

Towards a Logic of Graded Normativity and Norm Adherence

(Draft version, 03/11/2007)

Matthias Nickles

AI/Cognition Group, Computer Science Department
Technical University of Munich
Boltzmannstr. 3, D-85748 Garching b. Muenchen, Germany
nickles@cs.tum.edu

Abstract. A key focus of contemporary agent-oriented research and engineering is on *open multiagent systems* composed of truly autonomous, interacting agents. This poses new challenges, as entities in open systems are usually more or less mentally opaque (e.g., possibly insincere), and can enter and leave the system at will. Thus interactions among such black- or gray-box entities usually imply more or less severe contingencies in behavior: Among other issues, in principle, the adherence of agents to norms cannot be guaranteed in such systems. As a response to this issue, this paper proposes a logic-based approach based on the notion of (possibly probabilistic) behavioral expectations, which are stylized either as adaptive (i.e., predictive) or normative (i.e., prescriptive). Some features of this approach are the enabling of "soft norms" which are automatically weakened to some degree if contradicted at runtime, and the possibility to quantify norm adherence using the measurement of norm deviance.

Keywords. Norms, Modal Logic, Computational Expectations, Social AI, Belief Revision

1 Introduction

A key focus of contemporary agent-oriented research and engineering is on *open multiagent systems* composed of truly autonomous, interacting agents. This poses new challenges, as agents in open systems are usually more or less mentally opaque (e.g., possibly insincere), and can enter and leave the system at will. Among other issues, in principle, the adherence of agents to norms cannot be guaranteed in such systems. As a response, this paper proposes a logic-based approach to the modeling of open multiagent systems in form of probabilistic *behavioral expectations*, which are stylized either as adaptive (i.e., predictive) or normative (i.e., prescriptive), and which can be adapted dynamically at runtime. More concretely, we propose first a probabilistic logic for the representation of agent actions and other events, event sequences, and beliefs and intentions. Then, we define (event-related) normative and adaptive expectations on top of this logic. Doing so, our proposal is a more or less direct logic-based variant of our approach proposed in (Brauer, Nickles, Rovatsos, Weiß & Lorentzen 2002, Lorentzen

& Nickles 2002, Nickles, Rovatsos & Weiß 2005), adopting the sociological viewpoint regarding expectations and norms which was introduced in (Luhmann 1995a). With the approach presented in (Castelfranchi & Lorini 2003) it has in common that expectations are based on intentions and beliefs regarding future events, but otherwise the two approaches are unrelated.

Being a normative expectation is not sufficient to constitute a full-fledged social *norm* (Boella, van der Torre & Verhagen 2007) by itself (mainly because expectations do not by themselves lead to their announcement and enforcement), but all social norms are necessarily grounded in normative expectations (Luhmann 1995b, Lorentzen & Nickles 2002): Because only the behavior of an autonomous agent within some shared environment is visible for an observer, while his mental state remains obscure, *beliefs* and *demands* directed to the respective other agent can basically be stylized only as mutable behavioral expectations which are *fulfilled* or *disappointed* in future events (Luhmann 1995b). In the case of disappointment, an expectation can either be revised in order to consider the new perception accurately (so-called *adaptive* expectations), or the expecter decides to keep this expectation even contra-factually (so-called *normative* expectations), or to revise (resp. maintain) it only to a certain degree (*adaptive-normative* expectations). In the two latter cases, the expectation holder likely also decides to take action in order to make further disappointments of this expectation less probable (by, e.g., sanctioning unexpected - so-called *deviant* - behavior). And in any case, the expectation can be strengthened/weakened if an expected repeatable event turns out to be useful/useless afterwards.

Thus, we define normative expectations (and thus norms) via the degree of *resistance* to environmental (e.g., social) dynamics *in the course of time*, wrt. how somebody else should behave from the viewpoint of the expectation-holder. In addition, expectations can address the behavior of the expecter *himself* also, which can be useful for the expecter in order to model his self-commitments (intentions regarding own actions), and to communicate them to other agents in form of uttered expectations.

In order to make expectations expected (and thus socially relevant, e.g., as a norm), any kind of expectation needs of course to be *communicated* to others and to be armed with sanctions, if necessary, but concrete ways to do so are outside the scope of this paper.

As for the modeling of norms, we are mainly interested in representing behavioral norms as mental attitudes of some norm giver, such as the designer of the multiagent system (MAS). By representing even the designer of an agent-based application as an agent conceptually, we suggest that the designer of open MAS should not and can not be granted the omniscient, almighty position, as it is the case in by far most current frameworks (e.g., (Ndumu, Collins, Owusu, Sullivan & Lee 1999, Bellifemine, Poggi & Rimassa 2000, Bauer & Müller 2003)). Rather, we see her in the role of a *primus inter pares* among other agents, who, although equipped with more power than “real” agents, should aim for her goals socially (= communicatively) in interaction with the other agents as far as possible. In addition, the openness of open MAS suggests that the development of such systems can only be done in an evolutionary manner, with the need to monitor the system and to improve its model even after deployment during runtime. A way to put the conceptualization of system designers as agents into practice in

a semi-automatic manner is to assign the designer an intelligent, agent-like case tool, as we have proposed in (Brauer et al. 2002, Nickles, Rovatsos & Weiß 2005).

