What an Agent Ought To Do

A Review of John Horty's 'Agency and Deontic Logic'

Jan Broersen[1] and Leendert van der Torre[2]

[1] Department of Information and Computing Sciences, Utrecht University
[2] University of Luxembourgh

1 Introduction

John Horty's book 'Agency and deontic logic' appeared at Oxford Press in 2001. It develops deontic logic against the background of a theory of agency in non-deterministic time. Several philosophical reviews of the book appeared since then [1–5]. Our goal is to present the book to a general AI audience that is familiar with action theories developed in AI, classical decision theory [6], or formalizations of temporal reasoning like Computation Tree Logic (CTL) [7, 8]. Therefore, in contrast to the philosophical reviews, we discuss and explain several key examples in the book. We do not explicitly discuss the relevance for AI and law, because the book itself is not concerned with the application of the theory to the legal domain. However, the relevance of deontic logic and normative reasoning for legal reasoning is well established by a number of publications on deontic logic in AI and law, see for example the special issue of this journal on agents and norms (volume 4, 1999).

Horty presents a formal account of what individuals and groups of agents ought to do under various conditions and over extended periods of time, using the 'Seeing To It That' or STIT framework [9]. He explicitly develops a utilitarian / consequentialist perspective, which means roughly that an act is obligatory if performing it results in an optimal state of affairs. However, the question whether a state of affairs is 'optimal' is not a question that is exclusively linked to the deontic point of view. And also, seeing deontic necessity only from the perspective of an agent's welfare (optimality), might not suffice to model all subtleties involved in the semantics of deontic notions. Therefore, it is easy to be confused by the examples; sometimes it is not immediately clear why they are especially relevant for deontic logic.

Horty focusses on the common assumption that what an agent ought to do can be identified with the notion of what it ought to be that the agent does, and argues that this assumption is wrong. The assumption is based on the well-known conceptual distinction in deontic logic that concerns the notions of *ought-to-be* and *ought-to-do*. Roughly, ought-to-be deontic statements express a norm about the satisfaction of certain *conditions* at certain moments. Ought-to-do deontic statements apply to *actions*, which have been argued to fall in a different ontological category than conditions [10, 11]. The distinction between ought-to-do and ought-to-be forms the starting point for Horty's journey.

Dagstuhl Seminar Proceedings 07122
Normative Multi-agent Systems
http://drops.dagstuhl.de/opus/volltexte/2007/905

The central problem addressed in the first three chapters is the question whether ought-to-do deontic statements can be formalized within a STIT-framework that is extended with a Standard Deontic Logic or SDL-style ought-operator [12]. In particular, it is investigated whether it is intuitive to model 'agent α ought to do A' as 'it ought to be that agent α sees to it that A' in the STIT-framework. Horty argues that the answer is negative, and proposes, in chapter 4, a deontic operator that does formalize ought-to-do statements within the STIT-framework. In the remaining three chapters of his book Horty generalizes this theory to the conditional case, the group case, and the strategic case. The emphasis throughout the book is conceptual rather than technical, and as such the book is more aimed at offering food for thought for developers of deontic logic than at providing deontic logics which can directly be used in applications. Questions concerning deontic logic and the logic of agency are considered in parallel with issues from moral philosophy and the philosophy of action. This allows for a number of recent issues from moral philosophy to be set out clearly and discussed from a uniform point of view.

The review consists of two parts. In the first part we relate STIT-theory to standard decision trees, and explain concepts and ideas by selecting and discussing examples that are central to Horty's work. The critical discussion in the second part concerns three aspects: the examples, the concepts modelled by Horty's logic, and logical and technical issues.

2 Examples

At first sight, Horty's examples may seem innocent and their formalization straightforward. However, a more detailed analysis reveals that each example highlights a basic choice, which also is bound to appear in more detailed and realistic examples. The examples thus play the same role as simple examples in reasoning about action and change, like the widely discussed Yale shooting problem and stolen car problem that illustrate the frame problem [13, 14].

2.1 From decision trees to STIT models

Horty uses STIT-models to discuss a variety of examples and concepts. However, STIT-theory is not well-known outside the area of philosophical logic. Therefore we first explain STIT-theory by relating STIT-models to standard decision trees, which are well known in artificial intelligence. Roughly, each STIT-model contains a classical decision tree with decision nodes that abstracts from the probabilities. Horty describes the relation with decision theory as follows: 'The new analysis is based on a loose parallel between action in non-deterministic time and choice under uncertainty, as it is studied in decision theory' [15, p.4].

Decision trees are widely used to formalize decision problems. The branches of decision trees represent courses of time. Nodes without branches are called 'terminal nodes', to distinguish them from other nodes where time may advance and branch. Branching is either due to a choice made by the decision-making

agent, or due to the occurrence of events, where each possible event is associated with a probability. Branching nodes reflecting points of choice for the agent are called 'decision nodes'. Branching nodes that correspond to moments where events occur are called 'event nodes'. It is assumed that the decision-making agent knows what the probabilities are for each branch of an event node. The sum of the probabilities for the possible events at an event node is 1. Paths from the root to a terminal node correspond to sequences of choices and events in time. With each path a utility is associated, a kind of payoff under uncertainty. Rationality is defined as choosing an alternative that has the highest expected utility.

