Logique & Analyse **215** (2011), 327–342

DEONTIC REASONING WITH INCOMPLETE TRUST*

MARTIN MOSE BENTZEN

> Trust is a solution for specific problems of risk.
> Luhmann (1990)

1. *Introduction*

It is said, that we can make the world a better place, if we allow ourselves to trust one another. In this paper, I show situations where this is the case. I also show some situations, where it is not the case. The technical contribution of the paper is a generalization of John Horty's account of *individual ought to do*, see Horty (2001), based on what game theorists call *rationalization of choices*, see e.g. (Osborne; 2004, chapter 12). Conceptually, this means an extension of the stit framework to deal with situations of trust and in particular iterated reciprocal trust, e.g. a trusts that b trusts a. Consider the following examples.

Example 1: *The Victim is held up by the evil guy. He is wondering whether to attempt to resist the evil guy or not. The Hero is wondering whether to help or not. The best outcome is when the Victim tries to resist the evil guy and the Hero helps. Nobody gets hurt, the evil guy goes to jail. With the second best outcome, the Hero helps but the Victim remains inactive. Here the evil guy gets killed and the Hero and the Victim will both suffer some bad wounds. The third best outcome is when the Hero does not help and the Victim does not resist. The Victim will get killed but without too much suffering. The worst outcome is when the Victim tries to resist and is not helped. In this case the evil guy tortures him to death.*

What should the Hero and the Victim do in this situation? The Hero can reason with the sure thing principle as follows. Given that the Victim resists, it is better for me to help, in which case all is well, than not to help, which would yield the worst possible outcome. On the other hand, given that the

[*] I would like to thank Stig Andur Pedersen for numerous helpful suggestions.

Victim does not resist, it is still better for me to help, because the good guys suffering some wounds and killing the bad guy, is still better than letting the Victim die. What should the Victim do? It seems impossible to say. Of course, if the Hero helps, he is a lot better off resisting, the best possible outcome of the situation. On the other hand, if the Hero does not help, he will be tortured to death by making this choice, which would be absolutely terrible. If he does not resist, he might get rescued anyway if the Hero decides to help, but the rescue will come at a high cost. On the other hand, if the Hero decides not to help, he will at least die a clean death and not be tortured. It is really a predicament. But assume now, that the Victim trusts the Hero to be a good utilitarian. Suppose, in particular, that he trusts the Hero to not make a choice which is strictly dominated. In that case, the Victim trusts the Hero to help. The Victim can now apply sure thing reasoning as follows. If I resist, then we will easily overcome the evil guy together. On the other hand if I don't resist, I leave all the dirty work to the Hero who will have to kill the bad guy and we will both get hurt. It is thus better for me to resist.

Here is another example, which requires two levels of reasoning.

Example 2: *The Doctor needs to reach town fast from the jungle to get medicine. It is a difficult journey. She can walk through the mountains or travel by boat down the river. She can also decide to abandon the journey altogether. Nearby lives the Guide, who has heard about this. He has to decide whether to come and guide the Doctor on the journey. Naturally, if the Doctor stays home, he would rather stay home, too. But if the Doctor should decide to either walk or go by boat he will be able to get her there faster either way, possibly saving lives. In particular, if the Doctor goes by boat, the Guide's navigational skills makes him very useful. It would yield the best possible outcome, if he were to decide to come and the Doctor were to decide to go by boat.*

What should the Doctor do? Make preparations to go by boat, stay home or abandon the journey? In this example it is not enough that the Doctor trusts the Guide. This is so, because what the Guide should do, depends on what the Doctor does. If the Doctor decides to abandon the journey, the Guide should stay home. If she goes on the journey by foot or by boat, he should help. However, suppose staying home is a morally bad choice for the Doctor no matter what. If the Guide trusts the Doctor, he knows that the Doctor will either go by boat or walk. And if the Doctor trusts the Guide and *she trusts that the Guide trusts her*, then she trusts the Guide will come to help. So in that case, the Doctor ought to go by boat, ensuring the best possible outcome. In other words, because the Guide trusts the Doctor he ought to come help. And because the Doctor trusts the Guide to trust her and she trusts the Guide, she ought to go by boat.

I will present a way of formalizing the reasoning above. First, I will consider some important elements of an informal theory of trust developed by Niklas Luhmann.

2. *Luhmann's Theory of Trust*

In sociologist Niklas Luhmann's view trust presupposes a situation of risk. More specifically,

> If you choose one action in preference to others in spite of the possibility of being disappointed by the action of others, you define the situation as one of trust.(...) Moreover, trust is only possible in a situation where the possible damage may be greater than the advantage you seek. Otherwise, it would simply be a question of rational calculation and you would choose your action anyway...
> (Luhmann; 1990, pp. 97–98)

One example of trust given by Luhmann is hiring a babysitter for the evening and leaving him or her unsupervised. Clearly, this gives us a situation analogous to the informal examples spelled out above. As a way of contrast, Luhmann makes a distinction between *confidence*, which we may capture as an attitude to a wider and more basic class of situations, and *trust*, which is related to specific situations. As an example of this distinction, we need confidence in the use of the evaluative object money (perhaps this confidence is based on a social contract), but we need trust when entering into specific situations of investment. The theory developed here, really concerns what Luhmann calls trust. Whereas lack of confidence will result in alienation, Luhmann claims the following.