The rest of this paper is organized as follows: The next section introduces a formal language for the representation of mental attitudes of expectation-holders. Section 3 then defines expectations using the means of this language, and Section 4 outlines how expectations can be computed and adapted during runtime. Section 5 concludes.

2 A logic with modalities for intentions and uncertain beliefs

We use the languages proposed in (Cohen & Levesque 1990), Bacchus' logic of uncertain beliefs (Bacchus 1990) and own works (Nickles, Fischer & Weiß 2005, Fischer & Nickles 2005) as a basis. Most aspects of these formalisms have the advantage that they are well established and researched in Distributed AI in the context of modeling agent beliefs and intentions.

The deliberately rich language¹ \mathcal{L}^{probDL} we propose allows for the representation of

- **event sequences and test expressions.**
- **uncertain beliefs** denoting the an agent believes something with a certain degree. This requires us to use a different semantics compared to the standard belief-intention logics (i.e., for non-gradual beliefs).
- **agent intentions.** This is done in the same way as described above for standard Kripke-type belief-intention logics. Note that intentions might encode demands the agent is self-committed to and which are directed to other agents (e.g., using $Int(me, otheragent. done(someaction))$).
- **normative and non-normative expectations, and norm deviance.** These major enhancements are spelled out in the next section.

2.1 Syntax

Definition 1. *The syntax of well-formed \mathcal{L}^{probDL} formulas F, F_1, F_2, \dots and processes α, β, \dots is given by*

$$\begin{aligned}
 F, F_1, F_2, \dots ::= & p \mid \top \mid \perp \mid \neg F_1 \mid F_1 \wedge F_2 \mid F_1 \vee F_2 \mid F_1 \rightarrow F_2 \\
 & \mid F_1 \leftrightarrow F_2 \mid \exists x F_1 \mid \forall x F_1 \mid \langle Pred \rangle (x_1, \dots, x_n) \\
 & \mid done(\alpha) \mid happens(\alpha) \mid now(\langle Time \rangle) \\
 & \mid Bel(a, F) \mid Bel(a, F, d) \mid Bel(a, F|c, d) \mid Int(a, F) \\
 Expect(agent, normativity, event|context, strength) & \mid \Delta(event|context, deviance)
 \end{aligned}$$

$$\alpha, \beta, \dots ::= action \mid a_i.action \mid any \mid \alpha; \beta \mid \alpha \cup \beta \mid \alpha * \mid F?$$

Here,

¹ which is expected to be easily tailorable to concrete application needs in order to reduce complexity.

- $a, a_i \in \text{Agents}$ (cf. below), being agents;
- $x, y, z, i, j, k, x_i, y_i, z_i$ being variables;
- $\langle \text{Pred} \rangle$ being a predicate symbol;
- $\text{action}, a_i.\text{action} \in A$, with A denoting the set of elementary actions; every elementary action can be indexed with an agent (e.g., $\text{agent}_3.\text{assert}$ represents the communication act “assert” uttered by agent_3);
- any is an arbitrary action;
- $\langle \text{Time} \rangle$ is a time point, denoted as a natural number;
- $\alpha; \beta$ denotes sequential process combination (i.e., (sub-)process α is followed by β);
- $\alpha \cup \beta$ denotes non-deterministic choice between α and β ;
- α^* denotes zero or more iterations of α ;
- $F?$ is a test operation (i.e., the process proceeds if F holds true);
- $\text{Bel}(a, F, d)$ denotes that agent a (sincerely) believes with degree d that F , with d being a real valued probability measure and $0 \leq d \leq 1$;
- $\text{Bel}(a, F)$ is a shortcut for $\text{Bel}(a, F, 1)$;
- $\text{Int}(a, F)$ denotes that agent a (sincerely) intends that F ;
- $\text{Bel}(a, F|c, d)$ and $\text{Bel}(a, F|c) := \text{Bel}(a, F|c, 1)$ are shortcuts denoting contextualized beliefs (cf. below);
- $\text{Expect}(\text{agent}, \text{normativity}, \text{event}|\text{context}, \text{strength})$ (see next section), and
- $\Delta(\text{event}|\text{context}, \text{deviance})$ (see next section).

See below for the meaning of *done*, *happens*, and *now*. Further, let $\text{Agents} = \{a_i\}$ be the set of all agents (those currently in the MAS under observation, as well as all other possible agents), $\Theta = \{\alpha, \beta, \gamma, \dots\}$ the set of all syntactically valid processes, and $\Phi = \{F, F_1, F_2, \dots\}$ the set of all $\mathcal{L}^{\text{probDL}}$ formulas.

Meta- $\mathcal{L}^{\text{probDL}}$ statements are given in natural-like language (e.g., “If $\models F$ then $\models \text{Bel}(a, F, 1)$ ”). In addition, we will sometimes write calculations and other functional expressions directly within $\mathcal{L}^{\text{probDL}}$ formulas for simplicity, like in $\text{Bel}(a, F, f(1)/2)$.

