Ersch. in: Journal of Applied Logic ; 6 (2008), 2. - S. 251-266

http://dx.doi.org/10.1016/j.jal.2007.06.011

Living up to one's commitments: Agency, strategies and trust

Thomas Müller

Universität Bonn, Institut für Philosophie, Lennéstr. 39, 53113 Bonn, Germany

Abstract

In human social interaction, the notions of commitment and trust are strongly interrelated. A formal model for this interrelation will enable artificial agents to benefit from the associated reduction of complexity and improved possibilities for cooperation. We argue that the notion of living up to one's commitments, rather than actual fulfillment, plays the key role in the trust–commitment interrelation, and we propose a formal analysis of that notion in the framework of agents and strategies in branching time. The main technical innovation is a stringency ordering on an agent's strategies that allows one to classify what an agent does as more or less appropriate with respect to a given commitment, leading to a fine-grained assessment of trustworthiness.

Keywords: Promises; Contracts; Stit logic; Strategy; Commitment; Trustworthiness

1. Introduction

Commitment and trust are of key importance for our human social lives [1,2]: without trust, we would be paralysed and could not dare to act; without commitments, only trivial cooperative schemes would be possible. As human agents, we normally handle trust, commitments and their interaction smoothly. It appears plausible to assume that artificial agents capable of trusting and entering into commitments have an advantage over others, and much recent technical work testifies to the importance of these concepts [3–7]. One could approach these two concepts by starting from a given formal basis, e.g., some calculus for updating probabilities of future possibilities and some framework of deontic logic, and try to express as much as possible in these terms. In this paper, however, we follow a different approach to modelling commitments and trust: We first describe the phenomenology of commitment and trust in human society in some detail before moving to a consideration of which formal tools may be used to capture these notions with a view towards application and implementation.

Thus, in Section 2 we give an overview of the human practices of commitments, trust, and their interaction. We will point out that the notion of *living up to one's commitments* plays a key role in assessing trustworthiness. In Section 3, we propose a formal background theory in which one may hope to spell out formal counterparts of these notions, viz., the theory of agents and strategies in branching time. In Section 4 we move towards a formal characterisation of the notion of living up to a commitment. Section 5 pulls together the results of the previous sections and proposes a model for the interaction of commitments and trust.

E-mail address: Thomas.Mueller@uni-bonn.de.

2. Commitment and trust in human social interaction

The aim of this section is to give an overview of some key structural features of commitment and trust that can be discerned by looking at human social interaction. At this stage, we are mainly concerned with a phenomenologically adequate description of commitment and trust as actually practiced, not with a formal framework in which these phenomena could be embedded.

The literature on these subjects is immense. Rather than aiming at a balanced overview, we give a selective outline by presenting each concept from a single, coherent point of view.

2.1. Commitment

"Commitment" as used here refers to a directed (or relational) type of normativity: one person is committed towards another person. Deontic logic is usually concerned with a different type of normativity which is not directed, singling out "forbidden" or "allowed" states or actions. For a very useful, non-formal description of that distinction cf. [8], where relational normativity is called "bipolar", while the non-relational kind is called "monadic".

In [8] it is also pointed out that the formal structure of relational vs. non-relational normativity can be found in various normative spheres. For human beings, some relevant normative spheres are etiquette, the legal rules in effect in various states at various times, rules of clubs and societies, rules of games and, of course, the ethical sphere. Commitments in various spheres may interact, but they may also be independent. E.g., positive law can create commitments to do things that are morally insignificant, or even morally wrong. For the purpose of this paper, we will focus on relational commitments in a single normative sphere only: in a first approach to modelling the commitments of artificial agents, it appears justified to make this simplifying assumption. Later on, if applications so demand, additional normative spheres may be introduced and their interaction (or lack of interaction) characterised formally. Thus, in this paper we will ignore some of the bewildering complexity of our human notion of commitment.

How should one specify a relational commitment? Which aspects are structurally important? A relational commitment, or *normative relation*, between agents α and β is a certain type of normative constraint on the behaviour of α. Minimally, the *content* of the relation specifies what counts as fulfilling or violating the norm (and what is neutral with respect to this distinction), and often the content also specifies a deadline for fulfilling the commitment. Typically, β is also in a position to put a *sanction* on α for non-compliance (perhaps with the aid of an overarching normative metasystem such as the state and its legislation). Apart from the sanction, which is external to the normative relation, many normative relations also have an internal (quasi-moral) normative dimension: quite apart from any sanction, many norm violations are judged to be "just wrong".

Commitments are dynamical objects: They are created at a certain time during the agents' existence, and they usually expire at some later time—normally when they are fulfilled, but also when they are successfully revoked or given up. Commitments can also be frustrated, which may give rise to new commitments (e.g., to make good for one's fault); commitments usually do not vanish without a trace when they are violated. There are many sources of new commitments. Commitments typically arise as a consequence of what somebody does. E.g., if John takes away something that belongs to Joan, he is committed to giving it back to her. While such commitments that arise as an (indirect) consequence of some action are common, there is also a more direct source of commitments, which is most relevant for trust and trustworthiness: people can enter freely into relational commitments by consent, e.g., through promises or contracts.

Intuitively, we distinguish more from less stringent commitments, for which considerations of urgency, need, importance and the personal relation between the agents play a role. Whether we are willing to accept a specific commitment from somebody, is often a question of trust.

2.2. Trust

As Luhmann [1] points out, trust can be seen as a means for managing and reducing complexity. Any system exposed to an open, potentially unlimited environment has to face the fact that the complexity of the external environment is too large for the system's own, necessarily limited internal capacity for data acquisition, storage, and processing. Nobody can face the full complexity of the world. But systems have to cope if they want to sustain themselves successfully [1, p. 4].

The world's inherent complexity is increased as social interaction grows. Trust can balance this growth of complexity, thereby allowing socially interacting systems to develop despite the limited capacities of individual agents. Thus trust is an important means of human social development [1, p. 7].

Luhmann distinguishes personal trust from trust in systems. Trust starts as personal trust of one agent towards another. System trust develops on the basis of personal trust and allows for ever higher forms of social organisation. Luhmann convincingly points out that, e.g., the complexity-reducing aspects of money are best described in terms of trust in the monetary system, not in individuals [1, Chap. 7]. For the purpose of this paper, however, we will consider personal trust only.

Trust is future-directed. It allows us to act as if the future—or some relevant aspect of the future—was fixed [1, p. 10]. Of course the future is open; the more so if what the future brings for us depends on what other agents, with whom we interact, decide to do. While trust is directed to the future, it is effective in the present; it influences our present actions and choices.

Trust is different from planning. This shows itself most clearly when one considers the effect that trust vs. planning has on complexity. The more we plan, the more complex the possible future scenarios become that we have to deal with—there is so much that could go wrong. On the other hand, the more we trust, the less complex the possible future scenarios appear—so much is considered fixed. But of course trusting does not actually *fix* the future—it just means that presently we act as if the future was fixed in some relevant respects [1, p. 20]. It may turn out, afterwards (when it is too late), that we were wrong in putting our trust where we did. If it turns out that we were wrong, *we ourselves* will have to compensate for the external failure and find internal means of coping with the sudden increase in complexity. Thus trust shifts complexity from the outside to the inside of a system, presupposing but not specifying internal resources for coping. This point is important: when we trust, we normally do not have a concrete backup plan for what to do if things go wrong. If we did, the scenario would be one of complexity-increasing planning rather than one of complexity-reducing trust. Actual coping mechanisms can vary—e.g., external sanctioning mechanisms may be available, or we may put up with a loss and censure ourselves for having trusted where trust was in fact inappropriate. There are cases in which we should know better than to trust, and sometimes even our social environment will not pity us, but blame us when things go wrong [1, Chap. 5].