> The *lack of trust*, on the other hand, simply withdraws activities. It reduces the range of possibilities for rational action.
> (Luhmann; 1990, p. 104)

And further,

> Mobilizing trust means mobilizing engagement and activities, extending the range and degree of participation.
> (Luhmann; 1990, p. 99)

Although the present theory is an extension of stit theory, which gets its justification independently of Luhmann, I think the relation to Luhmann's theory is clear enough to be interesting. If we take a narrow definition of individual rational choice as resulting from reasoning by Savage's *sure thing principle* (I do not think this is too far from what Luhmann has in mind), it is clear that the theory we will present extends the possibilities for rational action. This is exactly what I mean, when I say this theory is a generalization of Horty's individual ought. Also, the informal examples given above fulfil the conditions given by Luhmann to be characterized as situations of trust. The agents cannot expect to get to the best outcomes by only trusting themselves. Furthermore, by trusting each other they risk greater damage than if they did not trust (e.g. the Victim risks to be tortured to death by trusting the Hero, a fate which he considers worse than simply dying). Moreover, in contrast to Luhmann, who does not emphasize this aspect, the theory makes it apparent that individuals trusting other individuals, in itself is not always enough. As the example with the Doctor's journey shows, the Doctor needs to trust that the Guide trusts the Doctor in order to make the choice that leads to the best outcome. Thus we really need reciprocal and iterated modes of trust - by the way, I trust that Luhmann would not deny the importance of this. The formal theory enables us to spell out such conditions clearly and to give reasons to trust based choices, which we make intuitively all the time. Before I turn to the formal frame work, I spell out a bit, what we mean by agents being in specific situations.

3. *Strategic Situations*

In the version of stit theory studied here, we do not consider time. Formally, it corresponds to stit theory reduced to a single moment, as studied, e.g. in (Belnap et al.; 2001, Chapter 16), Kooi and Tamminga (2006). Intuitively, since we use only operators, whose satisfaction (in the full stit framework including time) would not depend on histories, throwing away these histories from the models at the outset should not matter logically. It makes the model theory simpler, since we essentially reduce the models to standard relational models known from modal logic, see e.g. Chellas (1980), Blackburn et al. (2001). For a formal mapping between the two kinds of models, see, Herzig and Schwarzentruber (2008). Conceptually, I do not think we should consider the models as representing *single moments*. Rather, we should consider them as *strategic situations*, which is to say that agents act *independently* (meaning that each of their choices is consistent with any choice of any other agent), but not necessarily *simultaneously*. In the informal examples presented in this paper the agents are in different locations, and they cannot communicate. Further, in these models, each agent is aware of all *possible* consequences of the different combination of choices and this awareness is common knowledge. Also, agents agree on which utility to assign to outcomes (only ordinal aspects of utilities are used), and this evaluation is also common knowledge. The conceptual difference between the account given here and those based on instrumental rationality such as in game theory (a difference which really only exists at a meta level) is that we do not require the utilitarian values to correspond to the individual utility functions of agents (these are not part of the formal frame work). For more about knowledge and stit, see Broersen (2008), and for knowledge in strategic situations in game theory, see van der Hoek and Pauly (2007). What an agent is not aware of, is which particular choice the other agents will make. The examples suggest that there might be many situations, where these assumptions come quite naturally. The deontic operators presented here for the first time, represent reasoning about what such an agent *ought to do* in such situations given various levels of trust. What is also new in this paper is the construction of submodels using positive formulas. This construction is applied in the iterative removal of strictly dominated strategies.

4. *Utilitarian Strategic Models*

I presuppose rudimentary set theory and classical logic for the meta language. Otherwise, the following presentation of the formal framework is self contained. However, space does not permit me to cover the philosophical foundations of stit theory. Instead, I refer to Horty (2001), Belnap et al. (2001), see also Lindström and Segerberg (2007). For information about deontic logic and modal logic in general, a good, standard reference is Chellas (1980). A good introduction to standard game theory is Osborne (2004). Okay, let's get started!

Formally, we use *utilitarian strategic models* (sometimes we simply call them *models*) consisting of outcomes, agents, choices, a utility function on the outcomes and a valuation function. In the models we consider the choices, agents, outcomes, values and valuations are common knowledge amongst all agents.

Let Φ be a denumerable set of propositional variables. (The rest of the language will follow later).

Definition 4.0.1: (Utilitarian strategic model) *A utilitarian strategic model is a structure* $M = \langle W, Agent, \{\mathfrak{A}_i \mid a_i \in Agent\}, u : W \to \mathbb{R}, V\rangle$*, where*

1. W *is a nonempty set of* outcomes*.*
2. *We have finite, nonempty set of* agents*,* $Agent = \{a_1, \ldots a_n\}$

3. *For each agent* $a_i \in Agent$*, we have a finite, nonempty set* $\mathfrak{A}_i$ *of* actions *(subsets of* W*), i.e.* $\mathfrak{A}_i = \{A_{i1}, \ldots, A_{im_i}\}$, $0 < m_i$*, (agent* a_i *has* m_i *choices). Furthermore,*
   (a) *For any agent* a_i*, the elements of* $\mathfrak{A}_i$ *partition* W.
   (b) *Let* $A_{1j_1} \in \mathfrak{A}_1, \ldots, A_{nj_n} \in \mathfrak{A}_n$*, where* $1 \leq j_i \leq m_i$*. Then* $(A_{1j_1} \cap \ldots \cap A_{nj_n}) \neq \emptyset$*. (Independence of Agents)*
4. u *is a utility function assigning values to outcomes.*
5. V *is a valuation function from atomic propositions to subsets of* W*, i.e.* $V : \Phi \rightarrow \mathcal{P}(W)$