2.2 Model-based semantics

We propose the usual Kripke-style semantics of belief and intentions, with the “possible worlds” (future as well as past) consisting of finite sequences of events here. Some of these worlds are worlds the agents considers being true (i.e., consistent with the agents belief), others are those worlds the agents wants to become true, thus consistent with the agents intentions². Event sequences (or *courses*) are simply timely ordered, atomic events such as agent actions, with finite length, denoted as $\text{event}_1; \dots; \text{event}_n$. If we additionally assume some meaningful correlation or even causation among events, we speak about (event) *processes*.

A $\mathcal{L}^{\text{probDL}}$ model is a structure $(\Theta, \text{Events}, \text{Agents}, \text{Trajectories}, B, I, \Phi)$, where Θ is the universe of discourse, Agents is a set of agents as specified above, Events is a

² For simplicity, we do not explicitly introduce goals and desires. Instead, we assume that for the purpose of this work that long-term and, in case of failure, possibly re-established intentions could act as such.

set of events, $Trajectories \subseteq \{(n, e) : n \in \mathbb{N}, e \in Events\}$ is a linearly ordered, denumerable set of worlds in form of event sequences, $B : Agents \times Situations \rightarrow [0; 1]$ is a personalized and discrete probability measure over the worlds at all particular time points (so-called *situations*), with $Situations = \{(w, i) : i \in \mathbb{N}, w \in Trajectories\}$ and $\sum_{s \in Situations} B(agent, s) = 1$ for some particular agent $agent$, $I \subseteq Trajectories \times Agents \times \mathbb{N} \times Trajectories$ is the accessibility relation for the agent's intentions (extended with time points, cf. below), and Φ is a predicate interpreting function. B and I are serial, and B is in addition euclidian and transitive. This structure defines an agent-specific probability distribution over $2^{Situations}$ by $B(agent, S \subseteq Situations) = \sum_{s \in S} B(agent, s)$.

Let M be a \mathcal{L}^{probDL} model, σ a sequence of events (possible world), $n \in \mathbb{N}$ (a time point), ν be a set of variable bindings, relating elements of Θ , $Events$ and σ to variables. The satisfaction of some \mathcal{L}^{probDL} formula F by M, σ, ν is written as $M, \sigma, \nu, n \models F$. To express that an event occurs between two time points n and m , we write $M, \sigma, \nu, n \triangleright \alpha \triangleleft m$ (the exact meaning of this is given below).

With this, the model theoretic semantics of \mathcal{L}^{probDL} is then given by the following rules:

1. $M, \sigma, \nu, n \models now(\langle Time \rangle)$ iff $\nu(\langle Time \rangle) = n$.
2. $M, \sigma, \nu, n \models happens(\alpha)$ iff $\exists m, m \geq n : M, \sigma, \nu, n \triangleright \alpha \triangleleft m$ (i.e., α is here an event or event sequence which happens after time n).
3. $M, \sigma, \nu, n \models done(\alpha)$ iff $\exists m, m \leq n : M, \sigma, \nu, m \triangleright \alpha \triangleleft n$ (i.e., α is here an event or event sequence which happened just before time n).
4. $M, \sigma, \nu, n \models Bel(a, F, d)$ iff $B(a, \{s : M, \sigma, \nu, n \models F\}) = d$. This expresses that agent a believes F with strength d if and only if the personalized probability measure B equals d for all situations where F holds.
Since the relation B in \mathcal{L}^{probDL} models is a probability measure, $d = 1 - d'$ if $Bel(a, F, d) \wedge Bel(a, \neg F, d')$.
5. $M, \sigma, \nu, n \models Int(a, F)$ iff for all $\sigma^*, (\sigma, n, \nu(a), \sigma^*) \in I$. This rule states that F follows from the agents intentions iff F is true in all possible worlds (event sequences) accessible via I , at time n . Observe that it is not required that F is brought about by a . The intention to perform or let someone else perform some action can trivially be expressed with $Int(agent_1, done(agent_2.action))$.

The following defines the occurrence conditions for single and compound events:

1. $M, \sigma, \nu, n \triangleright \alpha \triangleleft n + i$ iff $\nu(\alpha) = \alpha_1; \dots; \alpha_i$ and $\sigma_{n+j} = \alpha_j, 1 \leq j \leq i$. This means that the event sequence α happens next to time point n in world σ .
2. $M, \sigma, \nu, n \triangleright \alpha \cup \beta \triangleleft m$ iff $M, \sigma, \nu, n \triangleright \alpha \triangleleft m$ or $M, \sigma, \nu, n \triangleright \beta \triangleleft m$ (i.e., either α or β occurs in the time interval $n \dots m$).
3. $M, \sigma, \nu, n \triangleright \alpha; \beta \triangleleft m$ iff $\exists k, n \leq k \leq m : (M, \sigma, \nu, n \triangleright \alpha \triangleleft k) \wedge (M, \sigma, \nu, n \triangleright \beta \triangleleft m)$ (i.e., β follows α).
4. $M, \sigma, \nu, n \triangleright F? \triangleleft m$ iff $M, \sigma, \nu, n \models F$. I.e., the test expression $F?$ occurs iff F is true.