Of this decision theoretic setting, Horty adopts the utilities associated with series of events and choices in a decision tree. Paths are called 'histories'. Roughly, STIT-models can be constructed from decision trees by dropping the event nodes, and consequently, the probabilities. The branches of a dropped event node are connected to the first decision node closer to the root in the tree, to form a non-deterministic action. Figure 1 visualizes such a transformation by showing a decision tree and the utilitarian STIT-model it reduces to. The decision tree should be read as follows. Boxes represent decision nodes and circles represent event nodes. Numbers at the end of paths through the tree represent utilities, and events are provided with probabilities. Expected utilities are represented at each node in italics.

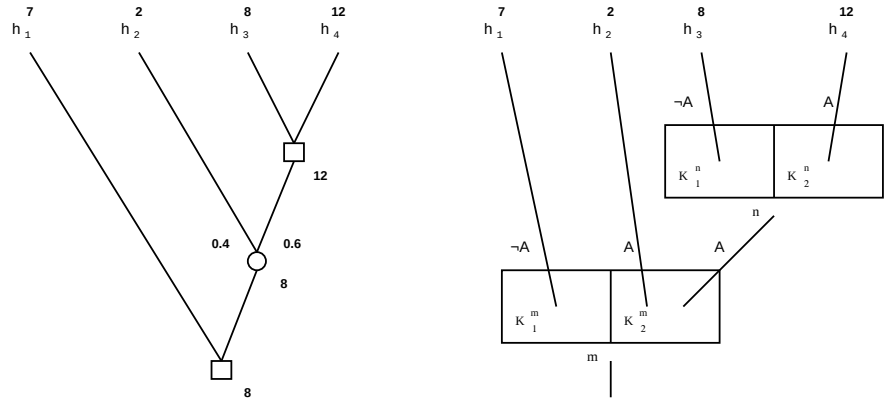

Fig 1. A decision tree and the corresponding utilitarian STIT-model

The two decision nodes of the decision tree correspond to the moments m and n in the STIT-models. Histories are series of events and choices from the root to the leaves of a STIT-model ($h_1 \dots h_4$ in figure 1). The event node in the decision tree on the left has turned into a non-deterministic action K_2^m, in the STIT-model on the right. Note that in the decision tree the choices in decision nodes are always deterministic. This is because in the decision tree, the non-determinism introduced by the events is 'temporally separated' from the decision nodes. Another distinction is that decision trees have no valuations of atomic propositions. In STIT-models there is a valuation of atomic propositions

for every moment-history pair. Horty denotes the set of histories through a moment m for which the atomic proposition A holds by $|A|_m$, and the utility of a history h by $Value(h)$.

Now we explain the semantics of some key concepts in Horty's STIT-formalism. A noticeable feature of the semantics is that formulas are evaluated with respect to moment-history pairs, and not with respect to moments, which is the viewpoint adopted in many temporal formalisms used in AI (e.g., CTL [7, 8]). This refinement of the unit of evaluation is induced by the basic assumption of the STIT-framework that actions constrain the possible future courses of time without actually 'taking' time. This means that we need to partition the histories of each moment according to the set of actions possible at it.

Some basic temporal formulas of Horty's utilitarian STIT-formalism are A and FA for 'the atomic proposition A' and 'some time in future A'. In particular, A is settled true at a moment-history pair m, h if and only if it is assigned the value 'true' in the STIT-model; FA is settled true at a moment-history pair m, h if and only if there is some future moment on the history, where A is settled true. On the STIT-model of figure 1 we have $\mathcal{M}, m, h_3 \models A$, which follows directly from the valuation of atomic propositions on moment-history pairs, and $\mathcal{M}, m, h_3 \models F\neg A$, which is due to the fact that the proposition $\neg A$ is true later on, at moment n, on the history h_3 through m.

An action formula is $[\alpha\ cstit : A]$, 'agent α Sees To It That A'. The 'c' in 'cstit' stands for 'Chellas', whose version of the STIT-operator [16] is predominant in Horty's work. $[\alpha\ cstit : A]$ is settled true at a moment history pair m, h if and only if A is settled true at all moment-history pairs through m that belong to the same *action* as the pair m, h, i.e., if $h \in K$ at m then $K \subseteq |A|_m$. Following Horty, we use a symbol like K both as a name of an action at a moment m and as a denotation for the set of 'admissible' histories determined by that action at moment m. In figure 1, we have $\mathcal{M}, m, h_3 \models [\alpha\ cstit : A]$, because A holds for all histories through m that belong to the action to which also h_3 belongs, that is, action K_2^m.

Finally a deontic formula $\bigcirc A$ stands for 'it ought to be that A'. $\bigcirc A$ is settled true at a moment history pair m, h if and only if there is some history h' through m such that A is settled true at all pairs m, h'' for which the history h'' has a utility at least as high as h', i.e., $\exists(m, h')$ such that $\forall(m, h'')$ for which $Value(h') \leq Value(h'')$ it holds that $h'' \in |A|_m$. In figure 1 we have $\mathcal{M}, m, h_3 \models \bigcirc A$ and $\mathcal{M}, m, h_3 \models \bigcirc[\alpha\ cstit : A]$. These two meta-propositions are true for the same reason: the history h_4 through m has the highest utility and satisfies both A and $[\alpha\ cstit : A]$ at m.

Note that this condition guarantees that on separate histories through a moment any ought formula evaluates to the same value, which is why ought-formulas are called 'moment determinate'. This semantic condition for the to-be ought is a utilitarian generalization of the standard deontic logic view (SDL [12]) that 'it ought to be that A' means that A holds in all deontically optimal worlds. Satisfaction of a formula A by a STIT-model can be defined as truth of A in all moment-history pairs of the model, and validity as satisfaction by all STIT-

models. Horty does not give these definitions explicitly, but this is the general STIT-view on validity (see, e.g., [9]).

