 = \{\}$.
- $\mathbb{A}_4$ represents Ted's fears (F). $\mathbb{A}_4 = \{\}$.
- $\mathbb{A}_5$ represents Ted's regrets (R). $\mathbb{A}_5 = \{\neg r\}$.
- $\mathbb{A}_6$ represents Ted's doubts (D). $\mathbb{A}_6 = \{\neg i\}$.
- $\mathbb{A}_7$ represents Ted's obligations (O). $\mathbb{A}_7 = \{\}$.
- $\mathbb{R}$ is the set of instances of the following rule schema where ϕ is a variable. In what follows, we eliminate the empty sets in the bridge rules and add to each set the first letter of the attitude it is representing. For example, the rule $\{\phi\}$,{},{},{},{},{},{} $\longmapsto$ {}, {}, {},{},{},{},{},$\{\phi\}$ will be written as $P = \{\phi\} \longmapsto O = \{\phi\}$.

  r1. $P = \{\phi\} \longmapsto O = \{\mathbf{I}(\phi)\}$
  r2. $O = \{\mathbf{I}(\phi)\} \longmapsto I = \{\phi\}$
  r3. $I = \{t\} \longmapsto I = \{c\}$
  r4. $B = \{l\} \longmapsto F = \{f\}$
  r5. $R = \{\neg r\} \longmapsto O = \{r\}$
  r6. $D = \{\neg i\} \longmapsto F = \{\neg r\}$
  r7. $F = \{\phi\} \longmapsto B = \{\phi\}$

  In the above rules, we use $\mathbf{I}(\phi)$ as a mnemonic equivalent to $\mathbf{A}_3(\phi)$ to denote that Ted intends ϕ. To illustrate how the bridge rules can be read, **r1** represents that if Ted promised to ϕ, then he is obliged to intend ϕ and **r2** represents that if Ted is obliged to intend any ϕ, then he intends ϕ. The rest of the rules can be read in a similar way.
- $\mathbb{S} = \{2\}$. This means that only the set of beliefs is closed under the meet/conjunction operation. □

The representation of a first-person variant of a BDI agent as a $Log_A\mathbf{M}$ theory should now be straightfoward. The corresponding $Log_A\mathbf{M}$ theory will contain a mental state $\mathbb{A}$ composed of three sets of attitudes representing the agent's beliefs, desires, and intentions. The bridge rules can be used to represent the axioms of BDI logics governing the interactions between the three mental attitudes.

4.2 From Syntax to Semantics

In this section, we present semantics for the syntax of $Log_A\mathbf{M}$ in addition to defining an interpretation function. We start by presenting a key element in the semantics of $Log_A\mathbf{M}$ which is the notion of a $Log_A\mathbf{M}$ *structure*.

Definition 8. *A* $Log_A\mathbf{M}$ structure *is a triple* $\mathfrak{S}_k = \langle \mathcal{D}, \mathfrak{A}, \mathcal{A}_k \rangle$*, where*

- $\mathcal{D}$ *is the domain of discourse containing a distinguished non-empty countable set of propositions* $\mathcal{P}$.
- $\mathfrak{A} = \langle \mathcal{P}, +, \cdot, -, \bot, \top \rangle$ *is a complete, non-degenerate Boolean algebra.*
- $\mathcal{A}_k = \{\mathfrak{a}_i \mid 1 \leq i \leq k\}$ *where* $\mathfrak{a}_i : \mathcal{P} \longrightarrow \mathcal{P}$, $1 \leq i \leq k$, *is a function modeling an mental attitude.*

A *valuation* $\mathcal{V}$ of a $Log_A\mathbf{M}$ language is a triple $\langle \mathfrak{S}_k, \mathcal{V}_f, \mathcal{V}_x \rangle$, where $\mathfrak{S}_k$ is a $Log_A\mathbf{M}$ structure, $\mathcal{V}_f$ is a function that assigns to each function symbol an appropriate function on $\mathcal{D}$, and $\mathcal{V}_x$ is a function mapping each variable to a corresponding element of the appropriate block of $\mathcal{D}$. An *interpretation* of $Log_A\mathbf{M}$ terms is given by a function $[\![\cdot]\!]^{\mathcal{V}}$.