When $M = \langle W, \ldots, V\rangle$ is a model, we sometimes write $dom(M)$ for W (the domain of M). It should be noted, that these models obey what Horty calls the *finite choice condition*, each agent considers only a finite number of choices. We call an action $A_{ij_i} \in \mathfrak{A}_i$, where $1 \leq j_i \leq m_i$, an *atomic action*. For an agent $a_i \in Agent$, we call the union of k_i ($k_i > 0$) actions from $\mathfrak{A}_i$, a *complex positive action* and we denote such a complex positive action α_i, i.e.

$$\alpha_i = A_{is_{i1}} \cup A_{is_{i2}} \cup \ldots \cup A_{is_{ik_i}} \; where \; 1 \leq s_{i1} \leq s_{i2} \leq \ldots \leq s_{ik_i} \leq m_i$$

(One may think of α_i as successively picking or leaving out each atomic action from $\mathfrak{A}_i$, possibly leaving out some, but picking at least one). Given a complex positive action α_i for each $a_i \in Agent$, we define an *action profile*, denoted P, as the intersection of the complex actions, i.e.

$$P = \bigcap_{a_i \in Agent} \alpha_i$$

When the action profile contains exactly one action for each agent we call it an atomic action profile (otherwise complex).

Let M be a model, and let P be an action profile. We define the sets of actions of agents restricted to the profile, denoted $\mathfrak{A}_i|P$ as follows, $\mathfrak{A}_i|P = \{A_{ij} \cap P \mid A_{ij} \cap P \neq \emptyset, j = 1, \ldots, m_i\}$. We now define the model restricted to P, denoted $M|P$, as follows.

Definition 4.0.2: $M|P = \langle P, Agent, \{\mathfrak{A}'_i \mid a_i \in Agent\}, u' : P \rightarrow \mathbb{R}, V'\rangle$, *where*

1. $\mathfrak{A}'_i = \mathfrak{A}_i|P$

2. $u' = u|P$ *(u restricted to P).*
3. *For each* $p \in \Phi, V'(p) = V(p) \cap P$.

We need to show that $M|P$ fulfils the conditions of Definition 4.0.1, in particular that the actions of agents partition $dom(M|P)$ and that $M|P$ fulfils the independence of agents condition. We show only the latter.

Proof. Independence of Agents Let $A'_{1j_1} \in \mathfrak{A}'_1, \ldots, A'_{nj_n} \in \mathfrak{A}'_n$. Each $A'_{ij_i} = A_{it_i} \cap P$ for some $A_{it_i} \in \mathfrak{A}_i$. We have $P = ((A_{1s_{11}} \cup A_{1s_{12}} \cup \ldots \cup A_{1s_{1k_1}}) \cap \ldots \cap (A_{ns_{n1}} \cup A_{ns_{n2}} \cup \ldots \cup A_{ns_{nk_n}})) \supseteq (A_{1t_1} \cap \ldots \cap A_{nt_n})$. Therefore, $(A'_{1j_1} \cap \ldots \cap A'_{nj_n}) = ((A_{1t_1} \cap P) \cap \ldots \cap (A_{nt_n} \cap P)) = ((A_{1t_1} \cap \ldots \cap A_{nt_n}) \cap P) = (A_{1t_1} \cap \ldots \cap A_{nt_n}) \neq \emptyset$ by Independence of Agents for M. □

From this we get the following.

Fact 4.0.3: *Let M be a strategic model and P an action profile. $M|P$ is a strategic model.*

Although restricting a model to an action profile is a sufficient condition for getting a new strategic model it is not necessary. The main thing is that the new model needs to fulfil the independence of agents condition. This rules out reductions based on the truth set of any formula, as in the public announcements considered in dynamic epistemic logic, see e.g. van Ditmarsch et al. (2007).

Given a model M (with W as its set of outcomes), we define the set M_P as the set of models resulting from restricting M to an action profile (atomic or otherwise) of M.

Definition 4.0.4: $M_P = \{M|P \mid P$ *is an action profile*$\}$

We order the set M_P as follows. $M|P' <_c M|P$, iff $P \subset P'$.

Fact 4.0.5: *M_P is finite. $<_c$ is a strict partial order on M_P with the atomic action profiles as maximal elements and M as minimal element.*

Proof. Since each agent has a finite number of actions, there is only a finite number of profiles, so M_P is finite. The strict partial order is forced by set inclusion. Obviously for any profile P, $P \subseteq W$. If P is an atomic profile there can be no profile P', such that $P' \subset P$, because that would require taking an action away from at least one agent, leaving us with an empty action for that agent, which violates the definition of an action profile. □

In order to get to sure thing reasoning we first lift the utility function on outcomes to a preference ordering on actions. The utilities of outcomes are lifted to arbitrary subsets of W, $S, T \subseteq W$, in the following way.