When should we trust? Whom should we trust? We intuitively assess people (or rather, people in their various roles)[1] in terms of their trustworthiness. Many factors appear to play a role in this assessment. However, one point that Luhmann stresses is crucial: *If trust can really play its role of complexity reduction, trust cannot be based on an in-depth formal calculation*. Of course utility considerations play a role when we decide whether to trust or not, and probabilistic estimates may play a role too. But trust must not be identified with an estimate of subjective probabilities or utilities [1, p. 30n4]. If that were so, trust would not reduce complexity, but rather increase it. Trust is based on "overdrawn information"; it involves a leap—as already pointed out by Simmel [9]. But trust isn't a leap in the dark either. An important aspect of trust is the law of meeting again: it is easier to trust people whom we know we will meet again. Also, established social (e.g., juridical) means of sanctioning trust-breaking make it easier to trust.

Trust is mostly implicit in a social relationship. In fact, as Luhmann points out, it usually hurts a trusting relationship if questions of trust are openly discussed at an early stage. The stranger who tells us to trust him is met with suspicion. But once some level of trust has been established, or some security is provided by an established sanctioning mechanism, it becomes possible to make trust explicit without hurting the trusting relationship. Only this makes it possible to enter into commitments: the free creation of a commitment presupposes a reflection on trust, as becomes obvious through the synonymity of "I promise to do X" and "Trust me, I will do X".[2]

[1] For many questions of social interaction, it is not so much people as their roles that are important [1, p. 41]. Luhmann gives the example that we may trust somebody as our co-worker (i.e., base our professional work on his results) while we would not trust him enough personally to lend him a little money [1, p. 91].

[2] Witness Grice's cancellation test: Is it possible to say, "I promise to do it, but don't trust me that I will", or "You can fully trust me that I will do it, but I don't promise"? Only a few special cases can give any plausibility to these expressions. To promise and yet not to invite trust might signal a person's knowledge that she is weak-willed, and thus not fully rational. To invite trust and yet not to promise might be a sign that one is forbidden to promise on religious grounds, as historically in Quakerism.

2.3. On the interaction of commitments and trust

In the interrelation between commitments and trust, both directions of influence play an important role. As everybody agrees, the way people treat their commitments influences our trusting relationships with them: broken commitments tend to have a negative influence on trust,[3] while promises kept and contracts fulfilled normally increase our level of trust with the other party. In the other direction, we just pointed out that freely created commitments presuppose an explicit reflection on trust and thus, an already existing trusting relationship. While this aspect of the trust–commitment interrelation is practically as important as the first, it is less discussed. For both aspects, a close look at actual human practices reveals important details that will later on be useful for formal modelling.

Trust plays a role in all commitments that we create by consent. By accepting a promise, we affirm that we trust the promisor to do what she promised.[4] As human beings are free agents, the question of whether they will in fact fulfill their commitments is necessarily one that we cannot answer with certainty. This is a conceptual point: it makes no sense to promise something that will happen anyway. Thus in accepting a promise and relying on its fulfillment, we act on overdrawn information: we trust the promiser.

Whether we will in fact accept a promise in a given situation, depends on whether the promiser appears trustworthy enough. Phenomenologically it is clear that we make rather fine distinctions in this respect. Trustworthiness is not an all-or-nothing affair. We may trust somebody in some respects and not in others (cf. Footnote 1), and both the importance and the urgency of the commitment we seek play a crucial role. Importance and urgency pull in different directions: The more important something is, the more careful we tend to be in selecting whom to trust, but the more urgent something is, the less selective we can afford to be. Thus whether we accept a commitment offered to us depends both on the nature and content of the commitment and on the trustworthiness of the one who offers the commitment to us.

In the other direction, when we consider the way in which we adjust our assessment of trustworthiness after a person has undertaken a commitment to us, interesting facts emerge as well. At first sight it might appear that all that counts for trustworthiness is whether people actually fulfill their commitments, where fulfillment is an extensional concept.[5] This may be adequate for certain closed systems, but it is not the way we assess trustworthiness in social interaction (as noted also by [4]). We know that all agents act under unavoidable (and unquantifiable) uncertainty, and we make fine distinctions when it comes to excusing others or attributing blame. Impossibility of fulfillment for external reasons such as an accident is often a good excuse. But there are also cases where we excuse people for breaking their commitments to us even though it was easily possible for them to fulfill them. In extreme cases, we might not just excuse, but praise such behaviour: a frustrated commitment can *increase* our trust in a person, if it was frustrated for the right kind of reasons.[6] On the other hand, an extensionally satisfied commitment may cause us to lose our trust in a person, if the commitment was satisfied for the wrong kind of reasons.[7] Rather than looking at extensional satisfaction or frustration of commitments, the important dimension in assessing trustworthiness seems to be whether people are *living up to their commitments*. Living up to a commitment may mean to break it, given exceptional circumstances, and extensionally satisfying a commitment may not qualify as living up to it. In the rest of this paper we are after a formal framework in which we can fruitfully represent that important notion.

At this point it may be in order to come back to a methodological point made in the introduction. Of course one can approach cases like the ones just mentioned with the conviction that it must be possible to capture all that is truly rational about them, in some framework of game theory, and dismiss all aspects that elude this pre-assigned

[3] Hume notoriously goes so far as to claim that one who makes a promise to anybody "subjects himself to the penalty of never being trusted again in case of failure" [10, p. 522].

[4] Of course, promises can also be accepted tactically—Raz [11, p. 213] gives the example of "a man who solicits a promise, hoping and believing that it will be broken, in order to prove to a certain lady how unreliable the promiser is". But this is a special case that can only function against a background of a non-tactical, basic practice.

[5] Thus people may fulfill a commitment without knowing that they do—especially if they have additional reasons for doing what in fact amounts to fulfilling the commitment.

[6] A case in point is the classical example of the person who promised to come for tea, but rather saved a drowning child on the way. Broken commitment, but what a good guy. In fact, *we* would be to blame morally if we made a fuss about our tea [12, p. 16].—It appears plausible to assume that the "right kind of reasons" here singles out actions that we would have done ourselves, or wish we would do ourselves.

[7] Consider John, who promised to meet Joan at the station at midnight, and does, but only because, even though he forgot all about her, that was the only place to get some beer at that time.

formal framework as superstition. We are far from assuming that there are no human practices that show elements of superstition that we should get rid of. However, we wish to capture the phenomena first, before moving on to the best formal description we can give. The unavoidable rest, the mismatch between formal model and actual behaviour, may then be classified as human superstition or formal inadequacy, as the case may be.

3. Towards a formal theory

The discussion of the previous section shows that a central notion in the dynamics of commitments and trust is that of living up to a commitment. If artificial agents are to benefit from a faithful analogue of the human practices of trust and commitment by consent, we need to find a way of characterising formally what it means for such an agent to live up to a commitment. We thus need a theoretical framework in which agents, their actions, and the contents of commitments can be represented. There appear to be at least two main candidate theories: the event calculus [13,14] and the theory of agency in branching time [15] that is used, e.g., for the stit ("seeing to it that") modal logic of agency. In view of the overtly indeterministic outlook of human agents, we opt for the latter, which has an explicit notion of agents and their choices in an indeterministic world. This should however not be taken to insinuate that we believe an event calculus approach would be pointless; rather, such an approach should be developed as well to learn more about the strengths and weaknesses of both frameworks. In the spirit of Belnap [16], one should let many flowers bloom.

Apart from the explicit indeterminism of branching time, we can profit from the fact that there is already some work on commitments in that framework. In fact, we will take a lead from the formal characterisation of commitments (promises vs. word-givings) given by Belnap et al. [15] (who acknowledge earlier, related work by Thomson [17]). Belnap specifies the content of a commitment in terms of an agent's strategy, employing the theory of agency and strategies in branching time.