Definition 9. Let $\mathcal{L}$ be a $Log_A\mathbf{M}$ language and let $\mathcal{V}$ be a valuation of $\mathcal{L}$. An *interpretation* of the terms of $\mathcal{L}$ is given by a function $[\![\cdot]\!]^{\mathcal{V}}$:

- $[\![x]\!]^{\mathcal{V}} = \mathcal{V}_x(x)$, for a variable x
- $[\![c]\!]^{\mathcal{V}} = \mathcal{V}_f(c)$, for a constant c
- $[\![f(t_1, \ldots, t_n)]\!]^{\mathcal{V}} = \mathcal{V}_f(f)([\![t_1]\!]^{\mathcal{V}}, \ldots, [\![t_m]\!]^{\mathcal{V}})$, for an m-adic ($m \geq 1$) function symbol f
- $[\![(t_1 \wedge t_2)]\!]^{\mathcal{V}} = [\![t_1]\!]^{\mathcal{V}} \cdot [\![t_2]\!]^{\mathcal{V}}$
- $[\![(t_1 \vee t_2)]\!]^{\mathcal{V}} = [\![t_1]\!]^{\mathcal{V}} + [\![t_2]\!]^{\mathcal{V}}$
- $[\![\neg t]\!]^{\mathcal{V}} = -[\![t]\!]^{\mathcal{V}}$
- $[\![\forall x(t)]\!]^{\mathcal{V}} = \prod_{a \in \mathcal{D}} [\![t]\!]^{\mathcal{V}[a/x]}$
- $[\![\mathbf{A}_i(t_1)]\!]^{\mathcal{V}} = \mathfrak{a}_i([\![t_1]\!]^{\mathcal{V}})$

In the rest of the paper, for any $\Gamma \subseteq \sigma_p$, we will use $[\![\Gamma]\!]^{\mathcal{V}}$ to denote $\prod_{p \in \Gamma} [\![p]\!]^{\mathcal{V}}$ for notational convenience.

4.3 Logical Consequence

Having defined the syntax and semantics of $Log_A\mathbf{M}$. What remains for us is to define logical consequence. Since we are taking the algebraic route, we employ our notion of multifilters from Section 3 to define a consequence relation for each mental attitude in a $Log_A\mathbf{M}$ theory.

In Section 3, we defined multifilters based on an arbitrary partial order $\preceq$. We start by defining how to construct such an order for the tuples of propositions in $\mathcal{P}$. The intuition is that the order is induced by the bridge rules in a $Log_A\mathbf{M}$ theory in addition to the natural order $\leq$ among the attitudes that observe the second condition of the definition of multifilters (closure under the meet/conjunction operation).

Definition 10. *Let $\mathbb{T} = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be a $Log_A\mathbf{M}$ theory and $\mathcal{V}$ a valuation. A $\mathbb{T}^{\mathcal{V}}$-induced order, denoted $\preceq_{\mathbb{T}^{\mathcal{V}}}$, is a partial order over $\mathcal{P}^k$ with the following properties.*

1. *If $i \in \mathbb{S}$ and $a \leq b$, then $(\{\top\}^{i-1} \times \{a\} \times \{\top\}^{k-i}) \preceq_{\mathbb{T}^{\mathcal{V}}} (\{\top\}^{i-1} \times \{b\} \times \{\top\}^{k-i})$.*
2. *If $(A_1, \ldots, A_k \longmapsto A'_1, \ldots, A'_k) \in \mathbb{R}$, then $([\![A_1]\!]^{\mathcal{V}}, \ldots, [\![A_k]\!]^{\mathcal{V}}) \preceq_{\mathbb{T}^{\mathcal{V}}} ([\![A'_1]\!]^{\mathcal{V}}, \ldots, [\![A'_k]\!]^{\mathcal{V}})$.*

At this point we observe that if the bridge rules in a $Log_A\mathbf{M}$ theory $\mathbb{T}$ observe some restrictions, then the order induced by the theory is classical (recall what it means for an order to be classical according to Definition 3).