5. $M, \sigma, \nu, n \triangleright \alpha * \triangleleft m$ iff $\exists n_1, \dots, n_k, n_1 = n, n_k = m \forall i, 1 \leq i \leq m : M, \sigma, \nu, n_i \triangleright \alpha \triangleleft n_{i+1}$. This states that $\alpha*$ occurs iff a sequence of α s occurs.

Like in temporal logic, we can express that F will eventually be true using $\exists \alpha \in A : happens(\alpha; F?)$. $\neg \exists \alpha \in A : happens(\alpha; \neg F?)$ says that F holds always.

We define logically contextualized beliefs by $Bel(a, F|c, d)$, $c \in \Phi \equiv Bel(a, c \rightarrow \exists \alpha : happens(\alpha; F?), d)$, and beliefs contextualized with event sequences α with $Bel(a, F|\alpha, d)$, $\alpha \in \{\alpha_1; \dots; \alpha_n : \alpha_i \in A\} \equiv Bel(a, happens(\alpha; F?), d)$. Sometimes, we abbreviate $Bel(a, done(\alpha), \dots)$ with $Bel(a, \alpha, \dots)$.

The semantics of $\langle Pred \rangle (\dots)$, \neg , \wedge , $=$, \vee , \top , \perp , \exists , \forall , \rightarrow and \leftrightarrow is as usual in FOL with equality.

We provide a partial axiomatization, focusing on belief and intention modalities. The \mathcal{L}^{probDL} belief axioms schema includes the well-known K45 (aka weak S5 plus consistency) modal logic axioms schema, adapted for personalized, probabilistic beliefs:

Axioms 1.1:

K (closure under consequence) $(Bel(a, F) \wedge Bel(a, F \rightarrow F')) \rightarrow Bel(a, F')$

D (consistency) $\neg Bel(a, \perp)$

4 (closure under positive introspection) $Bel(a, F) \rightarrow Bel(a, Bel(a, F))$

5 (closure under negative introspection) $\neg Bel(a, F) \rightarrow Bel(a, \neg Bel(a, F))$

Sometimes, a belief logic also includes the *necessity rule*: "If $\models F$ then $\models Bel(a, F, 1)$ ", which we do not adopt. I.e., our agents need not to be aware of valid formulas.

Contrary to the famous approach (Cohen & Levesque 1990), which focuses mainly on the interaction of intentions and goals, but in accordance with (Herzig & Longin 2002), we think that the relationship of intentions to the agent's belief is most important. It is governed by the following Bel-Int bridge axioms:

Axioms 1.2:

BelInt1 $Int(a, F) \rightarrow \neg Bel(a, F)$. Agent a intends F to become true only if she does not already believe that F is true already.

BelInt2 $Int(a, Bel(a, event|context))$

$\wedge \neg Bel(a, event|context) \Rightarrow Int(a, event)$.

BelInt2' (alternatively) $Int(a, Bel(a, event|context, e))$

$\wedge \neg Bel(a, event|context, e) \Rightarrow Int(a, occurs(event|context, e))$.

This probabilistic version of $Rel_{IntBel2}$ in (Herzig & Longin 2002) would express that disbelief in the occurrence of an event with probability e while intending to believe the event occurs with this probability forces the agent to intend the event to occur with probability e

(denoted as $Int(agent, occurs(event|context, e))$). This also expresses that in

case the agent has no particular belief regarding the occurrence of this event, she can bring about her introspective intention to belief in the event even without intending the event itself (e.g., by exploring new perceptions, or by improving her reasoning process). This axiom becomes very important later in the context of normative and adaptive-normative expectations. Unfortunately, the modality $Int(\dots occurs)$ is not part of \mathcal{L}^{probDL} , and maybe shouldn't be, since it is not clear what "shall occur with a certain probability" means exactly. Since we feel that adding such a modality could be problematic, we provided a simpler, harmless variant in form of $BelInt2$.

BelInt3 $Int(a, F) \rightarrow Int(a, Bel(a, F))$. Note that the opposite direction should not hold: There are other means than intentions to change one's belief, e.g., exploration.

BelInt4 $Bel(a, Int(a, F)) \rightarrow Int(a, F)$. This allows for introspection regarding an agent's intentions.

The following schematic axioms deal with uncertain belief:

Axioms 1.3:

$$\begin{aligned} Bel(a, F, d) &\rightarrow d \geq 0 \\ Bel(a, F, d) \wedge Bel(a, \neg F, d') &\rightarrow d' = 1 - d \\ Bel(a, F \wedge F', d) \wedge Bel(a, F \wedge \neg F', d') \wedge Bel(a, F, d'') &\rightarrow d'' = d + d' \\ \forall x : Bel(a, F, 1) &\rightarrow \bigwedge Bel(a, \forall x : F, 1) \end{aligned}$$

For δ and $Expect$, cf. the next section.

Being interested in open systems with truly autonomous agents only, we deliberately do not propose any axioms which would enforce sincerity, collaboration or other properties of benevolence.