The next two subsections introduce that framework. Section 3.1 gives a quick overview of branching time and agents' choices, and Section 3.2 addresses the formalities of strategies.

3.1. Background: Agents and choices in branching time

In this section we describe the theory of agency that forms the background of our approach. That theory is motivated and explained both formally and informally in Belnap et al.'s [15], which may be consulted for additional details and for its many examples.[8] (In what follows, the notation of [15] has been slightly altered in a number of places.)

Branching time is a theory of objective indeterminism that was first invented by Prior [21] and that has subsequently found many applications in philosophy [15,20], logic, and in computer science, where CTL ("computational tree logic") and its generalisations such as ATL ("alternating-time temporal logic") have found widespread use [22,23]. A branching time structure is a tree-like partial ordering of moments without backward branching:

Definition 1. A *branching time structure* is a partial ordering $\langle W, <\rangle$ such that there is no backward branching, i.e., if $m' < m$ and $m'' < m$, then $m' \leqslant m''$ or $m'' \leqslant m'$.

Note that we do not postulate a discrete ordering at this stage—the theory to follow can and will be developed in a more general setting. The elements of W are called *moments*, and $m < m'$ is read temporally, i.e., as "m is before m'". A *history* h is a maximal linear chain (set of pairwise comparable elements) in W, and for each moment m, the set H_m is the set of histories to which m belongs.

In view of the fact that we will later be dealing with commitments with deadlines, we need a way to talk about the time of these deadlines, i.e., clock times, in different histories. The easiest way to do this is to take clock times to be real numbers (think of some decimal coding), and to demand that each history be isomorphic to the set of clock times,

[8] In accord with Belnap [18,19], we are convinced that a really successful theory of agency must not stop at the level of resolution supplied by branching *time*. Rather, *branching space–times* models [20] will have to be employed. As the theory of agency in branching space–times is little developed so far, in this paper we will hold on to the technically well understood theory of agency in branching time.

or, to use Belnap's terminology, *instants*. Then a function $i_{(m)}$ can give the instant (clock time) associated with each moment m. Formally, we use the following definition:[9]

Definition 2. A structure of *branching time with instants* is a triple $\langle W, <, I\rangle$ where $\langle W, <\rangle$ is a branching time structure and I is a partition of W into instants i such that

- the intersection of each history h with each equivalence class (instant) i from I has exactly one member, which we write $m_{(i,h)} = i \cap h$,
- I preserves the ordering $<$, i.e., for any two instants i, i' and histories h, h', if $m_{(i,h)} < m_{(i',h)}$, then also $m_{(i,h')} < m_{(i',h')}$,
- the induced ordering on I ($i < i'$ iff $m_{(i,h)} < m_{(i',h)}$, where by the previous point, any history h will give the same verdict) is isomorphic to the reals, so that we will identify instants i with real numbers. We write $i_{(m)}$ for the (unique) instant to which the moment m belongs.

In a branching time structure (with or without instants), two histories $h_1, h_2 \in H_m$ can either split at m or be undivided at m:

Definition 3. Two histories $h_1, h_2 \in H_m$ are called *undivided at m* ($h_1 \equiv_m h_2$) iff they share a moment that is properly later than m, i.e., iff there is $m' \in h_1 \cap h_2$ s.t. $m < m'$. If, on the other hand, m is maximal in $h_1 \cap h_2$, we say that h_1 *splits off from* h_2 *at m* ($h_1 \perp_m h_2$).

The notation is suggestive: $\equiv_m$ is easily shown to be an equivalence relation, inducing a natural partition Π_m of H_m. Assuming $h \in H_m$, we write $\Pi_m\langle h\rangle$ for that unique element of Π_m that contains h. The Π_m are the metaphysical basis for all kinds of indeterminism in a branching time model, whether agency-related or not.

Given a set $A = \{\alpha_1, \ldots, \alpha_n\}$ of agents, the theory of agents and choices in branching time specifies, for each agent $\alpha \in A$ and each moment m, the set of choices open to α at that moment. That set, $Choice^{\alpha}_m$, partitions H_m and may be more coarse-grained, but not more fine-grained, than the natural partition Π_m at m. Metaphorically, an agent has at most as much, but possibly less, control over what can happen at a given moment as nature herself. Thus agency is seen as a way of exploiting the natural indeterminism inherent in our world, not as a supernatural addition.—Assuming $h \in H_m$, we write $Choice^{\alpha}_m(h)$ for that unique member of $Choice^{\alpha}_m$ to which h belongs.

Our main structural definition is the following:

Definition 4. A structure of *agents with choices in branching time with instants* is a quintuple $\langle W, <, I, A, Choice\rangle$, where $\langle W, <, I\rangle$ is a branching time structure with instants, A is a finite set of agents, and $Choice$ is a function that assigns a partition $Choice^{\alpha}_m$ of H_m to each pair $\langle \alpha, m\rangle$ in such a way that $Choice^{\alpha}_m$ is a coarse-graining of the natural partition Π_m, i.e.,

$$\text{for all } \alpha, m \text{ and } h \in H_m, \quad \Pi_m\langle h\rangle \subseteq Choice^{\alpha}_m(h).$$

It is often natural to assume that agents can choose what to do independently from each other. This demand can be spelled out as follows:

Postulate 1 *(Independence of agents). For all moments m and all functions f_m from the set of agents A such that $f_m(\alpha) \in Choice^{\alpha}_m$, we have*

$$\bigcap_{\alpha \in A} f_m(\alpha) \neq \emptyset.$$

For the preliminary purposes of this paper, we will be concerned with a single agent, so that Postulate 1 will play no role. It will however be important in a true multi-agent framework.

[9] This is similar to Belnap's definition [15, p. 194f]. The requirement of isomorphism with the reals, which simplifies some considerations below, is not present in Belnap's formulation.

We will only say a little about the formal language that is appropriate for the theory of agents and choices in branching time (cf. [15] for details). Formulae are built up from propositional variables by means of the truth-functional sentential connectives and modal operators for the tenses ("*was*" and "*will*"). The clock times are used for a modality "true at a clock time", "At_t". Modal operators for agency ("α *stit* : ϕ" for "agent α sees to it that ϕ") can also be defined, but these operators will not be used in our framework here, because we will be working with the more general concept of a strategy.

In accord with the ideas of Prior-Thomason semantics [24], formulae are evaluated not just at moments, but at moment-history pairs $\langle m, h\rangle$, where $m \in h$. Future-tense propositions are typically history-dependent, i.e., at moment m, the truth value of such a proposition also depends on the history of evaluation h. For the purposes of this paper, the most important concept is the modality of *settled truth*, which corresponds to universal quantification over histories: ϕ is settled true at $\langle m, h\rangle$ iff ϕ is true at $\langle m, h'\rangle$ for all $h' \in H_m$. Settled truth is thus independent of the history of evaluation.

The respective inductive semantic clauses for the mentioned modal operators are:

- $m, h \models was : \phi$ iff there is $m' < m$ s.t. $m', h \models \phi$ (note that $m' \in h$ follows by backwards linearity),
- $m, h \models will : \phi$ iff there is $m' \in h$ s.t. $m < m'$ and $m', h \models \phi$,
- $m, h \models At_t : \phi$ iff $m_{(t,h)}, h \models \phi$, i.e., iff ϕ is true at time t on history h,[10]
- $m, h \models settled : \phi$ iff for all $h' \in H_m$ we have $m, h' \models \phi$, and
- $m, h \models \alpha$ *stit* : ϕ iff (i) for all $h' \in Choice^{\alpha}_{m}(h)$, $m, h' \models \phi$ and (ii) there is some $h'' \in H_m$ (a "counter") for which $m, h'' \not\models \phi$.