2.2 'Ought-to-do' and the gambling problem

The central thesis of the book is that ought-to-do statements cannot be formalized as ought-to-be statements about action. More precisely, Horty claims that 'agent α ought to see to it that A' cannot be modeled by the formula $\bigcirc[\alpha\ cstit : A]$, whose reading is 'it ought to be that agent α sees to it that A'. Justification of this claim is found in what Horty calls the 'gambling problem' [15, p.53-58]. This example concerns the situation where an agent faces the choice between gambling to double or lose five dollar (action K_1^m) and refraining from gambling (action K_2^m). This STIT model is visualized in figure 2.

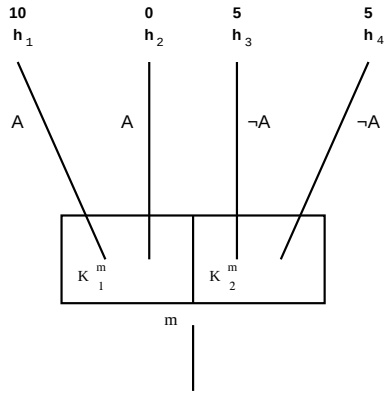

Fig 2. The gambling problem [15, Fig 3.8]

The two histories that are possible by choosing action K_1^m represent ending up with ten dollar by gaining five, and ending up with nothing by loosing all, respectively. Also for action K_2^m, the game event causes histories to branch. For this action the two branches have the same utility, because the agent is not taking part in the game, thereby preserving his five dollar. Note this points to redundancy in the model representation: the two branches are logically indistinguishable, because there is no formula whose truth value would change by dropping one of them.

The formula $\bigcirc[\alpha\ cstit : A]$ is settled true at m, because the formula $[\alpha\ cstit : A]$ is settled true for history h_1 and for all histories with a higher utility (of which there are none!). However, a reading of $\bigcirc[\alpha\ cstit : A]$ as 'agent α ought to perform action K_1^m' is counter-intuitive for this example. From the description of the gambling scenario it does not follow that one action is better than the other. In particular, without knowing the probabilities, we cannot say anything in favor of action K_1^m: by choosing it, we may either end up with more or with less money then by doing K_2^m. The only thing one may observe is that action

K_1^m will be preferred by more adventurous agents. But that is not something the logic is concerned with.

This demonstrates that 'agent α ought to see to it that A' cannot be modelled by $\bigcirc[\alpha\ cstit : A]$. The cause of the mismatch can be explained as follows. Adapting and generalizing the main idea behind SDL to the STIT-context, ought-to-be statements concern truth in a set of optimal histories. Optimality is directly determined by the utilities associated with the individual histories. If ought-to-be is about optimal histories, then ought-to-do is about optimal actions. But, since actions are assumed to be non-deterministic, actions do not correspond with individual histories, but with *sets* of histories. This means that to apply the idea of optimality to the definition of ought-to-do operators, we have to generalize the notion of optimality such that it applies to *sets* of histories, namely, the sets that make up the non-deterministic actions. More specifically, we have to obtain an ordering of non-deterministic actions that is based on the underlying ordering of histories. The ordering of actions suggested by Horty is very simple: an action is strictly better than another action if all of its histories are at least as good as any history of the other action, and not the other way around.

Having 'lifted' the ranking of histories to a ranking of actions, the utilitarian ought conditions can now be applied to actions. Thus, Horty defines the new operator 'agent α ought to see to it that A', written as $\bigodot[\alpha\ cstit : A]$, as the condition that for all actions not resulting in A there is a higher ranked action that does result in A, together with the condition that all actions that are ranked even higher also result in A. This 'solves' the gambling problem. We do not have $\bigodot[\alpha\ cstit : A]$ or $\bigodot[\alpha\ cstit : \neg A]$ in the gambling scenario, because in the ordering of actions, K_1^m is not any better or worse than K_2^m. So, it is not the case that the agent ought to gamble, nor is it the case that the agent ought to refrain from gambling.

2.3 The driving example

Horty generalizes the ordering on actions to the multi-agent context by imposing the so-called 'sure-thing principle' [6]. If there are only two agents, then at m for agent α action K_1^m is better than action K_2^m if for *each* action K_3^m by agent β it holds that $K_1^m \cap K_3^m$ is better than $K_2^m \cap K_3^m$. Here, an intersection like $K_1^m \cap K_3^m$ stands for a group action where agent α and agent β simultaneously perform K_1^m and K_3^m, respectively. The actions optimal for an agent α at a moment m are denoted $Optimal_\alpha^m$. The corresponding generalized operator $\bigodot[\alpha\ cstit : A]$ reflects what Horty calls 'dominance act utilitarianism'. The driving example [15, p.119-121] is used to illustrate the difference between dominance act utilitarianism and an orthodox perspective on the agent's ought. Dominance act utilitarianism says that α ought to see to it that A just in case the truth of A is guaranteed by each of the optimal actions available to the agent – formally, that $\bigodot[\alpha\ cstit : A]$ should be settled true at a moment m just in case $K^m \subseteq |A|_m$ for each $K^m \in Optimal_\alpha^m$. When we adopt the orthodox perspective, the truth or falsity of ought statements can vary from index to index. The orthodox perspective is that α should see to it that A at a certain index just in case the truth of A

is guaranteed by the available actions that are optimal given the circumstances in which he finds himself at this index. Horty uses the symbol $\bigoplus$ to denote the orthodox ought operator.