3 Expectations as combined mental attitudes with temporal dynamics

Expectations can be weighted in two ways, namely, w.r.t. their *strength* and w.r.t. their *normativity* (or inversely, their *adaptability*). The strength of an expectation indicates its "degree of expectedness" (also called *expectability*): the weaker (stronger) the expectation is, the less likely is or should be its expected fulfilment (violation). Against that, the normativity of an expectation (*both weak and strong expectations*) indicates its deliberate "degree of changeability": the more normative (adaptive) an expectation is, the smaller (greater) is the change in its strength when being contradicted by unpredicted actual actions. With that, the strength of a lowly normative expectation tends to change faster, whereas the strength of a highly normative expectation is maintained in the longer term even if it is obviously inconsistent with reality (e.g., some other agents' activities). Fully normative expectations (*normativity = 1*) ignore the actual occurrences of their modeled events completely, as long as they are not adapted "manually", whereas fully adaptive expectations (*normativity = 0*) follow the resp. beliefs of the expecting agents, given these beliefs follow themselves any incoming new information regarding the expected events. Thus it is assumed that there is a continuous transition

from weak to strong strength and from low to high normativity. The difference between the probability and the expectability (normativity-biased probability) of a certain event is called *deviance*. So, we can model both gradual and, to some degree, auto-adaptive normative expectations - in contrast to, e.g., binary-style modalities like obligation and permission as in deontic logic.

Some examples (adopted from (Brauer et al. 2002)) of combinations of expectation strength, normativity and deviance:

rules that govern criminal law (strong/non-adaptable/rather low deviance in western countries: even hundreds of actual murders will not alter the respective laws, and most people think of murder as a rather exceptional event);

habits (strong/adaptable, low deviance: before the times of fast food, people took full service in restaurants for granted, but as fast food became popular, they were willing to abandon this expectation);

adherence to public parking regulations (strong/hardly adaptable/high deviance: almost everyone violates them even if they are, in principle, rigid);

and *shop clerk friendliness* (weak/adaptable/indefinite deviance: most people expect bad service but are willing to change their view once encountering friendly staff).

Thus, the term “expectation” is inherently ambiguous, as it deliberately combines subjective, demanding expectations (reflecting the goals and intentions of the expecting agent) and the empirical likeliness of events (desired or not). In this regard it is worth to state that even the strengths of fully-adaptive expectations are not necessarily probabilities (from a frequentist point of view), because expectations are maintained (“expected”) as a part of the belief a subjective observer has, and do not necessarily take into account enough “real world” facts to determine expectation strengths objectively when he sets up his expectations. So, not only (adaptive-)normative, but also fully-adaptive expectations could theoretically be used to represent individual, contra-factual preferences (“desired probabilities”, so to say) instead of likelihoods. But such contra-factual yet non-normative expectations converge immediately to probabilities, since they are “willing to learn”, so to say.

Starting from these observations, we define the semantics of a so-called *normative* ($normativity = 1$) or *adaptive-normative* ($0 < normativity < 1$) expectation held by some agent as his intention to make (or keep) the strength of his belief regarding the (re-)occurrence of the expected event identical with the strength of this expectation. This can be weaker than to intend a certain probability of the event, but as we will see later, in the most common case we actually get by with defining (adaptive-)normative expectations as the intention to make the environment conforming to the expected state to some degree. In contrast expectations without any normativity, simply corresponding to uncertain beliefs, are called *adaptive expectations*.

At this, “intending a probability” can be understood as either aiming at bringing about a certain frequency of a repeatable event, or as the will to provide occurrence conditions for the event that make it probable to a certain degree.

Formally, an agent’s expectation (denoted as *Expect*) is a mental attitude, represented as a logic modality, and defined as follows:

Definition 2.

$$Expect(agent, \psi, event|context, e) :\Leftrightarrow \begin{cases} Bel(agent, event|context, e) \\ \vee Int(agent, Bel(agent, event|context, e)) \\ \text{if } \psi > 0 \\ Bel(agent, event|context, e) \text{ otherwise} \end{cases}$$

Hereby, e is the expectability, and $\psi \in [0; 1]$ is the normativity of the expectation. $\psi = 0$ leads to the special case of an adaptive expectation.

$event$ can theoretically be any proposition, but focusing on actions, if we use $event$, it should in fact be $done(event)$ (the $done$ operator omitted for simplicity).

For convenience, we set

$Expect_t(agent, \psi, event|context, e) \equiv Expect(agent, \psi, event|context, e) \wedge now(< t >)$ to denote expectations held at a certain time.

$\psi = 0$ leads to the special case of an adaptive expectation.

$Bel(agent, event|context, b)$ denotes that $agent$ believes that $event$ occurs with probability b in $context$ ³, and $Int(agent, p)$ denotes that agent intends p to become true (if $agent$ is not capable to bring about the desired fact or action directly by herself, this shall include the intention to make other agents bring about p etc., i.e., to use them like a tool)

We write $Expect(agent, event|context, e)$ as an abbreviation of

$Expect(agent, 0, event|context, e)$, and $Expect_t$ for $Expect$, when the time point t at which the expectation is held matters and can not be derived from the context (for ψ , Int and Bel analogously). Note that t is not the time point at which the event (should) occur(-s). If we would like to express that some event will or should happen at a certain time, we would have to encode this time within $context$.