The tense operators and the "settled" modality are the standard "Occamist" operators of Prior-Thomason branching time. The ("deliberative") stit operator is defined via two clauses, the first positive, the second negative: To say that at m, on history h (that is, by choosing $Choice^{\alpha}_{m}(h)$), α sees to it that ϕ, means that (i) by that choice, α really makes ϕ true and that (ii) ϕ is not guaranteed to be true anyway. This second clause answers to the intuition that it would be unnatural to say of someone that she sees to it that $2 + 2 = 4$. Our concept of (human) agency does not include such cases of free-riding.

There are other alternative stit semantics; most notably, the "achievement" stit defined by Belnap and Perloff and a simplified stit first defined by Chellas [25] and used, e.g., by Horty [26, p. 14]. The book [15] gives a comprehensive overview. One problem that all stit approaches have to face is that the framework seems unable to handle cases of temporally extended agency. The deliberative stit operator only considers a single choice at one point in time, while most actions involve a large number of choices at consecutive times. The most promising way out appears to be to introduce strategies: it seems possible to extend the stit framework so as to make it applicable to cases of temporally extended agency, if one employs the notion of a strategy in addition to the stit framework [27,28].

3.2. Background: Strategies

A strategy specifies defaults for an agent's future choices by fixing which choices of the agent count as in accord with the strategy. Strategies are useful for describing agency since most of what we actually do takes time. Thus, consider baking a cake, which takes about an hour. You can't at the beginning of that one hour period, m_0, act in such a way, or see to it, that no further choice of yours is required for finishing the cake. At any moment in the process, you have the choice to toss everything and leave. Nor can you make all the required future choices at the initial moment, m_0. It is a conceptual necessity that a choice can only be made once it is due—otherwise it wouldn't *be* a choice. This is not to say that an agent at m_0 is completely helpless, however. The agent can adopt a strategy, s, that prescribes default choices for future moments from m_0 on. It seems that temporally extended agency can be best described via strategies—cf. [27] for an attempt at spelling this out in a type of modal logic, and [28] for a similar approach that focuses on the multi-agent case and also addresses the question of uniform strategies.

Having available a means for describing temporally extended agency appears to be a prerequisite for a successful description of the content of a commitment. Almost all commitments pertain to a whole array of future choices—e.g.,

[10] Here we rely on a sloppy identification of the reals, terms standing for the reals, and the set of instants I which is isomorphic to the reals. We trust that no confusion will result. The proper distinctions for the general case are made in [15, pp. 194f].

hardly any promise can be fulfilled instantaneously. The whole point of commitments is to have a token now for something that cannot itself be had now: an agent's future choice(s).

The formal theory of strategies is laid out in detail in [15, Chap. 13]. We only give the main definitions:

Definition 5. A *strategy for* α is a partial function s on moments such that for each moment m for which s is defined, $s(m)$ is a subset of H_m that respects the available choices of α at m, i.e., for every $h \in H_m$, if $h \in s(m)$, then $Choice^{\alpha}_{m}(h) \subseteq s(m)$.

A strategy thus specifies what α should do, and in this, the strategy can give advice at most as fine-grained as the available choices for α at m allow (which, in turn, may be at most as fine-grained as the natural partition Π_m allows).

If an agent follows a strategy, some histories will be admitted (the strategy will advise to stay on the history), and others will be excluded. Technically, the definition is:

Definition 6. If s is a strategy for α, we say that s *admits* h iff for every m for which s is defined, if $m \in h$, then $h \in s(m)$. We say that s *excludes* h iff it does not admit it. The set of histories admitted by s is defined to be

$$Admh(s) = \{h \mid s \text{ admits } h\}.$$

The admitted histories are those histories that can happen, given that the strategy is followed. This however includes histories passing through moments at which s is not defined. In order to exclude these, the following restriction is useful:

Definition 7. A strategy s *guarantees* a set of histories H_0 iff

$$\big(Admh(s) \cap H_{\langle Dom(s)\rangle}\big) \subseteq H_0,$$

where $H_{\langle Dom(s)\rangle}$ is the set of histories passing through a moment at which s is defined.

By our definition, strategies are allowed to leave open choices for α—their advice does not have to be as fine-grained as possible. A strategy that gives the most detailed kind of advice possible, is called *strict*:

Definition 8. A strategy s for α is called *strict at* m iff it is defined at m and $s(m) \in Choice^{\alpha}_{m}$. The strategy s is called *strict* iff it is strict at every moment at which it is defined.

Strict strategies enjoy a special epistemic status: if an agent α is following a strategy that is strict at a moment m, then her actual choice at that moment is uniquely determined by the strategy. Thus, if we know that an agent is following a strict strategy, we can read off that strategy from her actual choices: in this case, behaviour is a perfectly reliable guide to the agent's strategy. This is not so if the agent is following a non-strict strategy:[11] at a moment m at which $s(m) \notin Choice^{\alpha}_{m}$, the actual choice that α makes (which corresponds to a single element of $Choice^{\alpha}_{m}$) does not reveal $s(m)$ completely (as $s(m)$ corresponds to two or more incompatible elements of $Choice^{\alpha}_{m}$). In such a case, we would have to ask α to tell us which strategy was on her mind, as behavioural clues will be of little help. However, if our task is to find out about α's strategy in order to assess her trustworthiness, asking α may appear to be circular: in order to rely on α's answer, we would have to know whether we can trust her, but that is exactly what we wish to find out. Studying α's behaviour, on the other hand, does not presuppose trusting α.

At this juncture, we face a methodological decision regarding our formal approach to the interaction of commitments and trust. It is clear that we can only ever identify an agent's strategy in full detail if that strategy is known

[11] Things are even worse if an agent *has* a strategy at m, but deviates from it. Metaphysically this is always possible, since adopting a strategy cannot mean making choices before they are due. In a case in which an agent deviates from her strategy, her actual choice reveals nothing about that strategy. However, in our setting, in which we identify strategies only *ex post facto*, the important thing is what the agent actually chose to do, and we can disregard the possibility of having deviated from a strategy as epistemically empty.—The fact that we identify strategies *ex post facto* also allows us to ignore the possibility that an agent's strategy may be undefined for some moments: looking back at a moment, by no backward branching it is always clear that *exactly one* of the available choices was made.

to be strict, and we need to know about the agent's strategy in order to assess trustworthiness (as spelled out below). May we assume that agents always follow strict strategies? It is not an easy task to weigh the pros and cons of that assumption. Assuming strict strategies allows for a smooth formal picture. Furthermore, the point can be made that action is after all our most fundamental guide to intention, or strategy [29, §4]. On the other hand, we can never be sure that an agent is really following a strict strategy, and presupposing strict strategies from the outset may appear to impose an unnecessary restriction on our formal theory. In what follows, we will take the case of a strict strategy to be basic, giving some indications of how to generalise for the case of non-strict strategies.

4. Commitments, strategies and stringency

Having laid out the formal background for our theory, we now turn again to human practices concerning commitments. Among these, promises and contracts are the most significant examples. We suggest that taking a close look at these practices opens up the view towards an important distinction that can be made formally precise within our framework: commitments come with various degrees of stringency. That distinction will be crucial for spelling out our target notion of living up to a commitment.

4.1. Promises vs. contracts

Is there a difference between a promise and a contract? The terminology is certainly not established so firmly as to allow for a straight yes. However, we wish to suggest that a typical promise and a typical contract are different creatures of the normative realm, and that it is useful for our purposes to distinguish the two notions sharply. In effect we will be arguing that there is a continuum of normative relations, one pole of which can be exemplified by a certain type of promise, while the other pole corresponds to a certain type of contract.

Promises are normally made between people who know and trust each other at least to some extent, and in most cases, there are no enforceable external sanctions connected with promise breaking. (However, the trusting relation between promiser α and promiser β is typically altered by a broken promise, which may count as a kind of internal sanction in some cases, and promises may give rise to *additional* legal relations that are associated with sanctions.) Promises normally concern matters of little economical importance, but of some personal, practical importance to β. They usually have a clear moral dimension, since for most promises it is judged to be morally wrong to break them.