According to Horty, the driver example is due to Holly Goldman [17], and it is also discussed by Humberstone in [18], a paper that sets out in a different context some of the fundamental ideas underlying the orthodox ought defined by Horty.

> "In this example, two drivers are travelling toward each other on a one-lane road, with no time to stop or communicate, and with a single moment at which each must choose, independently, either to swerve or to continue along the road. There is only one direction in which the drivers might swerve, and so a collision can be avoided only of the drivers swerves and the other does not; if neither swerves, or both do, a collision occurs. This example is depicted in Figure 3, where α and β represent the two drivers, K_1^m and K_2^m represent the actions available to α of swerving or staying on the road, K_3^m and K_4^m likewise represent the swerving or continuing actions available to β, and m represents the moment at which α and β must make their choice. The histories h_1 and h_3 are the ideal outcomes, resulting when one driver swerves and the other one does not; collision is avoided. The histories h_2 and h_4, resulting either when both drivers swerve or both continue along the road, represent non-ideal outcomes; collision occurs. The statement A, true at h_1 and h_2, expresses the proposition that α swerves." [15, p.119]

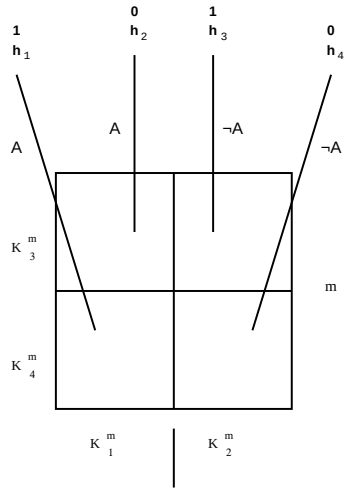

Fig 3. The driving example [15, Fig 5.6]

From the dominance point of view both actions available to α are classified as optimal, i.e. $Optimal_\alpha^m = \{K_1^m, K_2^m\}$, because the sure-thing principle does

not favor one of the actions over the other. Thus, one of the optimal actions available to α guarantees the truth of A and the other guarantees the truth of $\neg A$. Consequently $M, m \not\models \bigodot[\alpha\ cstit : A]$ and $M, m \not\models \bigodot[\alpha\ cstit : \neg A]$. But from the orthodox point of view, we have for example $M, m, h_1 \models \bigoplus[\alpha\ cstit : A]$ and $M, m, h_2 \models \bigoplus[\alpha\ cstit : \neg A]$, because A and $\neg A$ hold for all optimal actions given that agent α does K_1^m or K_2^m, respectively. So, α ought to do A or $\neg A$, depending at the index.

Horty also discusses the so-called Whiff and Poof example, an example with the same logical structure, introduced for example in [19–21]. In this example, there are two agents in the moral universe, who can each push a button or not. If both push the button the overall utility is 10, if neither push their button the utility is 6, and otherwise 0. Both the driver example and the Whiff and Poof example are instances of classical coordination games studied in game theory.

Horty concludes that from the standpoint of intuitive adequacy, the contrast between the orthodox and dominance deontic operators provides us with another perspective on the issue of moral luck, the role of external factors in our moral evaluations [15, p.121]. The orthodox ought can suitably be applied when an agent looks back in time and considers what he ought to have done in a certain situation. For example, when there has been a collision then α might say – perhaps while recovering from the hospital bed – that he ought to have swerved. The dominance ought is looking forward. Though the agent may legitimately regret his choice, it is not one for which he can be blamed, since either choice, at the time, could have led to a collision. The distinction corresponds to what has been called the diagnostic and the decision-theoretic perspective in [22], and can be related to Thomason's distinction between evaluative and judgmental oughts [23].

2.4 Procrastinate's choice

The example of Procrastinate's choices [15, p. 162] illustrates the notion of strategic oughts. A strategy is a generalized action involving a series of actions. Like an action, a strategy determines a subset of histories. The set of admissible histories for a strategy σ is denoted $Adh(\sigma)$. If a strategy σ is not more than a single action K^m at moment m, i.e. $\sigma = \{\langle m, K^m\rangle\}$, Horty simply writes K (assuming m is clear from the context) for $Adh(\{\langle m, K^m\rangle\})$.

A crucial new concept here is the concept of a *field*, which is basically a subtree of the STIT-model which denotes that the agent's reasoning is limited to this range. A strategic ought is defined analogous to dominance act utilitarianism, by replacing actions by strategies in a field. α ought to see to it that A just in case the truth of A is guaranteed by each of the optimal strategies available to the agent in the field – formally, that $\bigodot[\alpha\ cstit : A]$ should be settled true at a moment m just in case $Adh(\sigma) \subseteq |A|_m$ for each $\sigma \in Optimal_\alpha^m$. Horty observes some complications, and says that a 'proper treatment of these issues might well push us beyond the borders of the current representational formalism' [15, p.150].

Horty also uses the example of Procrastinate's choices to distinguish between actualism and possibilism, for which he uses the strategic oughts, and in particular the notion of a field. Roughly, actualism is the view that an agent's current actions are to be evaluated against the background of the actions he is actually going to perform in the future. Possibilism is the view that an agent's current actions are to be evaluated against the background of the actions that he might perform in the future; the available future actions. The example is due to Jackson and Pargetter [24].