The exact normativity (except from distinguishing if it is above zero or not) is not used in Definition 2, because the normativity prescribes how an expectability auto-evolves *in the course of time* with new information, if the expectability it is not set “manually”. If the normativity is zero, the expectation is set equal to the belief of the expecter immediately. Otherwise, the expectability adopts gradually to the belief when both differ, with a “learning rate” of the expectation inverse to the normativity.

Our definition of expectation is build straightforwardly upon probabilistic versions of the KD45 and belief-intention axioms usually used for multi-modal logics of mental attitudes (e.g. (Herzig & Longin 2002)), and is related to Sadek’s *want* attitude (Sadek 1992).

Given the agent’s belief (e.g., obtainable from an expectation via the so-called *deviance*, cf. below), the following proposition obviously holds, given

$Expect(agent, \psi, event|context, e)$:

Observation 1:

³We can also use this syntax to denote *expected expectations*: $Expect(agent_1, \dots, Expect(agent_2, \dots))$.

$$\begin{aligned} &Int(agent, Bel(agent, event|context, e)) \\ &\quad \text{if } (\psi > 0 \wedge Bel(agent, event|context, ne), ne \neq e) \\ &Bel(agent, event|context, e) \text{ otherwise} \end{aligned}$$

If we would either drop the usual $Bel(p) \rightarrow \neg Int(p)$ axiom in belief-intention logics, or introduce alternatively *maintenance intentions* (Bratman 1987) (denoted as Int^M), Definition 2 would change to

Definition 3. (alternatively to Definition 2)

$$Expect^{alt}(agent, \psi, event|context, e) :\Leftrightarrow \begin{cases} Int^M(agent, Bel(agent, event|context, e)) \\ \text{if } \psi > 0 \\ Bel(agent, event|context, e) \text{ otherwise} \end{cases}$$

The agent can achieve the intention to revise his belief in several ways, possibly even concurrently.

- i. Change the world** This is considered to be the usual way to enforce adaptive-normative and normative expectations, either by execution of the expected events by the expecting agent herself, or by bringing about the intended events indirectly (e.g., by asking other agents to do so).
- ii. Explore** The agent can try to obtain new perceptions in order to change his belief by exploration. Here, the (adaptive-)normative expectation serves as a kind of hypotheses, and the agents strives after new evidence in order to support or refute it.
- iii. Wait** This is actually not covered by the original intention at time t , but is a way to automatically decrease the “strength” of the intention (i.e., the degree and duration of the self-commitment) in consecutive time steps instead: If the normativity is below 1, in the longer term the expectation *learns* (i.e., adapts to the current probability), provided the probabilities of a certain event remain stable enough to be learnable (cf. 4). Practically, this happens if the expectation holder failed to decrease the deviance actively (due to insufficient social power, for example). The adaptation of the expectability to the probability in this case can nevertheless be desired, and it can even be a prerequisite for the enforcement of less flexible and thus likely more important expectations.
- iv. Ignore the deviance** Here, the agents simply believes that the expected event will occur, possibly ignoring reality thereby:

$Bel(agent, event|context, e) \wedge Expect(agent, \psi, event|context, e)$ holds in any case then.

Such deliberative ignorance appears to be irrational for intelligent agents, but is a common attitude of human agents and obviously somewhat functional for them. In any case, the identification of certain expectations with beliefs regardless of deviance might be reasonable for artificial agents in case the event belief is obtained from an unreliable source.

A less debatable use for such deliberative ignorance is to set the normativity greater zero in order to filter out (“flatten”) temporal and insignificant fluctuations of probabilities.

In all cases except from iv., we assume that the expectability is equal to the probability (in case the normativity is zero).

Note that even for the cases i.-iii. so far no assumptions have been made on how e has been obtained - an agent is basically free to hold any expectabilities she likes / is interested in from her subjective and possibly irrational viewpoint.

Definition 4. *The deviance Δ of an event regarding a certain expectation (or vice versa of an expectation regarding an event) is defined with*

$$\Delta(event|context) = e - p,$$

given that $Bel(agent, event|context, p)$ and $Expect(agent, \psi, event|context, e)$ holds.

We integrate deviance measures into \mathcal{L}^{probDL} using

$$M, \sigma, \nu, n \models \Delta(event|context, e-p) \text{ iff } \exists e, p, \psi : M, \sigma, \nu, n \models Bel(a, F|context, p) \wedge M, \sigma, \nu, n \models Expect(agent, event|context, \psi, e).$$

Sometimes we use $\Delta(event|context) = e - p$ as a syntactic variant.

A deviance can intuitively be seen as an indicator of the effort that would be required to make a normatively expected event happen, and as a measure for the compliance of the event-generating agent with the expectation, whereas the normativity is intuitively a kind of “stamina” of the intention (the strength of a self-commitment. Please remember in this regard, that we allow intentions also to be denoted as desired behavior of other agents).

Trivially, the deviance can be used to retrieve a probability p from an expectability.

There is also a conjunction with the *utilities* of events: If the normativity is larger zero, the utility for the agent to reach the specified probability is certainly larger zero also. The expectability *might* correspond to the utility of the event in this case (but this is to state a heuristic only, suggesting further research).