Contracts can be made between any two persons, but certain very close relationships may make contracts appear to be inappropriate, and some legal systems do not recognise certain types of contracts between, e.g., husband and wife. Most contracts come with enforceable external sanctions backed by the law. Contracts usually concern matters of some economical importance, while the personal importance of the content of a contract may be wholly derived from its economical aspects. Not all contracts need to have a moral dimension—there seem to be contracts that it may be costly, but not morally wrong, to break.

Based on these observations, we propose to assume a continuum of normative relations characterised by the degree of stringency (pure moral bindingness) vs. external sanction involved, as follows:

(1) The most solemn and most stringent case is a case of promising without any associated external sanction. In the philosophical literature, the typical example of this is a death-bed promise given without witnesses. If α promises something to dying β and there are no witnesses, there is no possibility of an external sanction. However, such promises seem to have great moral force—maybe just because of the sheer helplessness of β, who himself knows that there is no possibility of enforcing what is promised.
(2) In a typical case of promising, there is no or little external sanction possible, but a relatively large degree of moral force. If α promises β to meet her at the station and fails to do so (for no good reason), then there is usually little that β can *do*, but α's behaviour will be judged morally wrong. Whether it is good to fulfill a promise is mostly a matter of morality and only to a little extent a matter of α's personal utility.
(3) In a typical contract, there is a relatively large degree of external (legal) sanctioning possible, but (usually) little moral force. E.g., if α enters a contract with β to the effect that α deliver some goods at a specified time, but fails to deliver, the consequences are mostly external: α may be sued and have to pay for β's damages. This is not to say that such behaviour on α's side will be assumed to be morally neutral—there will usually be some negative

assessment in this dimension too. Whether it is good to fulfill a contract is to some degree a matter of morality, but mostly a matter of α's personal utility.

(4) Finally, there seem to be contracts the fulfillment or non-fulfillment of which is purely a matter of utility. In the extreme case, a freely levelled contract between α and β may give α a choice to either comply or opt out and pay a fine, so that payment of the fine is not even clearly a sanction (it might be viewed as an alternative way of fulfilling the contract). But there are also cases in which opting for what clearly is a sanction may be just a matter of utility. E.g., for most bills, there will be late charges if one does not pay in time, and these charges are clearly sanctions (contracts typically simply demand payment in time). But still, if the late charges are less than overdraft charges that my bank will take if I pay the bill before my paycheck comes, it may be just a matter of utility, and not a bad thing, to pay the bill later.

When α and β enter into a normative relation, there is normally some kind of agreement as to which level of stringency that relation should be assumed to have—and if there is not, such agreement can usually be reached by discussing the issue. (If such agreement cannot be reached, the agents may choose not to create the commitment.) Cases in which β is completely helpless (incapable of sanctioning) will tend to be more stringent morally, while free-market type agreements tend to have a low level of moral stringency.

The point of these observations about human practices is twofold. First, by showing that there are different types of normative relations in human social interaction, we wish to suggest that it may be beneficial to employ a correspondingly rich notion of normative relations in artificial normative systems, too. Certainly, distinctions of stringency of norms can be made for artificial normative systems—e.g., business transactions over the internet may be typical contracts, while giving root privileges to an agent may suggest a promissory commitment not to exploit that power in a harmful way. Secondly, the description given above already points towards a formal characterisation of stringency for strategies. We will develop that characterisation in the rest of this section. In Section 5, we will then employ our formal framework in order to address the question of trustworthiness of an agent. Roughly, the idea will be that an agent is trustworthy (because living up to her commitment) if her strategy is appropriate with respect to the degree of strictness of the commitment she has entered.

4.2. On the stringency of strategies

In employing strategies to analyse commitments, we take a lead from Belnap's analysis of promising, which in turn is based on the theory of strategies and agents and choices in branching time. In [15, Chap. 5C], the analytical target for promising is

$$\text{at } m_0, \alpha \text{ promises } \beta \text{ that } p,$$

where α and β are agents, m_0 is the moment of promising, and p is the content of the promise, which is typically a future-tense proposition that has something to do with an action, or actions, of α's. The difficult question of what type of proposition p should be exactly, is discussed at length in [15]. As Belnap et al. show, a stit analysis is bound to fail in all but the simplest cases, so that in their approach, promising is spelled out in terms of a strategy that α adopts at m_0. The content p is then taken to be a future-tense proposition that is to be evaluated at the moment m_0 at which the promise is made, along the lines of the idea of "double time references" propounded also, e.g., in [30]: At a moment $m_1 > m_0$, p is evaluated at m_0, but with reference to only those histories that also pass through m_1 (note that by no backward branching, $H_{m_1} \subseteq H_{m_0}$). One can thus define *settled truth at* m_0 *relative to* (later) m_1, as spelled out formally in [15, p. 122]. The main idea of promise-keeping, on that account, is that from m_0 onwards, the promise-keeping strategy advises α to choose so as to make p settled true (in the double time references sense just explained) if possible, and to keep p an open possibility otherwise.

In the context of his theory of promising, Belnap distinguishes two kinds of commitments, which he calls *word-giving* vs. *promising* (taking up a distinction from [17]). Word-giving he takes to be a less stringent commitment, which is expressed by the fact that α's strategy for word-giving does not advise α to do anything until the commitment is either fulfilled or violated.[12] In the latter case, the strategy advises α to compensate β, and that is all there is to it.

[12] Belnap et al. [15, p. 126f] suggest to choose more neutral terminology. According to their (non-standard) usage, promises are "satisfied" or "infringed", and mere word-givings are "vindicated" or "impugned". We will use the standard terminology of fulfillment vs. violation, but we wish to stress that the moral implications that these words normally suggest may be absent in the case of some commitments.

We wish to suggest that Belnap's analysis points in the right direction, but that it can be improved. First of all, it appears to us that the notion of a sanction or compensation that kicks in when a commitment has been violated, should be analysed differently. Secondly, rather than distinguishing just two types of strategies, we will be able to order strategies with respect to their stringency, thus allowing for a more fine-grained assessment of the adequacy of the strategy that an agent is following, relative to a given commitment. We will also simplify the content p somewhat by making use of the notion of instants to specify deadlines for fulfillment.

4.3. Commitments and sanctions

Commitments are normative relations between agents. Norms can be fulfilled or violated. There are two basic schemes that can be used to monitor and perhaps force compliance to norms. *Hard constraints* are such that it becomes (physically) impossible to violate the norm. E.g., in many parking garages you cannot leave without having paid—the gate just won't open. The norm to pay for parking is therefore monitored and enforced by a hard constraint. Quite another scheme is in effect in so-called *soft constraints*: Here, compliance to the norm is monitored, perhaps on a statistical basis, and non-compliance leads to sanctions if detected. E.g., in many schemes of paid on-street parking, you *can* park and retrieve your car without a valid ticket, but there will be a penalty if you are found not to have a ticket on display.

Most issues of our social lives are regulated via soft constraints. It is altogether impractical to try to enforce too many norms via hard constraints. E.g., how would you try to implement the rule not to steal via a hard constraint? This may be feasible in a few select circumstances (e.g., at vending machines), but generally, our society relies heavily on soft constraints. Indeed it seems difficult to imagine any society of human beings that would rely solely on hard constraints.

In the realm of artificial agents, hard constraints are often somewhat easier to impose than in the society of human beings—e.g., communication ports can be blocked relatively easily. However, even in a society of artificial agents, the more complex systems become, the less feasible does it become to rely on hard constraints only. Usually, hard constraints must be implemented and enforced centrally, and there are computational and protocol overheads as well as security issues speaking against relying on hard constraints exclusively.