Definition 4.0.6: $S \leq T$ *iff* $sup(u(o))_{o \in S} \leq inf(u(o'))_{o' \in T}$.

Informally, the (upper limit of the) utility of the outcomes with the highest utility in S, is lower than or equal to the (lower limit of the) utility of the outcomes with lowest utility in T. $\leq$ is a transitive relation on $\mathcal{P}(W)$. The strict ordering on propositions is defined as $S < T$ iff $S \leq T$ and not $T \leq S$.

We now use the preference ordering to define a *dominance* ordering on actions by means of Savage's *sure thing principle*. Let P be an action profile. By P_{a_k}, we mean the complex action α_k of P. By P_{-a_k} we mean the action profile P without any action specified for a_k, i.e. $P_{-a_k} = \bigcap_{a_i \in Agent - \{a_k\}} \alpha_i$. By (P_{-a_i}, α_i), we mean the set $P_{-a_i} \cap \alpha_i$, where α_i is some complex positive action for a_i. $(P_{-a_i}, \alpha_i) = P'$ determines an action profile. If P is an atomic action profile, and α_i is atomic, P' will be an atomic action profile. We define *sure thing dominance* in the following way. For actions, $A_{im}, A_{in} \in \mathfrak{A}_i$, A_{im} weakly dominates action A_{in} (denoted $A_{in} \preceq A_{im}$) iff for any atomic action profile P we have $(P_{-a_i}, A_{in}) \leq (P_{-a_i}, A_{im})$. Intuitively, this means that with any possible combination of choices for all the other agents, it is at least as good if a_i chooses A_{im} as if a_i chooses A_{in}. Strict dominance, denoted $A_{in} \prec A_{im}$, is defined as $A_{in} \preceq A_{im}$ and not $A_{im} \preceq A_{in}$.

We define the set of optimal choices for an agent a_i in a model M, denoted $optimal^{a_i}_M$ as the set of actions for that agent that are not strictly dominated, i.e. $optimal^{a_i}_M = \{A_{im} \in \mathfrak{A}_i$ |there is no $A_{in} \in \mathfrak{A}_i$, such that $A_{im} \prec A_{in}\}$.

Fact 4.0.7: 1. (Horty 2001) *For any agent a_i, $optimal^{a_i}_M \neq \emptyset$.*

2. *$\bigcup optimal^{a_i}_M$ is a complex positive action.*

Proof. We repeat Horty's proof of 1. in the current (atemporal) framework for the convenience of the reader. Assume $optimal^{a_i}_M = \emptyset$. Let $A_{in} \in \mathfrak{A}_i$. Since $A_{in} \notin optimal^{a_i}_M$, there is a different action $A_{il} \in \mathfrak{A}_i$ such that $A_{in} \prec A_{il}$. Since $A_{il} \notin optimal^{a_i}_M$ either, we can iterate the argument indefinitely, giving us an infinite subset of $\mathfrak{A}_i$ contradicting that an agent has only finitely many atomic actions. 2. is immediate from 1.. □

It follows that the intersection of all such actions, $\bigcap_{a_i \in Agent}(\bigcup optimal^{a_i}_M)$, is an action profile, which we denote $optimal_M$. So, by Fact 4.0.3, $M|optimal_M$ is a strategic model. Now, fix a model M. We define a model M_n with level of trust n as follows.

Definition 4.0.8: 1. $M_0 = M$.

2. $M_{n+1} = M_n|optimal_{M_n}$.

Fact 4.0.9: 1. $dom(M_{n+1}) \subseteq dom(M_n)$.

2. *If* $m < n$, *then* $dom(M_n) \subseteq dom(M_m)$.

3. *There is an* m, *s.t.* $M_n = M_m$, *for all* $n \geq m$. *We call this model* M_m, *optimus*$'$*.By optimus*$'^{a_i}$ *we mean* $optimal^{a_i}_{M_m}$.

Proof. 1. We write $\mathfrak{A}^n_i$ for $\mathfrak{A}_i|optimal_{M_n}$, A^n_{is} for an element of $\mathfrak{A}^n_i$. Obviously $optimal^{a_i}_{M_n} = \{A^n_{im} \in \mathfrak{A}^n_i$ |there is no $A^n_{is} \in \mathfrak{A}^n_i$, such that $A^n_{im} \prec A^n_{is}\} \subseteq \mathfrak{A}^n_i$. Hence the largest possible reduction is to $\bigcap_{a_i \in Agent}(\bigcup \mathfrak{A}^n_i) = dom(M_n)$, and (since for any M, $M|dom(M) = M$), we have $dom(M_{n+1}) \subseteq dom(M_n)$.

2. Since $n > m$, M_n is obtained from M_m in a finite number of steps for each of which the previous argument holds, so we have $dom(M_n) \subseteq dom(M_m)$.

3.For any n, either $M_{n+1} = M_n$ or for some agent some action is dominated, in which case $optimal_{M_{n+1}} \subset optimal_{M_n}$, i.e. $M_n <_c M_{n+1}$. Now, since for any n, $M_n \in M_P$, and $<_c$ yields a finite partial order (see Definition 4.0.4 and Fact 4.0.5), this process must come to an end eventually, at the latest when it hits a maximal element (an atomic action profile, i.e. each agent is down to one non-dominated action). The models can only get smaller and they never become empty. □

Furthermore, we have the following.