Observation 2:

Except from the case iv. above (belief despite ignorance of event occurrences)

$$Int(agent, \forall t_i, t \leq t_i \leq t + h : \Delta_{t+i}(event|context) = 0)$$

holds at time step t . At this, h is a possibly infinite intention horizon which determines how long the expectation is maintained, and Δ_{t+i} is defined analogously to $Expect_t$.

Finally, we want to further simply the semantics in case the probability of an intended event is irrelevant:

Observation 3:

$$(Expect(agent, \psi, event|context, e) \wedge Bel(agent, event|context, en), en), \\ \rightarrow Int(agent, event), \text{ if } en < e$$

4 Computational adaptation of expectations at runtime

The expectability of an event is a function of event probability and normativity, whereby the normativity can be interpreted as the "stubbornness" of the expectation, or, inversely, its flexibility. After the expectabilities and normativities of adaptive-normative expectations have been obtained from goals and intentions, they are exposed to reality, so to say. One driving force for the run time adaptation of such expectations is the active influencing of the domain of the expected events in order to enforce normative and adaptive-normative expectations, another is to let such expectations adopt to empirical expectations passively. The following shows how this can be done in dependance from the normativity. As important special cases, the following definition covers expectations with normativity zero and one also.

To this end, it is assumed that for an event $event|context$ corresponding to a certain EN node an initial expectation strength $\theta(event, context) = P_0(event|context)$ exists. We define thereby for convenience $Bel_t(a, F|context, d) \equiv Bel(a, F|context, d) \wedge now(< t >)$ and $P_t(a|context) = d \Leftrightarrow Bel_t(a, F|context, d)$, denoting a probability stated at time t (not the probability of an event happening at time t) (cf. Definition 2 for $Expect_t$). Given a normativity ψ_t and a probability $P_t(event|context)$ (e.g., in form of a belief) obtained empirically at time step t , the expectation strength at this time step can be calculated recursively as follows. This way to calculate $Expect_t$ is not obligatory, other ways to calculate adaptive-normative expectations could be reasonable too, depending from the concrete application.

Definition 5. With $E_t(agent, \psi_t, event|context) = e \Leftrightarrow Expect_t(agent, \psi_t, event|context, e)$, $E_t(agent, event|context, \psi_t) =$

$$\begin{cases} \theta(event, context) & \text{if } t < 1 \\ E'_{t+1}(agent, \psi_t, event|context) & \text{otherwise} \end{cases}$$

with $E'_t(agent, \psi_t, event|context) =$

$$\begin{cases} E'_{t-1}(agent, \psi_t, event|context) \\ \quad - \Delta'_{t-1}(event|context)(1 - \psi_t) \\ \quad \text{if } t > 0 \\ \theta(event, context) & \text{otherwise} \end{cases}$$

$\Delta'_t(event|context)$ is calculated as $E'_t(agent, \psi_t, event|context) - P_t(event|context)$ ⁴.

This (non-mandatory) way to calculate $Expect_t$ reminds of the econometrics technique of *Exponential Smoothing* used for the smoothing and extrapolation of non-linear

⁴ Calculating $Expect_t(\dots)$ using $Expect'_{t+1}(\dots)$ is done just in order to get rid of the delay of one time step in the adaptation of $Expect_t(\dots)$ to $P_t(\dots)$ that would exist otherwise.

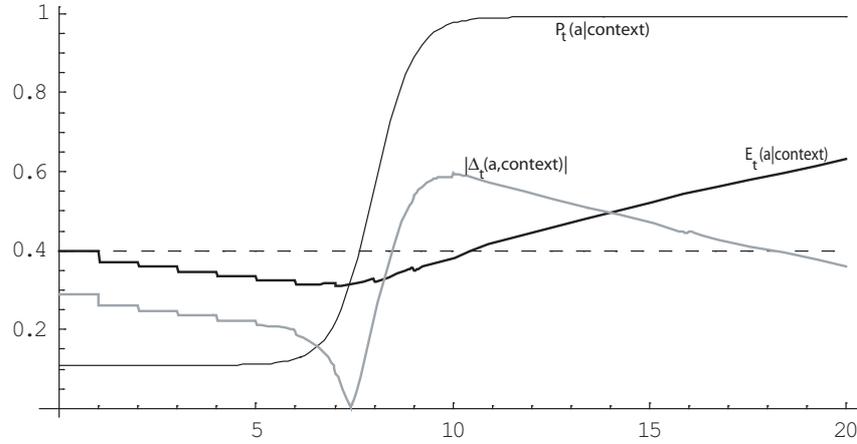


Fig. 1. Unattended adaptation of an expectability ($t \rightarrow$)

time series. It calculates a flattened version (with a flattening degree depending on the normativity) of the graph of $P_t(event|context)$, and lets $Expect_t(agent, \psi_t, event|context)$ converge to $P_t(event|context)$ at least if $P_t(event|context)$ remains constant with increasing t , and ψ_t remains constant also. The normativity (i.e., the expectation adaptation rate) itself does not change.

If, e.g., $\psi_t = 1$, the expectation strength e in $Expect_t(agent, \psi_t, event|context)$, $e = \theta(context, event)$ will remain constant, whatever the empirical evidence is. In contrast, if $\psi_t(agent, context, event) = 0$, $Expect_t(agent, \psi_t, event|context, e)$ with $e = P_t(event|context)$ applies at all time steps.