Commitments are conceptually tied to soft constraints: If an agent is under a commitment, she is normally free to fulfill or violate the commitment (of course, influences outside the agent's control can have an impact on these possibilities). Once a commitment is violated, depending on the type of commitment, some sanction is appropriate. The question we now wish to address is how to model this sanction.

In human interactions, first of all, not all violations of commitments are detected. Secondly, the sanctions imposed upon (detected) norm violation are usually again enforced via soft constraints. E.g., if you fail to pay one of your bills, there will be a penalty, but it will again be up to you to pay the penalty or not. The situation can however escalate by taking a number of turns at court, and in the end, if you fail to comply persistently, you might be put in prison, which means that a truly hard constraint would be triggered. In our daily lives, we thus mostly operate with soft constraints, but legally binding norms are in the end backed by hard constraints.

In our formal model, we will employ an idealisation: We will assume that commitments are subject to soft constraints, but that upon violating a commitment, detection is certain, and a hard constraint kicks in to make sure the sanction has its desired effect. Thus, we will assume that sanctions are automatic.[13] One effect of this is that an agent must take the cost of sanctions into her utility considerations from the outset. This is how mere considerations of personal utility can lead to compliance with norms. We wish to suggest that in some cases, this will be judged good enough, whereas in other cases, it will not—it depends on the stringency of the commitment in question. This will allow us to capture an important aspect of our human practices that we pointed to in Section 4.1.

[13] This assumption is closely related to Anderson's [31] early proposal of a "reduction of deontic logic to alethic modal logic", according to which the deontic-logical "it is obligatory that p" is reduced to the alethic "necessarily, if non-p, then SANCTION". The main difference is that we do not suggest an overall reduction of anything to anything else, and that any sanction is relative to the commitment to which it is attached, whereas Anderson seems to have thought of a single, all-purpose sanction.

4.4. Ordering strategies by stringency

Having discussed the question of how sanctions take their effect, we can now address the question of when it is appropriate to call one strategy more stringent than another.

The stringency of a strategy is always relative to a given commitment. We specify a commitment c as a tuple $c = \langle m_0, \alpha, \beta, t, \psi, Z\rangle$, where m_0 is the moment at which the commitment is created, α is the agent who is committed, β is the addressee of the commitment, and Z is the sanction attached to non-fulfillment, which might be given as a negative utility or in some other form. The content of the commitment we specify via a later time (deadline) t (a real number s.t. $i_{(m_0)} < t$) and a non-modal proposition ψ (i.e., a proposition that can be expressed in pure predicate logic); the idea is that the content of a commitment should specify some state of affairs ψ holding at a time t. Laying out things that way is less elegant and less general than Belnap's employment of the idea of double time references— not all commitments have a content of this kind (even though many do). We opt for this type of specification in order to give a simple picture of what we are after.

In our approach, a commitment is monitored via soft constraints. Thus, what α *can do* is not directly influenced by the commitment she has entered—unless the commitment is violated, in which case the hard constraint enforcing the sanction Z kicks in. However, what α *does* will be more or less *appropriate* relative to the commitment c. The "good" histories with respect to the commitment are the ψ-at-t-histories passing through m_0:

$$H(m_0, \psi, t) = \{h \in H_{m_0} \mid m_{(t,h)}, h \models \psi\}.$$

An agent's attitude towards a commitment may be expressed in terms of the agent's choices in relation to the set of histories $H(m_0, \psi, t)$.

As mentioned, Belnap proposes to distinguish promising, which requires a strategy that actively aims at fulfilling the commitment if that is at all possible, from mere word-giving, which only requires a strategy that takes care of accepting the sanction—which in our case, in which sanctions are automatic, does not require any specific choices by the agent at all. These two strategies can be seen to indicate two poles in an ordering of strategies with respect to their stringency.[14]

After the moment m_0 at which the commitment c in question is created, it makes sense to qualify the choices available to α at a moment m with respect to the effect that they have on c. Obviously, after the deadline, it is too late:[15] what α chooses to do can no longer matter for c. Thus we only need to consider moments m for which $m_0 \leqslant m$ and $i_{(m)} < t$. Depending on the circumstances, α might have a number of choices at such m. As α's choices correspond to sets of histories, we let $h \in H_m$ and $C = Choice^{\alpha}_{m}(h)$. We can then classify the choice C as follows:

(1) Via C, α might make $At_t : \psi$ settled true (i.e., for all $h' \in C$, $m, h' \models At_t : \psi$, i.e., $C \subseteq H(m_0, \psi, t)$);
(2) there might be available at m a strategy s for making $At_t : \psi$ settled true (a strategy s that guarantees $H(m_0, \psi, t)$ in the sense of Definition 7), and C might be the choice required by that strategy at m ($C = s(m)$);
(3) there might be some $h' \in C$ such that for some $m' > m$ with $m' \in h'$, there will be available a strategy guaranteeing $H(m_0, \psi, t)$ from m' on;
(4) there might be some $h' \in C$ such that $m, h' \models At_t : \psi$ (i.e., $C \cap H(m_0, \psi, t) \neq \emptyset$);
(5) via C, α might make $At_t : \psi$ settled false (i.e., $C \cap H(m_0, \psi, t) = \emptyset$).

This classification is exhaustive: (4) and (5) are exhaustive already and exclude each other, being contradictories. (1)–(3) imply (4), so that the classification is not exclusive. However, the following classification based on (1)–(5) *is* exhaustive and exclusive:

(a) (1);
(b) not (1), but (2);

[14] Belnap explicitly says that "[t]he theory of strategies permits us to make many [...] distinctions" [15, p. 125], which remark may be seen as foreshadowing the present development.

[15] Only rarely will it be too late for making good. However, we can specify reparational commitments (so-called "contrary-to-duties", cf. [32]) as *new* commitments that arise at or after the deadline of the commitment that was violated—and these reparational commitments will then have their own deadlines.

(c) neither (1) nor (2), but (3);
(d) neither (1) nor (2) nor (3), but (4);
(e) (5).

Thus, for any moment m with $m_0 \leqslant m$ and $i_{(m)} < t$ and for any $C \in Choice^{\alpha}_{m}$, exactly one of (a)–(e) holds. Now intuitively, that ranking of possible choices at m is already in a natural ordering with respect to stringency: Facing a commitment c, the best one can do with respect to that commitment is to choose in such a way as to guarantee fulfillment (a). The next best thing would be to follow a strategy guaranteeing fulfillment (b)—this is less stringent than (a), because fulfillment will still depend on future choices that cannot be fixed yet. After that, one might choose so as to make it possible that later on one might have a strategy guaranteeing fulfillment, but without guaranteeing fulfillment strategically at m itself (c). If one chooses neither of these options, choosing in such a way as to keep fulfillment an open possibility (d) must obviously count as more stringent than choosing in such a way as to make fulfillment impossible (e).

Of course not all of these options will be available at any given m. But for any choice C, one of these options always holds, so that our list (a)–(e) serves as a pointwise partial ordering of momentary choices of α's with respect to the given commitment c. Formally, we define:

Definition 9. For a commitment c, a moment m with $m_0 \leqslant m$ and $i_{(m)} < t$ and for $C, C' \in Choice^{\alpha}_{m}$, we say that C *is more stringent with respect to c at m than* C', or that C' *is less stringent with respect to c at m than* C ($C' \prec^{m}_{c} C$) iff C is rated higher in the above list (a)–(e) than C'.

If C and C' are both rated the same, the ordering relation will not hold either way.[16] Fig. 1 illustrates this pointwise ordering of choices for a simple case.