Fact 4.0.10: *For any* n, a_i, $\{A_{ij} \in \mathfrak{A}_i \mid A_{ij} \cap optimal^{a_i}_{M_n} \neq \emptyset\} \neq \emptyset$.

In words there is a non-empty subset of actions from the original model consistent with the actions of the model restricted to $optimal_{M_n}$. We denote the union of the elements of this set $optimal^{\mathfrak{A}_i}_{M_n}$, i.e. $optimal^{\mathfrak{A}_i}_{M_n} = \bigcup\{A_{ij} \in \mathfrak{A}_i \mid A_{ij} \cap optimal^{a_i}_{M_n} \neq \emptyset\}$, call it *the* n*-optimal action for agent* a_i. It is in fact a complex positive action in M for a_i. Similarly, for $\bigcup\{A_{ij} \in \mathfrak{A}_i \mid A_{ij} \cap optimus'^{a_i}\}$, we write $optimus'^{\mathfrak{A}_i}_M$. We have the following.

Fact 4.0.11: 1. $optimal^{\mathfrak{A}_i}_{M_n} \subseteq optimal^{\mathfrak{A}_i}_{M_m}$*, for* $m < n$*.*

2. *There is an* m*, such that* $optimal^{\mathfrak{A}_i}_{M_m} = optimal^{\mathfrak{A}_i}_{M_n}$*, for any* $n > m$*.*

5. *New Deontic Operators*

We are now going to do deontic logic with the models constructed above. Based on the set of propositional variables, Φ, we build a language by the following rule. We use agents as names for themselves.

Definition 5.0.12: $a_i \in Agent$*,* $p \in \Phi$*.*
$\phi ::= p \mid \perp \mid \phi_1 \rightarrow \phi_2 \mid \bigcirc\phi \mid \mathsf{A}\phi \mid [a_i\ cstit]\phi \mid \bigodot[a_i\ cstit]_n\phi \mid \bigodot[a_i\ cstit]\phi$

We define the rest of the propositional connectives, $\neg, \wedge, \vee, \leftrightarrow$, in the standard way, ($\neg\phi$ is defined as $\phi \rightarrow \perp$, and so on). As usual, we write $M, o \vDash \phi$, for ϕ is true with outcome o of model M. By $|\phi|_M$ we mean the proposition expressed by ϕ, also called the truth set of ϕ, i.e. the set $\{o \mid M, o \vDash \phi\}$. The truth conditions for atomic sentences, the propositional constant, and propositional connective are standard:

Definition 5.0.13: 1. $M, o \vDash p$ *iff* $o \in V(p)$*, where* $p \in \Phi$ *(*p *is atomic).*

2. $M, o \vDash \perp$ *never.*

3. $M, o \vDash \phi \rightarrow \psi$ *iff, if* $M, o \vDash \phi$*, then* $M, o \vDash \psi$*.*

As usual, by ϕ being true in a model, written $M \vDash \phi$, we mean that ϕ is true with all outcomes of that model (for any $o \in dom(M)$, $M, o \vDash \phi$). By ϕ being *valid*, written $\vDash \phi$, we mean true in all models (for any model M, $M \vDash \phi$). By *logical consequence*, written $\Gamma \vDash \phi$, where Γ is a set of formulas and ϕ is a formula, we mean that for any outcome o, of any model M, if $M, o \vDash \psi$ for all $\psi \in \Gamma$, then $M, o \vDash \phi$. The standard deontic operator $\bigcirc$ has the following truth condition (recall that u is a utility function assigning values to outcomes).

$$M, o \vDash \bigcirc\phi \text{ iff. there is an outcome } o', \text{ such that } M, o' \vDash \phi \text{ and for all } o'', \text{ such that } u(o') \leq u(o''), M, o'' \vDash \phi$$

This is a normal modal operator, validating e.g. D ($\neg(\bigcirc\phi \wedge \bigcirc\neg\phi)$) and 4 ($\bigcirc\phi \rightarrow \bigcirc\bigcirc\phi$). Let a_i be an agent and o an outcome. By $choice^{a_i}_M(o)$ we mean the unique action $A_{im} \in \mathfrak{A}_i$, such that $o \in A_{im}$. The Chellas stit operator[1] has the following truth condition. (The following definitions and validities apply to any $a_i \in Agent$).

$$M, o \vDash [a_i\ cstit]\phi \text{ iff. } choice^{a_i}_M(o) \subseteq |\phi|_M.$$

The A operator is a universal modality.

$$M, o \vDash \mathsf{A}\phi \text{ iff. for any } o' \in dom(M),\ M, o' \vDash \mathsf{A}\phi.$$

Its dual E is defined as $\neg\mathsf{A}\neg$. The Chellas stit operators and the universal modality are both $\mathsf{S5}$ operators, and further:

Fact 5.0.14: $\vDash \mathsf{A}\phi \rightarrow [a_i\ cstit]\phi$.