> "Professor Procrastinate receives an invitation to review a book. He is the best person to do the review, has the time, and so on. The best thing that can happen is that he says yes, and then writes the review when the book arrives. However, suppose it is further the case that were he to say yes, he would not in fact get around to writing the review. Not because of incapacity or outside interference or anything like that, but because he would keep on putting the task off. (This has been known to happen.) Thus although the best thing that can happen is for Procrastinate to say yes and then write, and he *can* do exactly this, what *would* happen in fact were he to say yes is that he would not write the review. Moreover, we may suppose, this latter is the worst thing which may happen.
> [...]
> According to possibilism, the fact that Procrastinate would not write the review were he to say yes is irrelevant. What matters is simply what is possible for Procrastinate. He can say yes and then write; that is best; that requires *inter alia* that he says yes; therefore, he ought to say yes. According to actualism, the fact that Procrastinate would not actually write the review were he to say yes is crucial. It means that to say yes would be in fact to realize the worst. Therefore, Procrastinate ought to say no." [24, p.235]

Horty represents the example by the STIT-model in Figure 4. Here, m_1 is the moment at which Procrastinate, represented as the agent α, chooses whether or not to accept the invitation: K_1 represents the choice of accepting, K_2 the choice of declining. If Procrastinate accepts the invitation, he then faces at m_2 the later choice of writing the review or not: K_3 represents the choice of writing the review, K_4 another choice that results in the review not being written. For convenience, Horty also supposes that at m_3 Procrastinate has a similar choice whether or not to write the review: K_5 represents the choice of writing, K_6 the choice of not writing. The history h_1, in which Procrastinate accepts the invitation and then writes the review, carries the greatest value of 10; the history h_2, in which Procrastinate accepts the invitation and then neglects the task, the least value of 0; the history h_4, in which he declines, such that a less competent authority reviews the book, carries an intermediate value of 5; and the peculiar h_3, in which he declines the invitation but then reviews the book anyway, carries a slightly lower value of 4, since it wastes his time, apart from doing no one else any good. The statement A represents the proposition that he accepts the invitation; the statement B represents the proposition that Procrastinate will write the review.

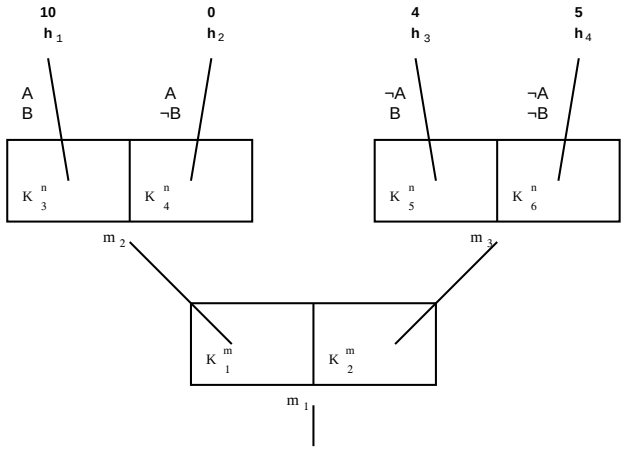

Fig 4. Procrastinate's choices [15, Fig 7.6]

Now, in the possibilist interpretation, $M = \{m_1, m_2, m_3\}$ is the background field. In this interpretation, Procrastinate ought to accept the invitation because this is the action determined by the best available strategy – first accepting the invitation, and then writing the review. Formally, $Optimal_{\alpha}^{M} = \{\sigma_6\}$ with $\sigma_6 = \{\langle m_1, K_1\rangle, \langle m_2, K_3\rangle\}$. And since $Adh(\sigma_6) \subseteq |A|_m$, the strategic ought statement $\bigodot[\alpha\ cstit : A]$ is settled true in the field M. In the actualist interpretation, the background field may be narrowed to the set $M' = \{m_1\}$, which shifts from the strategic to the momentary theory of oughts. but In this case, we have $\bigodot[\alpha\ cstit : A]$ is settled false. It is as if we choose to view Procrastinate as gambling on his own later choice in deciding whether to accept the invitation. However, from this perspective, this should not be viewed as a gamble; an important background assumption – and the reason that he should decline the invitation – is that he will not, in fact, write the review.

3 Discussion

3.1 The examples

The examples in Horty's book are meant to provoke discussion. In this section we raise some issues ourselves.

According to Horty [15, p.57], the gambling example "seems to reflect a real difficulty with the strategy of identifying even a purely utilitarian notion of what an agent ought to do with the notion of what it ought to be that the agent does – at least on the basis of any theory conforming to the standard deontic idea that whatever holds in all the best outcomes is what ought to be. Any such theory would have the result that, in this situation, it ought to be that the agent gambles; after all, gambling is a necessary condition for achieving the best outcome, the outcome with the greatest utility." This observation is the basis of

the formal distinction between a logic for ought-to-be and a logic for ought-to-do. However, the quote also indicates a way in which the two may be identified anyway, namely by leaving the idea that whatever holds in all the best outcomes is what ought to be! This idea of so-called standard deontic logic, a modal logic proposed by von Wright in 1951, has been criticized during the last five decades by many authors for various reasons, and many alternative deontic logics have been proposed. Horty does not discuss the question whether his example is also a problem for these logics. For example, in preference-based deontic logics [25] an obligation for p is formalized by a preference of p over $\neg p$, i.e. $O(p) = p \succ \neg p$. For most preference orderings we do not seem to have the result that, in this situation, it ought to be that the agent gambles. This suggests that the gambling problem may not occur in such settings.