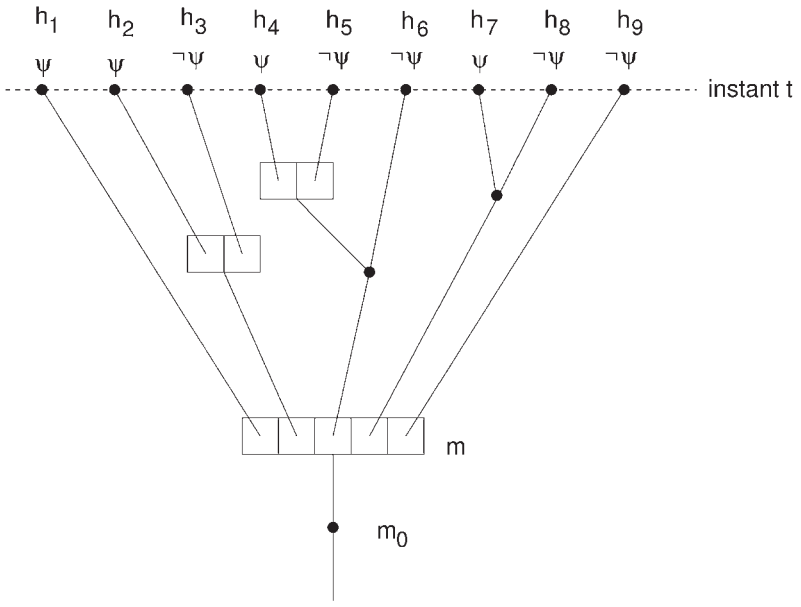

Fig. 1. Example of the pointwise partial ordering of choices for α at m according to Definition 9. Following the conventions of [15], the compartments of a box represent α's choices, whereas a simple dot indicates vacuous choice for α ("nature chooses" at such moments). We have $\{h_9\} \prec^{m}_{c} \{h_7, h_8\} \prec^{m}_{c} \{h_4, h_5, h_6\} \prec^{m}_{c} \{h_2, h_3\} \prec^{m}_{c} \{h_1\}$, illustrating the five cases (e)–(a), in that order.

[16] Our definition takes into account only the basic modal facts. If probabilities are available, these might naturally be used as tie-breakers. E.g., if C and C' are both rated (d) (both choices keep the possibility of fulfillment open, but none of them possibly leads to, or belongs to, a strategy for fulfillment), then the choice with respect to which fulfillment is more probable, will be more stringent with respect to the given commitment. In this paper, we refrain from introducing probabilities (or utilities, for that matter). Cf. [33] for an attempt at introducing probabilities into the branching space–times framework, which also applies to the (simpler) theory of branching time employed here.—Another tie-breaker might be discussed in case (c): arguably, the sooner a strategy for fulfillment can be available, the better for the commitment.

We can easily generalise this pointwise definition to a partial ordering of full strategies:

Definition 10. Let a commitment $c = \langle m_0, \alpha, \beta, t, \psi, Z\rangle$ be given, and let s and s' be two strategies for α defined on the future of m_0. We say that *s is more stringent with respect to c than s'* ($s' \prec_c s$) iff there is some $m \geqslant m_0$ such that (i) $s'(m) \prec_c^m s(m)$ and (ii) for all m' for which $m_0 \leqslant m'$ and $m' < m$, the strategies coincide ($s(m') = s'(m')$).

A further generalisation to the case of non-strict strategies is also straightforward: in comparing two non-strict choices $s(m)$ and $s'(m)$ locally, the stringency ordering holds ($s'(m) \prec_c^m s(m)$) iff at least one of the strict sub-choices C of $s(m)$ is more stringent that a strict sub-choice C' of $s'(m)$, and the remaining ones are tied with respect to $\prec_c^m$; the above definition then extends this local partial ordering to a global partial ordering of non-strict strategies as well.

5. Modelling the interaction of commitments and trust

How trustworthy is an agent? The discussion of Section 2.3 has shown that it is not enough to assume that agents are trustworthy if they fulfill their commitments with high probability, which could be monitored statistically in terms of actual fulfillment. We pointed out that it was *living up to a commitment*, rather than (extensional) fulfillment, that was most important.

Given the theory of agency in branching time plus strategies sketched in Section 3, together with the stringency ordering on strategies defined in the previous section, we are now in a position to sketch a formal model of the interrelation between commitments and trust. As we pointed out above, *both* directions of interaction play a role:

- In order to assess trustworthiness, one can monitor how agents treat the commitments they are under. If an agent is living up to her commitments, i.e., if the strategy she is following shows the required degree of stringency, that counts in favour of trustworthiness. Failing to live up to a commitment, on the other hand, counts against trustworthiness.
- In order to assess whether it would be appropriate to accept a commitment by some agent, one can compare the required level of trustworthiness with the overall trustworthiness of the agent, as exemplified by her record of living up to, or failing to live up to, commitments she was under.

Above we have characterised a commitment in the form

$$c = \langle m_0, \alpha, \beta, t, \psi, Z\rangle,$$

where m_0 is the moment at which the commitment is created, α is the agent who is committed, β is the addressee of the commitment, t is the deadline and ψ the content of the commitment, and Z is the sanction attached to non-fulfillment. In Section 4.1 we pointed out that a commitment is usually created by two agents who agree on the appropriate level of stringency of the commitment, which will be higher for typical promises than for typical contracts. Having available the stringency ordering on an agent's strategies, we can now spell out the level of stringency appropriate for a given commitment of α's towards β in the form of an antichain (i.e., a set of mutually *incomparable* strategies) S in the partial ordering of α's strategies relative to that commitment.[17] Thus, two agents do not just create a commitment of the form c above, but a *rated commitment* r, which consists of a commitment c (the subject matter of the commitment) and the rating (appropriate level of stringency) of that commitment in the form of an antichain S:

$$r = \langle c, S\rangle.$$

In view of what was said above about the perils of making considerations of trust explicit (cf. Section 2.2), S will often be tacitly understood and not discussed openly (the same holds for the sanction Z)—but if necessary, these matters can be discussed and brought to light.

We do not give a fully specified theory of stringency here. Belnap's proposal of differentiating between promising and word-giving, referred to in Section 4.2 above, amounts to assigning maximal stringency to those commitments

[17] Some simplifications are possible if we demand that a *maximal* antichain be specified. However, that demand may be too much for some applications, so we do not wish to impose it generally.

that arise from promises. Technically, assuming a finite set of available strategies for α this amounts to taking S to be the maximal antichain of maximal elements in the stringency ordering. While this gives a clear formal definition, realistically, only few commitments we enter into will be that strict.

If probabilities are available, it may be possible to devise a quantitative measure of stringency. However, in the light of the discussions of Section 2, the (required) probability of fulfillment of the commitment will not be the right measure.

Relative to a prospective application, we will need to have available both a way of assigning levels of stringency for commitments and a way of updating an agent's level of trustworthiness. Given all this machinery, we suggest the following, interestingly asymmetric scheme for the interaction of commitments and trust:

Assessing trustworthiness: Each agent may be assigned some initial level of trustworthiness to get things going. The level of trustworthiness of an agent will be updated at any instant with respect to all commitments she is under. The main idea is to identify the strategy that the agent is actually following by looking at what she chooses to do. As we pointed out above, a strict strategy completely reveals itself through the agent's choices, and if the agent is following a non-strict strategy, the strict strategy read off from her actions is still our best guess as to what her strategy actually is. The initial segment of the agent's strategy thus identified, s, may then be compared to the members of the class S_i of appropriate strategies for each commitment $r_i = \langle c_i, S_i \rangle$ in effect at the given instant. If the agent's strategy is itself appropriate (a member of S_i), or at least as stringent as one of the appropriate strategies, we say that the agent is *living up to that commitment*, which counts in favour of trust. On the other hand, if the agent's strategy is less stringent than appropriate, this counts against trusting the agent. The information thus gathered for all current rated commitments r_i can be used to update the agent's level of trustworthiness. Note that in this scheme of assessing trustworthiness, it is really the notion of living up to a commitment and not the actual fulfillment that is important.