Observation 1. *Let $\mathbb{T} = \langle (\mathbb{A}_1, \ldots, \mathbb{A}_k), \mathbb{R}, \mathbb{S} \rangle$ is a $Log_A\mathbf{M}$ theory, $\mathcal{V}$ a valuation, and $\preceq_{\mathbb{T}^{\mathcal{V}}}$ be a k partial-order on $\mathfrak{A}$. For every $(A_1, ..., A_k \longmapsto A'_1, ..., A'_k) \in \mathbb{R}$, and for every i, j such that $1 \leq i, j \leq k$ and $j \neq i$, $\preceq_{\mathbb{T}^{\mathcal{V}}}$ is classical in i if and only if $A'_i \neq \{\}$ just in case $A_j = A'_j = \{\}$ and $[\![A_i]\!]^{\mathcal{V}} \leq [\![A'_i]\!]^{\mathcal{V}}$.*

We next utilise a multifilter based on a $\mathbb{T}^{\mathcal{V}}$-induced order to define an extended logical consequence relation for each mental attitude.

Definition 11. *Let $\mathbb{T} = \langle(\mathbb{A}_1, ..., \mathbb{A}_k), \mathbb{R}, \mathbb{S}\rangle$ a $Log_A\mathbf{M}$ theory and $\preceq_{\mathbb{T}^{\mathcal{V}}}$ be a $\mathbb{T}^{\mathcal{V}}$-induced order. For every $\phi \in \sigma_P$, ϕ is an A_i consequence of $\mathbb{T}$ for $1 \leq i \leq k$, denoted $\mathbb{T} \models_{A_i} \phi$, if, for every valuation $\mathcal{V}$, $[\![\phi]\!]^{\mathcal{V}} \in \mathcal{F}_i$ where $\langle \mathcal{F}_1, ..., \mathcal{F}_k\rangle = \mathfrak{F}_{\preceq_{\mathbb{T}^{\mathcal{V}}}}(\langle [\![\mathbb{A}_1]\!]^{\mathcal{V}}, ..., [\![\mathbb{A}_k]\!]^{\mathcal{V}}\rangle, \mathbb{S})$.*

We now inspect the properties of our extended consequence relations. The following theorem states that each $\models_{A_i}$ is *monotonic* and has the distinctive properties of classical Tarskian logical consequence. Further, if $i \in \mathbb{S}$, then $\models_{A_i}$ observes a variant of the deduction theorem.

Theorem 2. *Let $\mathbb{T} = \langle(\mathbb{A}_1, ..., \mathbb{A}_k), \mathbb{R}, \mathbb{S}\rangle$ and $\mathbb{T}' = \langle(\mathbb{A}'_1, ..., \mathbb{A}'_k), \mathbb{R}', \mathbb{S}'\rangle$ be $Log_A\mathbf{M}$ theories with $\mathbb{S} = \mathbb{S}'$*

1. *If $\phi \in \mathbb{A}_i$ for some $\mathbb{A}_i \in \mathbb{A}$, then $\mathbb{T} \models_{A_i} \phi$.*
2. *If $\mathbb{T} \models_{A_i} \phi$, $\mathbb{A}_j \subseteq \mathbb{A}'_j$ for all $1 \leq j \leq k$, and $\mathbb{R}' \subseteq \mathbb{R}$, then $\mathbb{T}' \models_{A_i} \phi$.*
3. *Let $\mathbb{A}'_i = \mathbb{A}_i \cup \{\psi\}$ for some i such that $1 \leq i \leq k$, $\mathbb{A}'_j = \mathbb{A}_j$ for all $j \neq i$, and $\mathbb{R}' = \mathbb{R}$. If $\mathbb{T} \models_{A_i} \psi$ and $\mathbb{T}' \models_{A_i} \phi$, then $\mathbb{T} \models_{A_i} \phi$.*
4. *Let $\mathbb{A}'_i = \mathbb{A}_i \cup \{\phi\}$ for some i such that $1 \leq i \leq k$. If $i \in \mathbb{S}$, $\mathbb{R}' = \mathbb{R}$, and $\mathbb{T}' \models_{A_i} \psi$, then $\mathbb{T} \models_{A_i} \phi \Rightarrow \psi$.*

In the remainder of this section, we go back to our running example showing the consequences of the $Log_A\mathbf{M}$ theory of Example 2. In what follows, let $\mathbb{A}'_i = \{\phi \mid \mathbb{T} \models_{A_i} \phi\}$ for $1 \leq i \leq k$.