We give the 'ought to do' operator with a level of trust $n > 0$ the following truth condition.

$$M, o \vDash \bigodot[a_i\ cstit]_n\phi \text{ iff. } optimal^{\mathfrak{A}_i}_{M_n} \subseteq |\phi|_M.$$

The intuition behind this operator is that if ϕ being true is a necessary condition for a_i to perform the optimal action given a level of trust n (the n-optimal action), then ϕ is obligatory for a_i. The n-optimal action might be complex, in which case we should think of it as a free choice between the action tokens it contains. We define the individual ought to do operator without subscript on $optimus'$. We give the individual ought to do operator without subscript the following truth condition.

$$M, o \vDash \bigodot[a_i\ cstit]\phi \text{ iff. } optimus'^{\mathfrak{A}_i}_M \subseteq |\phi|_M.$$

The following facts contain some validities for these operators.

Fact 5.0.15: *For any $n, m > 0$,*

1. $\vDash \phi$ *implies* $\vDash \bigodot[a_i\ cstit]_n\phi$
2. $\vDash \bigodot[a_i\ cstit]_n(\phi \rightarrow \psi) \rightarrow (\bigodot[a_i\ cstit]_n\phi \rightarrow \bigodot[a_i\ cstit]_n\psi)$
3. $\vDash \neg(\bigodot[a_i\ cstit]_n\phi \wedge \bigodot[a_i\ cstit]_n\neg\phi)$

[1] Our syntax deviates a bit from the one found in Horty (2001), Belnap et al. (2001), where the Chellas stit is written $[a_i\ cstit : \phi]$.

4. $\vDash \bigodot[a_i\ cstit]_n\phi \rightarrow \bigodot[a_i\ cstit]_m\phi$, *for* $n < m$.
5. $\vDash \bigodot[a_i\ cstit]_n\phi \rightarrow \bigodot[a_i\ cstit]\phi$
6. *There is some* m, *such that for all* $n \geq m \vDash \bigodot[a_i\ cstit]_n\phi \leftrightarrow \bigodot[a_i\ cstit]\phi$

Proof. 1. Assume $\vDash \phi$. Let $o \in dom(M) = W$ for some M. Since $|\phi|_M = W$, $optimal^{\mathfrak{A}_i}_{M_n} \subseteq |\phi|_M$, so $M, o \vDash \bigodot[a_i\ cstit]_n\phi$. Hence $\vDash \bigodot[a_i\ cstit]_n\phi$. 2. Assume $M, o \vDash \bigodot[a_i\ cstit]_n(\phi \rightarrow \psi)$ and $M, o \vDash \bigodot[a_i\ cstit]_n\phi$. Let $o' \in optimal^{\mathfrak{A}_i}_{M_n}$. Since $optimal^{\mathfrak{A}_i}_{M_n} \subseteq |\phi|_M$ and $optimal^{\mathfrak{A}_i}_{M_n} \subseteq |\phi \rightarrow \psi|_M = -|\phi|_M \cup |\psi|_M$, $o' \in |\psi|_M$. Hence, $optimal^{\mathfrak{A}_i}_{M_n} \subseteq |\psi|_M$, so $M, o \vDash \bigodot[a_i\ cstit]_n\psi$. 3. Assume for the sake of a contradiction that there is some model M and some outcome o, such that $M, o \vDash \bigodot[a_i\ cstit]_n\phi \wedge \bigodot[a_i\ cstit]_n\neg\phi$. Hence $optimal^{\mathfrak{A}_i}_{M_n} \subseteq |\phi|_M$ and $optimal^{\mathfrak{A}_i}_{M_n} \subseteq |\neg\phi|_M$, so $optimal^{\mathfrak{A}_i}_{M_n} \subseteq (|\phi|_M \cap |\neg\phi|_M) = \emptyset$. So, $optimal^{\mathfrak{A}_i}_{M_n} = \emptyset$, which contradicts Fact 4.0.10. 4. Assume $M, o \vDash \bigodot[a_i\ cstit]_n\phi$ and $n < m$. By Fact 4.0.11 $optimal^{\mathfrak{A}_i}_{M_m} \subseteq optimal^{\mathfrak{A}_i}_{M_n}$, so, $optimal^{\mathfrak{A}_i}_{M_m} \subseteq |\phi|_M$, and hence $M, o \vDash \bigodot[a_i\ cstit]_m\phi$. 5. Assume $M, o \vDash \bigodot[a_i\ cstit]_n\phi$. Since $optimus'^{\mathfrak{A}_i}_M \subseteq optimal^{\mathfrak{A}_i}_{M_n} \subseteq |\phi|_M$, $M, o \vDash \bigodot[a_i\ cstit]\phi$. 6. Take m to be such that $optimal_{M_m} = optimus'_M$. For any $n > m$, $optimus'^{\mathfrak{A}_i}_M = optimal^{\mathfrak{A}_i}_{M_n}$, so $\vDash \bigodot[a_i\ cstit]_n\phi \leftrightarrow \bigodot[a_i\ cstit]\phi$. □

Since the set of valid formulas is obviously closed under Modus Ponens (we have $\{\phi, \phi \rightarrow \psi\} \vDash \psi$), 1. (Necessitation) and 2.(K) show that $\bigodot[a_i\ cstit]_n$ is a normal modal operator. It is easily shown for the operator without subscript, $\bigodot[a_i\ cstit]$, as well. 3. is the characteristic deontic formula, saying that if we are on one level of trust, there can be no moral conflicts. Again, it holds for $\bigodot[a_i\ cstit]$, also. 4. shows that there is also consistency across levels of trust, in the sense that no obligation is lost when going to higher levels of trust. From these two validities, it follows that no obligation can be contradicted on a higher level of trust. 5. and 6. show some rather obvious interactions between the subscripted and non-subscripted operators. That any obligation is preserved by the non-subscripted operator, and that there is a finite level of trust from which adding more levels of trust is unnecessary, since it just gives the same obligations. In stit theory *ability* is expressed by $\mathsf{E}[a_i\ cstit]\phi$. One can think of this formula as expressing 'a_i has the choice to enforce ϕ.' We have the following important principle of *ought implies can*.