Moreover, Danielsson [3] observes that situations with the same structure as the gambling problem also appear in examples with no actions involved. He discusses an example in which a window may be open or not, and the wind may bring something good or bad if the window is open. Finally, McNamara [**?**] observes that the gambling problem is closely related to an example discussed by Feldman [26]. Feldman imagines that it will rain tomorrow, and that given that it rains it is best for the reporter to predict that fact. However, there is no good reason now to think it will rain rather than that it won't. It is in fact indeterminate without probabilities. So although it is ideal for the reporter to report that it will rain, it does not folow that he should do so.

The driver example is, as Horty observes, a classical coordination game as studied in classical game theory. This raises the question whether the techniques used in game theory are relevant for the analysis of this example. For example, what is the role of Nash equilibria is the analysis of the example? Moreover, is Horty's philosophical study relevant for game theory, and if so, why?

Procrastinate's choices also raises questions. For example, is the notion of a field related to the notion of bounded or limited reasoning as studied in, amongst others, artificial intelligence? Moreover, Horty does not discuss that the notion of strategic ought can be applied to the most famous of all deontic examples, Chisholm's contrary-to-duty paradox, as has been suggested by van der Torre and Tan [27]. Horty observes [15, p.40] that STIT-models can deal with reperational oughts (contrary-to-duty oughts), but he does not discuss the paradox. If in Figure 4 we read A as 'the man tells his neighbors that he will come' and B as 'the man goes to the assistance of his neighbors', and the utility of history h_2 is raised to for example 8, then the STIT-model seems to reflect a variant of Chisholm's paradox:

1. A certain man is obliged to go to his neighbour's assistance;
2. If he goes, he should tell them he will come;
3. If he does not go, then he should tell them that he does not come;
4. He does not go.

3.2 The relation with other motivational concepts

As we mentioned in the introduction, Horty explicitly develops a utilitarian perspective, which means roughly that an act is right or obligatory if it is a best promoter of (social) welfare. Danielsson [3] emphasizes that it is also a consequentialist perspective, which means that an act is right or obligatory if it is a best act for achieving a highest ranked state of affairs. Danielsson also observes that Horty apparently sees no need to discuss possible important differences between rules of rational behaviour, moral rules, and rules of semantics, which makes the whole project somewhat unclear.

The decision-theoretic setting used by Horty to define obligations has also been used to define goals in knowledge-based systems in artificial intelligence, and to define desires in belief-desire-intention (BDI) systems in agent theory, see e.g. [28]. In such settings, the basic distinction between obligations on the one hand and goals and desires on the other hand is that the former are external motivations, whereas the latter are internal motivations of the agent.

Now, some of Horty's examples can also be interpreted in terms of goals and desires. For example, in another example [15, p.49] an agent is discussed who wishes to buy a horse which costs \$15,000 whereas the agent only has \$10,000. The problem in this example is whether the agent should bid \$10,000 for the horse or not. In this example, it seems that we might as well say that the agent desires to buy the horse for \$10,000. Horty mentions that his "characterization of values, or utilities, as abstract, and intended to accommodate a variety of different approaches. It says nothing about what is ultimately taken as a measure of the individual agent's utility – pleasure, mental states of intrinsic worth, happiness, money, an ndex of basic goods" [15, p.38]. These measures seem related to goals and desires.

Horty acknowledges this problem, when he observes that his notion of ought is completely utilitarian, whereas our intuitive idea that an agent α ought to see to it that A often seems to be sensitive to non-utilitarian considerations. Our conception of what we ought to do is often influenced, not only by the utility of the outcomes that might result from our actions, but also by considerations involving a number of additional concepts, such as rights or personal integrity. If Smith makes a promise to Jones, for example, Jones has a right, a claim against Smith, that Smith should keep the promise, even if the outcome that would result from Smith's keeping the promise carries less utility than the outcome that would result if the promise were broken.

Horty's answer to such objections is pragmatic. Such objections, he says, are perhaps too broad to be illuminating. The objection is directed not so much against the analysis itself as against the utilitarian framework within which the analysis is developed. Rather than attempting to model our ordinary, common sense notion of what an agent ought to do, governed as it is by a variety of considerations, he instead restricts his attention only to those oughts generated by considerations of utility. His goal, then is to model only a more limited, utilitarian notion of what an agent ought to do, a notion of what the agent ought to do on the basis of utilitarian considerations alone [15, p.54].

3.3 Logical and technical issues

Since Horty's book is about *logic*, one may expect that the logical repercussions of the semantic definitions in the book are studied in depth. However, the book mentions most logical considerations only briefly.

For instance, it is mentioned that the logic of the composed operator $\bigcirc[\alpha\ cstit : A]$ is similar to the logic of $\odot[\alpha\ cstit : A]$. Horty [15, p.79]: 'Although perhaps already apparent, it is worth noting explicitly that the notion carried by the new operator of what an agent ought to do is logically neither weaker nor stronger than the notion of what it ought to be that the agent does, but incomparable.' Horty demonstrates this incomparability in various ways, since it is directly related to his central thesis about the irreducibility of the ought-to-do operator. But in our opinion, the other part of the claim, i.e., that the first operator is neither weaker or stronger than the second, requires a proof. It is not enough just to observe and prove that the operators both satisfy some properties that are typical for normal modal logics.

The second issue we raise in this section concerns the 'intuitiveness' of the orderings used for actions. This concerns Horty's choice for the definition of an ordering of actions in terms of the ordering of the underlying sets of histories. We argue that this 'lifting' of the ordering of histories to an ordering of actions can also be defined intuitively in another way.