Example 3. Recall the $Log_A\mathbf{M}$ theory $\mathbb{T}$ in Example 2. In the following, we demonstrate the effect of the application of the bridge rules.

1. Initially, the applicable bridge rules are **r1**, **r4**, **r5**, and **r6**. This causes $\mathbf{I}(t)$ to be an obligation consequence of $\mathbb{T}$, f a fear consequence of $\mathbb{T}$, r an obligation consequence of $\mathbb{T}$, and $\neg r$ a fear consequence of $\mathbb{T}$.
2. Once $\mathbf{I}(t)$ becomes an obligation consequence and f and $\neg r$ become fear consequences, **r2** and **r7** become applicable. This causes t to be an intention consequence of $\mathbb{T}$, and $\neg r$ and f to be belief consequences of $\mathbb{T}$. Adding $\neg r$ to the belief consequences adds m to the belief consequences as well due to the belief $\neg r \Rightarrow m$.
3. Finally **r3** becomes applicable. This causes c to become an intention consequence of $\mathbb{T}$.

The following are the final consequences of $\mathbb{T}$.

- **Promises:** $\mathbb{A}'_1 = \{t\}$.
- **Beliefs:** $\mathbb{A}'_2 = \{\neg r \Rightarrow m, m \Rightarrow f, l, i \Rightarrow r \wedge t, \neg r, f, m\}$
- **Intentions:** $\mathbb{A}'_3 = \{t, c\}$
- **Fears:** $\mathbb{A}'_4 = \{f, \neg r\}$
- **Regrets:** $\mathbb{A}'_5 = \{\neg r\}$
- **Doubts:** $\mathbb{A}'_6 = \{\neg i\}$
- **Obligations:** $\mathbb{A}'_7 = \{\mathbf{I}(t), r\}$ □

5 Incorporating Non-Monotonicity in $Log_A\mathbf{M}$

According to Theorem 2, the consequence relations for the different mental attitudes of $Log_A\mathbf{M}$ have a monotonic nature. This means that newly acquired propositions in the different mental attitudes can never invalidate previous propositions. Moreover, the consistency of the mental attitudes of the agent is not guaranteed. These are inconvenient assumptions for some mental attitudes. For example, a natural consequence of typical incomplete knowledge about the world is that newly acquired beliefs can invalidate previous beliefs. Consequently, some intentions might be dropped as their supporting beliefs are not believed anymore. Furthermore, it would also make sense that the beliefs and intentions (for instance) of the agents are always collectively consistent if we are to model a rational agent. For these reasons, in this section we informally describe a non-monotonic extension of $Log_A\mathbf{M}$ where the consistency of selected mental attitudes is preserved. The extension we are proposing is a generalization of a framework we developed in [6] for non-monotonic practical reasoning with beliefs and motivations.

Towards incorporating non-monotonicity in $Log_A\mathbf{M}$, the first thing we do is that we associate *grades* with the different mental attitudes. The grades are reified objects with some total order on them and are taken to represent measures of trust or preference. Moreover, we define the agent's *character* as a total order over the mental attitudes that we wish to maintain their collective consistency. Whenever inconsistencies arise, the agent character and the grades of the contradictory propositions are utilised to resolve them. The agent character orders the attitudes from the least preffered to the most preferred. The agent always prefers to give up propositions from the least preferred attitude. Similarly, the least preferred proposition is the proposition with the least grade. We also enforce that the consequents of the bridge rules are only graded propositions to make sure that any newly added proposition has a grade that can be inspected if this newly added proposition causes a contradiction. To illustrate this, we go back to the $Log_A\mathbf{M}$ theory in Example 2. We first show the modified theory after we add grades to the different attitudes and modify the consequents of the bridge rules.