Fact 5.0.16: *For any* $n \vDash \bigodot[a_i\ cstit]_n\phi \rightarrow \mathsf{E}[a_i\ cstit]\phi$.

Proof. Assume $M, o \vDash \bigodot[a_i\ cstit]_n\phi$. Then $optimal^{\mathfrak{A}_i}_{M_n} = \bigcup\{A_{ij} \in \mathfrak{A}_i \mid A_{ij} \cap optimal^{a_i}_{M_n} \neq \emptyset\} \subseteq |\phi|_M$. Let $A_{ij} \subseteq \bigcup\{A_{ij} \in \mathfrak{A}_i \mid A_{ij} \cap optimal^{a_i}_{M_n} \neq \emptyset\}$ (by Fact 4.0.10, there is such an A_{ij}.) Let $o' \in A_{ij}$. Now $choice^{a_i}_M(o') = A_{ij} \subseteq |\phi|_M$, hence $M, o' \vDash [a_i\ cstit]\phi$. Hence $M, o \vDash \mathsf{E}[a_i\ cstit]\phi$. □

This validity says, that we do not demand too much of the agents in the following sense. If an agent ought to do ϕ, she in fact can see to it that ϕ. Furthermore, all deontic operators are *settled* in the sense that they are either true in the whole model or false in the whole model. I.e.

Fact 5.0.17: *For any level of trust* m *and any* $o \in W$, $M, o \vDash \bigodot[a_i\ cstit]_m\phi$ *iff* $M \vDash \bigodot[a_i\ cstit]_m\phi$.

Proof. Right to left is trivial. For left to right, assume $M, o \vDash \bigodot[a_i\ cstit]_m\phi$ and let $o' \in dom(M)$. Since $optimal^{\mathfrak{A}_i}_{M_m} \subseteq |\phi|_M$, $M, o' \vDash \bigodot[a_i\ cstit]_m\phi$. □

Thus we are justified in talking about what agents ought to do at the level of *models*, i.e. in a *strategic situation*, rather than just with particular outcomes of such a situation. (Naturally, we *can* talk about the latter as well, but the fact shows that there is no difference).

6. *Formalizing the examples*

The first example is represented by figure 1. The atomic formula R is true iff the Victim resists. The atomic formula H is true iff the Hero helps.

		$Victim\ (a_1)$	
		$Resist$	$Don't\ resist$
$Hero$	$Help$	$o_1 : R, H, u(o_1) = 4$	$o_2 : H, u(o_2) = 3$
	$Don't\ help$	$o_3 : R, u(o_3) = 1$	$o_4 : u(o_4) = 2$

Figure 1. Hostage Situation

Considered as a formal model, M, we have $M \nvDash \bigodot[a_1\ cstit]_1 R$, with agents who only trust themselves, it is not the case that the Victim ought to resist. On the other hand we have $M \vDash \bigodot[a_1\ cstit]_2 R$, with agents trusting themselves and each other, the Victim ought to resist. In this case, since we are down to one action per agent, we have $M_2 = \mathit{optimus}'$, so $M \vDash \bigodot[a_i\ cstit]_2\phi \leftrightarrow \bigodot[a_i\ cstit]\phi$. Adding further levels of trust will not give us any more obligations. This example also shows that for some model, $M \nvDash \bigodot[a_i\ cstit]\phi \rightarrow \bigodot[a_i\ cstit]_1\phi$. The account thus generalizes Horty's individual ought to do, Horty (2001), which is our $\bigodot[a_i\ cstit]_1$, because more propositions may be obligatory on this account. It is of course a matter of context, whether agents are justified in trusting each other and the indexed operator gives us flexibility to meet different modeling needs in this respect. The second example is treated in a similar way. It is represented by figure 2. B is a propositional atom meaning that the Doctor goes by boat. Here we have $dom(M_1) = \{o_1, \ldots, o_4\}$ (The Doctor staying home is dominated), and $dom(M_2) = \{o_1, o_3\}$ (The Guide staying home is dominated). We have $dom(M_3) = \{o_3\}$ (The Doctor walking is dominated). Since we are down to an atomic action profile, no further levels of trust will subtract more from the model. We thus have $M \vDash \bigodot[Doctor\ cstit]B$, the Doctor ought to go by boat.