Assessing whether to accept a commitment: When it comes to deciding whether to accept a commitment offered by an agent with some established level of trustworthiness, both trustworthiness and actual prospects of fulfillment will play a role. It is not normally wise to accept a commitment by an honest agent who is obviously incapable of fulfilling it, even though we can be sure that the agent will do her best. When we accept commitments, we do not just want to deal with good people, we also want things done. Thus, depending on the nature and content of the commitment, appropriate trustworthiness of an agent will be a necessary, but not a sufficient condition of acceptance.

6. Conclusions

In this paper we have supplied detailed phenomenological motivation and a number of formal steps towards technical implementation of a general framework for commitment and trust. Our initial focus on actual human practices has opened the view for the importance of the notion of living up to a commitment. We have argued that it is that notion, rather than actual fulfillment, that is at the core of our understanding of trustworthiness. While a notion of trustworthiness tied exclusively to extensional fulfillment has been criticised by others before [4], we claim to have identified a deeper reason for and a phenomenologically more appropriate description of our practices of trust and distrust.

The key technical innovation of this paper is to employ the theory of agents and strategies in branching time to define a stringency ordering of strategies, relative to a given commitment. That ordering in turn plays the central role in our scheme for identifying whether an agent is living up to a commitment. Further work is needed to find appropriate formats for characterising and updating agents' trustworthiness as exemplified by their record of living up to, or failing to live up to, commitments they are under. On the basis of the methodology exemplified in this work, we would propose that such modelling should also start from a broad phenomenological overview of actual human practices before considering technical details. Even though trust helps us reduce the complexity of our environment, it is itself a complex phenomenon. One should not expect to be able to capture it too easily. The theory of agents and choices in branching time, as employed here, appears to have the necessary resources for carrying our work further.

Acknowledgements

We would like to thank Michael Perloff, the referees for DEON2006 and for this journal, and the audience at DEON2006 in Utrecht for helpful suggestions. Support by the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

References

[1] N. Luhmann, Trust and Power, Wiley, New York, 1979, English translation of "Vertrauen" (1968) and "Macht" (1975).
[2] A.B. Seligman, The Problem of Trust, Princeton University Press, Princeton, NJ, 1997.
[3] M.P. Singh, An ontology for commitments in multiagent systems: Toward a unification of normative concepts, Artificial Intelligence and Law 7 (1999) 97–113.
[4] R. Falcone, C. Castelfranchi, Trust dynamics: How trust is influenced by direct experiences and by trust itself, in: Autonomous Agents and Multiagent Systems (AAMAM 2004), ACM Press, New York, 2004, pp. 740–747.
[5] G. Boella, L. van der Torre, Normative multiagent systems and trust dynamics, in: R. Falcone (Ed.), Trusting Agents, in: LNAI, vol. 3577, Springer, Heidelberg, 2005, pp. 1–17.
[6] G. Boella, L. van der Torre, A game theoretic approach to contracts in multiagent systems, IEEE Transactions on Systems, Man, and Cybernetics C 36 (1) (2006) 68–79.
[7] J. Huang, M.S. Fox, An ontology of trust: formal semantics and transitivity, in: Proceedings of the 8th International Conference on Electronic Commerce, in: ACM International Conference Proceeding Series, vol. 156, ACM Press, New York, 2006, pp. 259–270.
[8] M. Thompson, What is it to wrong someone? A puzzle about justice, in: R.J. Wallace, P. Pettit, S. Scheffler, M. Smith (Eds.), Reason and Value. Themes from the Moral Philosophy of Joseph Raz, Oxford University Press, Oxford, 2004, pp. 333–384.
[9] G. Möllering, The nature of trust: From Georg Simmel to a theory of expectation, interpretation and suspension, Sociology 35 (2001) 403–420.
[10] D. Hume, A Treatise of Human Nature, London, 1739/40, ed. by L.A. Selby-Bigge and P.H. Nidditch, Oxford University Press, Oxford, 1978.
[11] J. Raz, Promises and obligations, in: P. Hacker, J. Raz (Eds.), Law, Morality, and Society. Essays in Honour of H.L.A. Hart, Oxford University Press, Oxford, 1977, pp. 210–228.
[12] A.I. Melden, Rights and Persons, University of California Press, Berkeley, CA, 1977.
[13] M. van Lambalgen, F. Hamm, The Proper Treatment of Events, Blackwell, Oxford, 2005.
[14] P. Yolum, M.P. Singh, Reasoning about commitments in the event calculus: An approach for specifying and executing protocols, Annals of Mathematics and Artificial Intelligence 42 (2004) 227–253.
[15] N. Belnap, M. Perloff, M. Xu, Facing the Future. Agents and Choices in Our Indeterminist World, Oxford University Press, Oxford, 2001.
[16] N. Belnap, Bressan's type-theoretical combination of quantification and modality, in: H. Lagerlund, S. Lindström, R. Sliwinski (Eds.), Modality Matters. Twenty-Five Essays in Honour of Krister Segerberg, Uppsala Universitet, Uppsala, 2006, pp. 31–53.
[17] J.J. Thomson, The Realm of Rights, Harvard University Press, Cambridge, MA, 1990.
[18] N. Belnap, Branching histories approach to indeterminism and free will, preprint, http://philsci-archive.pitt.edu/documents/disk0/00/00/08/90, 2002.
[19] N. Belnap, Agents and agency in branching space–times, in: D. Vanderveken (Ed.), Logic, Thought and Action, Kluwer, Dordrecht, 2005, pp. 291–313.
[20] N. Belnap, Branching space–time, Synthese 92 (1992) 385–434.
[21] A.N. Prior, Past, Present and Future, Oxford University Press, Oxford, 1967.
[22] E.A. Emerson, Temporal and modal logic, in: J. van Leeuwen (Ed.), Handbook of Theoretical Computer Science, vol. B: Formal Models and Semantics, Elsevier, Amsterdam, 1990, pp. 996–1072.
[23] W. van der Hoek, M. Wooldridge, Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications, Studia Logica 75 (1) (2003) 125–157.
[24] R.H. Thomason, Indeterminist time and truth value gaps, Theoria (Lund) 36 (1970) 264–281.
[25] B. Chellas, The logical form of imperatives, PhD thesis, Stanford University, 1969.
[26] J.F. Horty, Agency and Deontic Logic, Oxford University Press, Oxford, 2001.
[27] T. Müller, On the formal structure of continuous action, in: R. Schmidt, I. Pratt-Hartmann, M. Reynolds, H. Wansing (Eds.), Advances in Modal Logic, vol. 5, King's College Publications, London, 2005, pp. 191–209.
[28] J. Broersen, A. Herzig, N. Troquard, A STIT-extension of ATL, in: M. Fisher, W. van der Hoek, B. Konev, A. Lisitsa (Eds.), Logics in Artificial Intelligence (Proceedings of JELIA 2006), in: LNAI, vol. 4160, Springer, Heidelberg, 2006, pp. 69–81.
[29] G.E.M. Anscombe, Intention, second ed., Harvard University Press, Cambridge, MA, 1963.
[30] N. Belnap, Double time references: Speech-act reports as modalities in an indeterminist setting, in: F. Wolter, H. Wansing, M. de Rijke, M. Zakharyaschev (Eds.), Advances in Modal Logic, vol. 3, World Scientific, Singapore, 2002, pp. 37–58.
[31] A.R. Anderson, A reduction of deontic logic to alethic modal logic, Mind 67 (1958) 100–103.
[32] J. Carmo, A.J.I. Jones, Deontic logic and contrary-to-duties, in: D. Gabbay, F. Guenthner (Eds.), Handbook of Philosophical Logic, vol. 8, second ed., Kluwer, Dordrecht, 2002, pp. 265–343.
[33] T. Müller, Probability theory and causation. A branching space–times analysis, British Journal for the Philosophy of Science 56 (2005) 487–520.