Notice first that in Horty's formalism, the utilities associated with the histories are relevant in as far as they determine relative strengths. So, the absolute values of the utilities have no meaning. In particular the following two models are indistinguishable for Horty's logic:

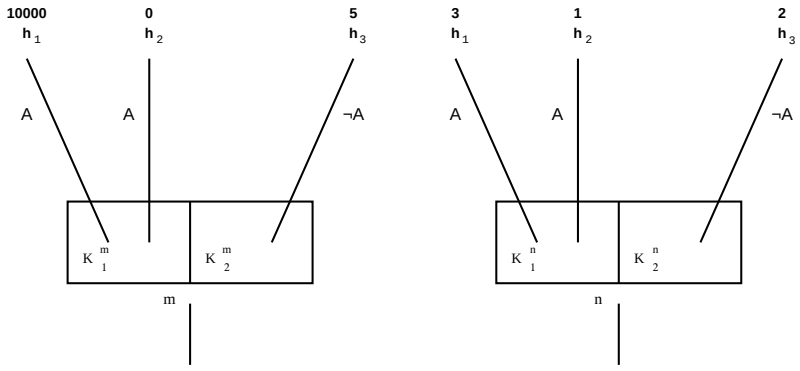

Fig 5. Two models that cannot be distinguished in Horty's logic.

The value of the numbers is only used to decide whether a history is better or worse than another history, which means that any linear order will do. We emphasize this point, because when being presented such example models one is inclined to attach meaning to the absolute values. In particular, when one is used to work in a classical decision theoretic setting, one could easily reason that the high value in the left model will inevitably influence decisions, culminating

in some formulas being evaluated differently. But, for Horty's theory the two choice situations are identical.

The above observation is important for our discussion on the lifting of the ordering of histories to an ordering of actions. Consider the two choice situations sketched in figure 6.

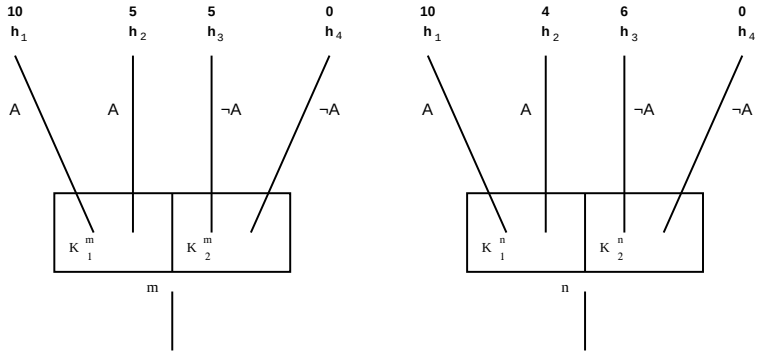

Fig 6. Two more choice situations

In the situation on the left, Horty's ordering on actions gives that action K_1^m is better than action K_2^m, resulting in satisfaction of $\bigodot[\alpha \; cstit : A]$ at m, i.e., the agent ought to perform K_1^m. This is intuitive, since any possible outcome of performing K_1^m is at least as good as any outcome of K_2^m. But in the choice situation n on the right, Horty's ordering gives no decision: there is a possible outcome of K_1^n, namely history h_2, for which there is an outcome of K_2^n, namely h_3, that is better. So, it is not the case that the agent ought to do K_1^n, nor is it the case that he ought to refrain from K_1^n (i.e. do K_2^n). However, we think that in the utilitarian setting put forward by Horty, it is very well possible to defend that action K_1^n is actually *better* than action K_2^n. Let us analyze the information contained in the model. As argued above, we should not attach any meaning to the absolute values of the utilities. Then, all the information that is available is that the highest utility can be reached by doing K_1^n and the lowest by doing K_2^n, and what's more, the highest utility *cannot* be reached by doing K_2^n, and the lowest cannot be reached by doing K_1^n. If an agent is presented with such a choice, he should choose K_1^n, for two good reasons:

1. it is the *only* choice that *might* result in the best possible history, and
2. it is the *only* choice by which he can be *sure* to avoid the worst possible history.

This line of reasoning cannot be countered by claiming that such arguments should account for probabilities concerning the occurrence of separate histories. As said, Horty simply does not consider a logic for situations where the probabilities are known; the logic is only about choices, non-determinism and utilities. It can also not be countered by claiming that there can be (causal) dependencies

between the histories of separate actions. Such information is not represented in the models, meaning that we cannot account for it in the logic.

We do not suggest that the above two conditions are each individually sufficient for concluding that an action is better. But following the line of reasoning, we can define a more subtle way in which an ordering of actions is derived from an underlying ordering of histories. In [29] we show how to define such an ordering, and apply it to the semantics of deontic modalities in a dynamic logic setting. If we apply this ordering to the present STIT-theory, we get a weaker utilitarian ought-to-do-operator (weaker in the sense that it allows more models) that also solves the gambling problem of fig. 2.

4 Conclusion

John Horty's book 'Agency and deontic logic' develops deontic logic against the background of a theory of agency in non-deterministic time. Horty tells a self-contained story without loosing momentum by diving into the conceptual and technical details that are met along the way. He formulates precise and clear, and takes his time to put forward a wealth of concepts and ideas. The book itself is not concerned with the application of the theory to the legal domain, but the relevance of deontic logic and normative reasoning for legal reasoning is well established.