Example 4. In what follows, we use $\mathbf{P}$ for $\mathbf{A}_1$, $\mathbf{B}$ for $\mathbf{A}_2$, $\mathbf{I}$ for $\mathbf{A}_3$, $\mathbf{F}$ for $\mathbf{A}_4$, $\mathbf{R}$ for $\mathbf{A}_5$, $\mathbf{D}$ for $\mathbf{A}_6$, and $\mathbf{O}$ for $\mathbf{A}_7$ for readability. A possible *graded* $Log_A\mathbf{M}$ theory representing Example 1 is
$\mathbb{T} = \langle(\mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3, \mathbb{A}_4, \mathbb{A}_5, \mathbb{A}_6), \mathbb{R}, \mathbb{S}\rangle$ where:

- $\mathbb{A}_1$ represents Ted's promises. $\mathbb{A}_1 = \{\mathbf{P}(t, 3)\}$.
- $\mathbb{A}_2$ represents Ted's beliefs. $\mathbb{A}_2 = \{\mathbf{B}(\neg r \Rightarrow m, 1), \mathbf{B}(m \Rightarrow f, 5), \mathbf{B}(l, 10), \mathbf{B}(i \Rightarrow r \wedge t, 4)\}$.
- $\mathbb{A}_3$ represents Ted's intentions. $\mathbb{A}_3 = \{\}$.
- $\mathbb{A}_4$ represents Ted's fears. $\mathbb{A}_4 = \{\}$.
- $\mathbb{A}_5$ represents Ted's regrets. $\mathbb{A}_5 = \{\mathbf{R}(\neg r, 6)\}$.
- $\mathbb{A}_6$ represents Ted's doubts. $\mathbb{A}_6 = \{\mathbf{D}(\neg i, 3)\}$.
- $\mathbb{A}_7$ represents Ted's obligations. $\mathbb{A}_7 = \{\}$.
- $\mathbb{R}$ is the set of instances of the following rule schema where ϕ and g are variables. These are a modified version of the rules in Example 2 to add grades to the consequences of the bridge rules.

  r1. $P = \{\mathbf{P}(\phi, g)\} \longmapsto O = \{\mathbf{I}(\phi, g)\}$

r2. $O = \{\mathbf{I}(\phi, g)\} \longmapsto I = \{\mathbf{I}(\phi, g)\}$
r3. $I = \{\mathbf{I}(t, g)\} \longmapsto I = \{\mathbf{I}(c, g)\}$
r4. $B = \{\mathbf{B}(l, g)\} \longmapsto F = \{\mathbf{F}(f, g)\}$
r5. $R = \{\mathbf{R}(\neg r, g)\} \longmapsto O = \{\mathbf{O}(r, g)\}$
r6. $D = \{\mathbf{D}(\neg i, g)\} \longmapsto F = \{\mathbf{F}(\neg r, g)\}$
r7. $F = \{\mathbf{F}(\phi, g)\} \longmapsto B = \{\mathbf{B}(\phi, g)\}$

– $\mathbb{S} = \{2\}$. □

Now consider the above graded $Log_A\mathbf{M}$ theory and suppose that we only care that Ted's collective beliefs, promises, obligations, and intentions are consistent. Given that, for instance, Ted believes $\mathbf{B}(l, 10)$ and does not believe $\neg l$, it would make sense for him to accept l despite his uncertainty about it. Similarly, it would make sense for Ted to add t to his promises if they do not conflict with other beliefs, promises, obligations, or intentions. However, if we only use multifilters, we will never be able to reason with those nested graded attitudes as they are not themselves in the agent's theory but only grading propositions thereof. For this reason, we extend our notion of multifilters into a more liberal notion of *graded multifilters* to enable the agent to conclude, in addition to the consequences of the initial theory, attitudes graded by the initial attitudes (like l and t). Should this lead to contradictions, the agent's character and the grades of the contradictory propositions are used to resolve them. In what follows, we show how graded multifilters are used to get the set of consequences for each mental attitude.

Example 5. We first apply the bridge rules to $\mathbb{T}$ just like we did in Example 3. We get the following updated mental state.