Example:

Figure 1 shows the time and normativity dependent expectabilities (abbreviated with E_t) of an event a , with $\psi_{0..20} = 0.95$ and $\theta(a, context) = 0.4$. Being a fictive event, the potential effect the announcement of these values to the event generator (a communication partner of the agent, for example) would have, is not considered. The *agent* parameter has been omitted.

5 Conclusion

Most traditional formalisms for normative systems clearly restrict the agents' autonomy without any concept of flexibility. In demarcation from such approaches, we aim at avoiding this by accepting autonomy even from social norms as a necessary characteristic of agency that must not be ruled out, and sometimes even can not be ruled out

at all, as it is typical for truly open multiagent systems. Based on Luhmann's theory of social systems and our previous works (Brauer et al. 2002), this is in the line of Castelfranchi's view: A socially oriented perspective of engineering order in agent systems is needed and most effective (Castelfranchi 2000). In addition to that, this sociological grounding also makes our approach different from approaches that apply sociological concepts and terminology in a comparatively superficial and more or less ad-hoc manner. Thus we hope that the introduction of adaptive-normative expectations opens a new perspective of multiagency and normative systems.

References

- Bacchus, F. (1990), *Representing and Reasoning with Probabilistic Knowledge*, MIT Press, Cambridge, Massachusetts.
- Bauer, B. & Müller, J. (2003), Using UML in the context of agent-oriented software engineering, in P. Giorgini, J. Müller & J. Odell, eds, 'Agent-oriented software engineering. Proceedings of the Fourth International Workshop (AOSE-2003)', Lecture Notes in Artificial Intelligence, Vol. 2935, Springer-Verlag, pp. 1–24.
- Bellifemine, F., Poggi, A. & Rimassa, G. (2000), Developing multi-agent systems with JADE, in C. Castelfranchi & Y. Lespérance, eds, 'Intelligent Agents VII, Proceedings of the Seventh International Workshop on Agent Theories, Architectures, and Languages (ATAL-2000)', Springer-Verlag.
- Boella, G., van der Torre, L. & Verhagen, H. (2007), Introduction to normative multiagent systems, in G. Boella, L. van der Torre & H. Verhagen, eds, 'Proceedings of the Dagstuhl Seminar on Normative Multiagent Systems 2007'.
- Bratman, M. (1987), *Intentions, Plans and Practical Reasoning*, Harvard University Press, Cambridge, MA.
- Brauer, W., Nickles, M., Rovatsos, M., Weiß, G. & Lorentzen, K. (2002), Expectation-oriented analysis and design, in M. Wooldridge, G. Weiß & P. Ciancarini, eds, 'Agent-oriented software engineering. Proceedings of the Second International Workshop (AOSE-2001)', Lecture Notes in Artificial Intelligence, Vol. 2222, Springer-Verlag, pp. 226–244.
- Castelfranchi, C. (2000), Engineering social order, in 'Working Notes of the First International Workshop on Engineering Societies in the Agents' World (ESAW-00)'.
- Castelfranchi, C. & Lorini, E. (2003), Cognitive anatomy and functions of expectations, in 'Proceedings of IJCAI'03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions'.
- Cohen, P. & Levesque, H. (1990), 'Intention is choice with commitment', *Artificial Intelligence* **42**, 213–261.
- Fischer, F. & Nickles, M. (2005), Towards a notion of generalised statement-level trust, Technical Report FKI-249-05, Institut für Informatik, Technical University of Munich.
- Herzig, A. & Longin, D. (2002), A logic of intention with cooperation principles and with assertive speech acts as communication primitives, in 'Proceedings of the first international joint conference on Autonomous agents and multiagent systems (AAMAS 2002)'.
- Lorentzen, K. & Nickles, M. (2002), Ordnung aus Chaos - Prolegomena zu einer Luhmann'schen Modellierung dezentropisierender Strukturbildung in Multiagentensystemen, in T. Kron, ed., 'Luhmann modelliert. Ansätze zur Simulation von Kommunikationssystemen', Leske & Budrich.
- Luhmann, N. (1995a), *Social systems*, Stanford University Press, Palo Alto, CA. Originally published in 1984.

- Luhmann, N. (1995*b*), *Social Systems*, Stanford University Press, Palo Alto, CA. (originally published in 1984).
- Ndumu, D., Collins, J., Owusu, G., Sullivan, M. & Lee, L. (1999), 'ZEUS: A toolkit for building distributed multi-agent systems', *Agentlink News* **2**, 6–9.
- Nickles, M., Fischer, F. & Weiß, G. (2005), Communication attitudes: A formal approach to ostensible intentions, and individual and group opinions, *in* W. van der Hoek & M. Wooldridge, eds, 'Proceedings of the Third International Workshop on Logic and Communication in Multiagent Systems (LCMAS 2005)'.
- Nickles, M., Rovatsos, M. & Weiß, G. (2005), 'Expectation-oriented modeling', *Engineering Applications of Artificial Intelligence (EAAI)* **18**.
- Sadek, M. (1992), A study in the logic of intention, *in* 'Proceedings of the third international conference on principles of knowledge representation and reasoning (KR'92)'.