We presented the book to a general AI audience that is familiar with action theories developed in AI, classical decision theory, or formalizations of temporal reasoning. We discussed three representative examples: the gambling paradox, the driving example and Procrastinate's choice. The first illustrates the distinction between ought-to-do and ought-to-be, the second illustrates the distinction between dominance act utilitarianism and an orthodox perspective on the agent's ought. The third example illustrates the distinction between actualism and possibilism. The reader who is intrigued by one of the examples, or the distinctions they illustrate, should read Horty's book for the full story, and for other instructive examples and distinctions.

The book does not study the developed logics in any depth, and there are no axiomatizations. Moreover, Horty does not discuss why utilities should be used for obligations, in contrast to for example goals and desires. Finally, the relation between his logic and related work in for example logics of action in AI, classical decision theory, and temporal logic is not studied. This may be judged as an omission, but also as an opportunity.

In this review we indicated how classical decision trees can be related to STIT models, and we have given an alternative way to lift the ordering on histories to a dominance relation on actions. We believe that the book is a good starting point for other comparisons that relate philosophical logic to theories developed in AI. We strongly recommend anyone interested in the philosophical and logical aspects of reasoning about oughts, agency and action to get hold of a copy of this book.

Acknowledgements

Thanks to Paul McNamara for very useful comments on an earlier version of this review.

References

1. Bartha, P.: A review of john horty's 'agency and deontic logic'. Notre Dame Philosophical Reviews (2002.02.01) (2002) ndpr.icaap.org.
2. Czelakowski, J.: John f. horthy, agency and deontic logic. Erkenntnis **58**(1) (2003) 116–126
3. Danielsson, S.: A review of john horty's 'agency and deontic logic'. The Philosophical Quarterly (2002) 408–410
4. McNamara, P.: Agency and deontic logic by john horty. Mind **112**(448) (2003)
5. Wansing, H.: A review of john horty's 'agency and deontic logic'. Journal of Logic, Language and Information (2003) to appear.
6. Savage, L.: The Foundations of Statistics. John Wiley and Sons (1954)
7. Clarke, E., Emerson, E., Sistla, A.: Automatic verification of finite-state concurrent systems using temporal logic specifications. ACM Transactions on Programming Languages and Systems **8**(2) (1986)
8. Emerson, E.: Temporal and modal logic. In Leeuwen, J.v., ed.: Handbook of Theoretical Computer Science, volume B: Formal Models and Semantics. Elsevier Science (1990) 996–1072
9. Belnap, N., Perloff, M., Xu, M.: Facing the future. Oxford University Press (2001)
10. Castañeda, H.N.: The paradoxes of deontic logic: the simplest solution to all of them in one fell swoop. In Hilpinen, R., ed.: New Studies in Deontic Logic: Norms, Actions and the Foundations of Ethics. D. Reidel Publishing Company (1981) 37–85
11. Castañeda, H.N.: Aspectual actions and davidson's theory of events. In E. LePore, B.M., ed.: Actions and Events: Perspectives on the Pholosophy of Donald Davidson. Basil-Blackwell (1985) 294–310
12. Wright, G.v.: Deontic logic. Mind **60** (1951) 1–15
13. Hanks, S., Dermott, D.M.: Default reasoning, nonmonotonic logics, and the frame problem. In: Proceedings of the National Conference on Artificial Intelligence (AAAI86), Morgan Kaufmann Publishers (1986) 328–333
14. Kautz, H.: The logic of persistence. In: Proceedings of the National Conference on Artificial Intelligence (AAAI86), Morgan Kaufmann Publishers (1986) 401–405
15. Horty, J.: Agency and Deontic Logic. Oxford University Press (2001)
16. Chellas, B.: The Logical Form of Imperatives. PhD thesis, Philosophy Department, Stanford University (1969)
17. Goldman, H.: Dated rightness and moral imperfection. The Philosophical Review **85** (1976) 449–487
18. Humberstone, I.: The background of circumstances. Pacific Philosophic Quarterly **64** (1983) 19–34
19. Gibbard, A.: Rule-utilitarianism: merely an illusionary alternative? Australasian Journal of Philosophy **43** (1965) 211–220
20. Sobel, J.: Rule-utilitarianism. Australasian Journal of Philosophy **46** (1968) 146–165
21. Regan, D.: Utilitarianism and Co-operation. Clarendon Press (1980)

22. Torre, L.v.d., Tan, Y.: Diagnosis and decision making in normative reasoning. Artificial Intelligence and Law **7** (1999) 51–67
23. Thomason, R.: Deontic logic as founded on tense logic. In Hilpinen, R., ed.: New Studies in Deontic Logic. D. Reidel Publishing Company (1981) 165–176
24. Jackson, F., Pargetter, R.: Oughts, options and actualism. Philosophical Review **99** (1986) 233–255
25. Torre, L.v.d., Tan, Y.: Contrary-to-duty reasoning with preference-based dyadic obligations. Annals of Mathematics and Artificial Intelligence **27** (1999) 49–78
26. Feldman, F.: Doing the Best We Can. D. Reidel Publishing Company (1986)
27. Torre, L.v.d., Tan, Y.: The temporal analysis of chisholm's paradox. In: Proceedings of 15th National Conference on Artificial Intelligence (AAAI'98). (1998) 650–655
28. Lang, J., Torre, L.v.d., Weydert, E.: Utilitarian desires. Autonomous Agents and Multi-Agent Systems **5**(3) (2002) 329–363
29. Broersen, J., Dastani, M., Huang, Z., Torre, L.v.d.: Trust and commitment in dynamic logic. In Shafazand, H., Tjoa, A.M., eds.: Eurasia-ICT 2002: Information and Communication Technology. Volume 2510 of Lecture Notes in Computer Science., Springer (2002) 677–684