– **Promises:** $\mathbb{A}'_1 = \mathbb{A}_1$.
– **Beliefs:** $\mathbb{A}'_2 = \mathbb{A}_2 \cup \{\mathbf{B}(f, 10), \mathbf{B}(\neg r, 3)\}$.
– **Intentions:** $\mathbb{A}'_3 = \{\mathbf{I}(t, 3), \mathbf{I}(c, 3)\}$.
– **Fears:** $\mathbb{A}'_4 = \{\mathbf{F}(f, 10), \mathbf{F}(\neg r, 3)\}$.
– **Regrets:** $\mathbb{A}'_5 = \mathbb{A}_5$
– **Doubts:** $\mathbb{A}'_6 = \mathbb{A}_6$
– **Obligations:** $\mathbb{A}'_7 = \{\mathbf{I}(t, 3), \mathbf{O}(r, 6)\}$.

Next, we admit the graded attitudes in the initial theory. The following becomes the updated mental state of the agent.

– **Promises:** $\mathbb{A}''_1 = \mathbb{A}'_1 \cup \{t\}$.
– **Beliefs:** $\mathbb{A}''_2 = \mathbb{A}'_2 \cup \{\neg r \Rightarrow m, m \Rightarrow f, l, i \Rightarrow r \wedge t, f, \neg r, m\}$.
– **Intentions:** $\mathbb{A}''_3 = \{t, c\}$.
– **Fears:** $\mathbb{A}''_4 = \{f, \neg r\}$.
– **Regrets:** $\mathbb{A}''_5 = \mathbb{A}'_5 \cup \{\neg r\}$.
– **Doubts:** $\mathbb{A}''_6 = \mathbb{A}'_6 \cup \{\neg i\}$.
– **Obligations:** $\mathbb{A}''_7 = \{r\}$.

Note that we add m to the agent's beliefs not because it was extracted out of a graded belief, but because it follows from $\neg r \Rightarrow m$ and $\neg r$ that was just extracted out of $\mathbf{B}(\neg r, 3)$. At this point, Ted's beliefs and obligations are contradictory as his beliefs

include $\neg r$ and his obligations include r. This is where the agent's character come into play. If Ted's character prefers to give up his beliefs, then $\neg r$ will be retracted from his beliefs. Otherwise, r will be given up as an obligation.

Now suppose that Ted acquires the new belief that he will not be fired $\mathbf{B}(\neg f, 15)$. We extract $\neg f$ and add it to Ted's beliefs. Once we do this, Ted's set of beliefs becomes contradictory as it contains $\neg f$ and f. Since the contradiction is now within the same attitude, we allude to the grades of the contradictory propositions. Since $\neg f$ has the grade of 15 and f has the grade of 10, then f will be kicked out of Ted's beliefs resolving the contradiction. □

In general, to resolve inconsistencies among the attitudes we select to be consistent, we always remove the propositions with the lowest grade in the least preferred attitude according to the agent character. Next, any propositions supported only by the removed propositions are removed as well. If two contradictory propositions have the same grade, they both go away.

6 Conclusion

In this paper, we presented a general algebraic framework for reasoning with mental states. We also provided semantics for an algebraic logic, $Log_A\mathbf{M}$, where any set of mental attitudes can be represented. We defined a monotonic consequence relation for each mental attitude and showed that the consequence relations observe the distinctive properties of Tarskian logical consequence. Moreover, we informally described how $Log_A\mathbf{M}$ can be extended to handle non-monotonic reasoning with graded mental attitudes. We are currently working on a proof theory for the non-monotonic version of $Log_A\mathbf{M}$. Reasons for the different mental attitudes are to be computed in the same way reason-maintenance systems compute supports for beliefs. Hence, the end result will be a versatile framework for reasoning with graded mental attitudes giving rise to an explainable AI system.

References

1. James Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154, 1984.
2. George Bealer. Theories of properties, relations, and propositions. *The Journal of Philosophy*, 76(11):634–648, 1979.
3. Ronen I. Brafman and Moshe Tennenholtz. Modeling agents as qualitative decision makers. *Artificial Intelligence*, 94(1-2):217–268, 1997.
4. Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on Autonomous agents*, pages 9–16, 2001.
5. Alonzo Church. On carnap's analysis of statements of assertion and belief. *Analysis*, 10(5):97–99, 1950.
6. Nourhan Ehab and Haythem O. Ismail. Algebraic foundations for non-monotonic practical reasoning. In Ivan José Varzinczak and María Vanina Martínez, editors, *Proceedings of the 18th International Workshop on Non-Monotonic Reasoning (NMR2020)*, 2020. To appear.