	Walk	$o_1 : u(o_1) = 5$	$o_2 : u(o_2) = 4$
Doctor	*Go by boat*	$o_3 : B, u(o_3) = 6$	$o_4 : B, u(o_4) = 3$
	Stay home	$o_5 : u(o_5) = 1$	$o_6 : u(o_6) = 2$
		Go help	*Stay home*
		Guide	

Figure 2. The Doctor's Journey

7. *The Meinong-Chisholm thesis*

The Meinong-Chisholm thesis[2] is the following claim:

> An agent a ought to see to it that ϕ, if and only if, it ought to be the case that the agent a sees to it that ϕ.

The Meinong-Chisholm thesis stands refuted with the theory presented here. There are cases where it ought to be that the agent sees to it that ϕ is still not equivalent to that the agent ought to see to it that ϕ, for instance, we might have $M, o \vDash \bigcirc[a_i\ cstit]\phi$, but not $M, o \vDash \bigodot[a_i\ cstit]\phi$. In certain

[2] See (Lindström and Segerberg; 2007, p.1204). This thesis was originally called the *Meinong/Chisholm Analysis* by Horty, see (Horty; 2001, p. 45).

situations, genuine group reasoning (individuals act as parts of groups) can get us closer, e.g. in *hi-low* scenarios. Here is an example of such a scenario.

Example 3: *Two friends, Friend 1 and Friend 2, (who cannot communicate beforehand) both face the choice of going to town to meet their buddy. It would yield the best outcome if they both went. However, going to town alone is futile and a big waste of energy. So the two outcomes, where one friend goes and the other stays home, are the worst. If both friends stay home, it is better than if one goes in vain, but not as good as both meeting up in town.*

Formally, the situation looks like in Figure 3, where G_1 is a propositional atom, which means 'Friend 1 goes to town.'

		$Friend\ 2$	
		Go	$Stay$
$Friend\ 1$	Go	$o_1 : G_1, u(o_1) = 3$	$o_2 : G_1, u(o_2) = 1$
	$Stay$	$o_3 : u(o_3) = 1$	$o_4 : u(o_4) = 2$

Figure 3. Friends Meeting in Town

Even though this situation appears to have a similar structure to the first example, they are in fact essentially different. The difference is simply that none of the actions of either agent are strictly dominated. Therefore $M|optimal_M = M$. It follows that e.g. $M, o_1 \nvDash \odot[Friend_1\ cstit]G_1$, we cannot say that Friend 1 ought to see to it, that he goes. On the other hand going is a necessary condition for obtaining the best outcome in the situation, so we have $M, o_1 \vDash \bigcirc[Friend_1\ cstit]G_1$. It ought to be that Friend 1 sees to it that he goes. One way of getting there is to extend the theory to also cover agents trusting in *groups*. This extension, which is postponed for further research, could be based on Horty's group ought operator, see Horty (2001). Even such an account, however, would not validate the Meinong-Chisholm thesis, since there are pure coordination situations, where the agents simply cannot know what they ought to do, whether they identify with a group or not. We can transform the situation above into a pure coordination situation, by assuming that the utility of both agents staying home is exactly the same as of meeting up in town. Here, conditional accounts of ought to do (given that Friend 1 stays, Friend 2 ought to stay, etc), seem appropriate, but this too, is beyond the scope of this paper. It is clear, though, that such conditional oughts cannot tell agents what to do in a pure coordination situation *as a whole*, but only when fixing certain circumstances, which we take as the antecedent of the conditional ought.

Roskilde University
The Department of Culture and Identity
Philosophy and Science Studies
Universitetsvej 1, P.6
DK-4000, Roskilde
Denmark
E-mail: `mamobe@ruc.dk`

REFERENCES

Belnap, N., Perloff, M. and Xu, M. (2001). *Facing the Future*, Oxford University Press.

Blackburn, P., de Rijke, M. and Venema, Y. (2001). *Modal Logic*, Cambridge University Press.

Broersen, J. (2008). A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing', *Proceedings of the Ninth International Workshop on Deontic Logic in Computer Science (DEON 2008)*, Lecture Notes in Computer Science, Springer, pp. 140–154.

Chellas, B. F. (1980). *Modal Logic*, Cambridge University Press.

Herzig, A. and Schwarzentruber, F. (2008). Properties of logics of individual and group agency, *Advances in Modal Logic*, Vol. 7.

Horty, J. F. (2001). *Agency and Deontic Logic*, Oxford University Press.

Kooi, B. and Tamminga, A. (2006). Conflicting obligations in multi-agent deontic logic, *DEON 06*, pp. 175–186.

Lindström, S. and Segerberg, K. (2007). Modal logic and philosophy, *in* P. Blackburn, J. van Benthem and F. Wolter (eds), *Handbook of Modal Logic*, Elsevier, pp. 1150–1214.

Luhmann, N. (1990). Familiarity, confidence, trust: Problems and alternatives, *in* D. Gambetta (ed.), *Trust: Making and Breaking Cooperative Relations*, Blackwell Publishers, pp. 94–107.

Osborne, M. J. (2004). *An Introduction to Game Theory*, Oxford University Press.

van der Hoek, W. and Pauly, M. (2007). Modal logic for games and information, *in* P. Blackburn, J. van Benthem and F. Wolter (eds), *Handbook of Modal Logic*, Elsevier, pp. 1077–1148.

van Ditmarsch, H., van der Hoek, W. and Kooi, B. (2007). *Dynamic Epistemic Logic*, Springer.