7. Nourhan Ehab and Haythem O. Ismail. $Log_A\mathbf{G}$: An algebraic non-monotonic logic for reasoning with graded propositions. *Annals of Mathematics and Artificial Intelligence*, 2020.
8. Nourhan Ehab and Haythem O. Ismail. Reasoning with artificial mental states: An algebraic approach. Technical report, German University in Cairo, 2020. https://met.guc.edu.eg/Repository/Faculty/Publications/950/FCR2020-Appendix.pdf.
9. Kenneth M. Ford, Patrick J. Hayes, Clark Glymour, and James Allen. Cognitive orthoses: toward human-centered AI. *AI Magazine*, 36(4):5–8, 2015.
10. Harry G. Frankfurt. Freedom of the will and the concept of a person. In *What is a person?*, pages 127–144. Springer, 1988.
11. Haythem O. Ismail. $Log_A\mathbf{B}$: A first-order, non-paradoxical, algebraic logic of belief. *Logic Journal of the IGPL*, 20(5):774–795, 2012.
12. Haythem O. Ismail. Stability in a commonsense ontology of states. *Proceedings of the Eleventh International Symposium on Logical Formalization of Commonsense sense Reasoning (COMMONSENSE 2013)*, 2013.
13. Haythem O. Ismail. The good, the bad, and the rational: Aspects of character in logical agents. In Alia ElBolock, Yomna Abdelrahman, and Slim Abdennadher, editors, *Character Computing*. Springer, 2020.
14. Hector J. Levesque. Making believers out of computers. *Artificial Intelligence*, 30(1):81–108, 1986.
15. John McCarthy. Ascribing mental qualities to machines. *Philosophical Perspectives in Artificial Intelligence, Humanities Press*, pages 161–195, 1979.
16. John McCarthy. The little thoughts of thinking machines. *Psychology Today*, 17(12):46–49, 1983.
17. John McCarthy and Patrick Hayes. Some philosophical problems from the standpoint of artificial intelligence. In D. Meltzer and D. Michie, editors, *Machine Intelligence*, volume 4, pages 463–502. Edinburgh University Press, Edinburgh, Scotland, 1969.
18. Paul McNamara. Deontic logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.
19. Allen Newell et al. The knowledge level. *Artificial intelligence*, 18(1):87–127, 1982.
20. Pietro Panzarasa, Nicholas R. Jennings, and Timothy J. Norman. Formalizing collaborative decision-making and practical reasoning in multi-agent systems. *Journal of logic and computation*, 12(1):55–117, 2002.
21. Terence Parsons. On denoting propositions and facts. *Philosophical Perspectives*, 7:441–460, 1993.
22. Donald Perlis. Languages with self-reference II: Knowledge, belief, and modality. *Artificial Intelligence*, 34(2):179–212, 1988.
23. Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In *ICMAS*, volume 95, pages 312–319, 1995.
24. H.P. Sankappanavar and Stanley Burris. A course in universal algebra. *Graduate Texts Math*, 78, 1981.
25. Stuart C. Shapiro. Belief spaces as sets of propositions. *Journal of Experimental & Theoretical Artificial Intelligence*, 5(2-3):225–235, 1993.
26. Yoav Shoham. An overview of agent-oriented programming. *Software agents*, 4:271–290, 1997.
27. Yoav Shoham and Steve B. Cousins. Logics of mental attitudes in AI. In *Foundations of Knowledge Representation and Reasoning*, pages 296–309. Springer, 1994.
28. Hans van Ditmarsch, Joseph Halpern, and Barteld Kooi. An introduction to logics of knowledge and belief. In Hans van Ditmarsch, Joseph Halpern, Wiebe van der Hoek, and Barteld Kooi, editors, *Handbook of Epistemic Logic*, pages 1–51. College Publications, 2015.