Contextual Deontic Logic: Normative Agents, Violations and Independence

Leendert van der Torre [a]

[a] *Department of Artificial Intelligence, Vrije Universiteit De Boelelaan 1081a, 1081 HV Amsterdam, the Netherlands*
E-mail: torre@cs.vu.nl

In this paper we discuss when and how to use deontic logic in multi agent systems. Our central question is how to proceed once a norm has been violated or defeated, a key issue of deontic logic applications in multi agent systems. To bridge the logical analysis of norms in philosophy with applications in agent theory, we propose a practical approach based on violation contexts and independence statements. In particular, we introduce a combination of two traditional deontic logics, which we extend with so-called deontic and factual independence assumptions. We show how different notions of violability and defeasibility can be encoded in the logic by defining different ways in which independence assumptions are derived from the explicit manner of presentation. We also show how our approach can be used to give a new analysis of several notorious paradoxes of deontic logic.

1. Introduction

Norms play an important role in multi agent systems, but it is an open question when agents reason about these norms and when and how they represent them explicitly. The obvious candidate to represent norms is deontic logic, the logic of norms, obligations, prohibitions and permissions. It is usually presented as a branch of modal logic, in which $\bigcirc p$ is read as 'p ought to be (done),' Fp as 'p is forbidden to be (done)' and Pp as 'p is permitted to be (done).' We believe that deontic logic should be used in agent logics and architectures when (1) norms can be violated and (2) the agents reason explicitly about these violations *and their consequences.* What to do when violations of norms occur has to be addressed in the multi agency area, in particular if we consider applications where agents act within a normative system.

We call agents that can reason about norms and obligations *normative agents* [23,75]. We believe that norms and their violability are relevant for several aspects of multi agent systems (such as deliberation, autonomy and control), as explained in Section 2. Our contribution in this paper is twofold:

- Our starting point is a combination of two classical deontic logics. In contrast to other approaches, we do not further complicate the definition of obligation, but instead we introduce independence statements. Note that the distinction between operators and independence statements has already been tested and proven successful in reasoning about uncertainty, in particular in Bayesian networks [50]. By introducing notions of independence we also introduce notions of context, as is explained in detail in Section 3.3.
- We also show how our approach can be used to give a new analysis of several notorious paradoxes of deontic logic. Our claim is that once the violation contexts and independence assumptions are made explicit, the deontic paradoxes do not arise. Since we test our formal systems against the benchmark examples given in the philosophical literature, we do not introduce new examples of norms in multi agent systems.

The main goal of this paper is to bridge work in philosophical logic to multi agent systems applications. We focus on what is usually considered to be the main problem of reasoning about norms, recovering from violations. Another issue discussed in philosophical logic which is important for multi agent systems applications is the relation between so-called ought-to-be and ought-to-do obligations [32]. In this paper we use Boutilier's simple but useful distinction between controllable and uncontrollable propositions [10]. We do not discuss the emergence of norms in multi agent systems, or the use of norms in the emergence of coordination, cooperation and trust.

This paper is organized as follows. In Section 2 we discuss which aspects of multi agent systems the paper formalizes. In Section 3 we discuss the problems which have been identified in formalizing violability. In Section 4 we discuss the relation between deontic and contextual reasoning and we introduce a simple contextual deontic logic. In Section 5 we discuss deontic independence and we analyze two deontic paradoxes. Finally in Section 6 and Section 7 we show how independence assumptions can be derived from the explicit manner of representation, which also shows the distinction between deontic independence and world-based factual independence.

2. Normative agents and multi agent systems

We first discuss the role of norms (Section 2.1) and deontic logic (Section 2.2) in agent theory. We then discuss norms in agent architectures (Section 2.3) and agent specification and verification languages (Section 2.4). Finally in Section 2.5 we compare logics of obligations with logics of desires.

2.1. Agents, multi agent systems and norms

Although there is no consensus on what exactly an 'agent' is, agent theory has turned into a major branch of artificial intelligence and computer science. This is in particular due to the following two reasons for respectively individual autonomous agents and multi agent systems.[1]

Cognitive concepts like beliefs, desires and intentions are proper abstraction tools for characterizing the behavior of agent systems – just like an object and a method are the right abstractions for object-oriented programming. In particular it has been found that the behavior of any agent is mainly governed by the specific way in which it handles the rational balance between its beliefs, desires and intentions, where beliefs describe its informational state, desires its motivational state, and intentions its deliberative state [11,16,54,59].

Social concepts like coordination and cooperation are proper abstraction tools for characterizing the behavior of multi agent systems. Interest in multi agent systems – as well as in normative agents – has risen in accordance with the growing interest in the internet, auctions, electronic commerce and the new economy.

[1] This transposition of concepts from cognitive and social sciences to computer science is very useful for the latter, but it does *not* imply that the developed computational tools can therefore also be used in these cognitive and social sciences (though it also not excluded!). Criticism based on the fact that psychological experiments have shown that humans do *not* reason with these concepts from folk psychology or according to the developed computational models therefore miss the point of the transposition. In other words, the success of this intentional stance [18] is not measured in psychological validity, but in for example its ability to specify and verify complex computer systems efficiently and effectively. This success just can be *explained* by the psychological fact that humans are used to ascribe mental states to artificial systems. Note that this kind of transposition of foreign concepts into computer science has been very successful and has even led to whole new research areas, such as e.g. the transposition of biological models into neural and evolutionary computing, or the transposition of physical models into quantum computing.

In agent theory, the gap between individual autonomous agents and multi agent systems – i.e. the gap between cognitive and social concepts – is an important open problem. Normative agents bridge an agent and the multi agent system it is part of, in the sense that the *cognitive* concept of obligation is the main building block of more complex *social* notions like coordination, cooperation, trust and reputation. An autonomous agent may deliberate which norms to adopt or to violate, and a system may vote for a norm or such norms may emerge spontaneously. This role of norms and obligations as the bridge between individual agents and multi agent systems is explained by the compositional model of multi agent systems in Figure 1. This compositional multi agent system consists

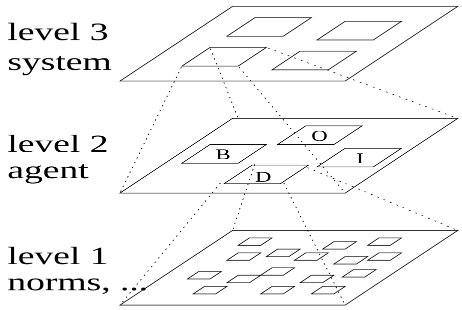

Figure 1. Compositional Multi Agent System

of three levels: components within an agent such as its obligations (level 1), the agent themselves (level 2), and the multi agent system or society (level 3). Questions can be asked for each level, for example how obligations interact with each other, how obligations interact with desires, or how normative agents interact. Moreover, questions can be asked for the relations between these levels:

Between level 1 and level 2 are the issues of decision-making (also called practical reasoning). How do the agent's beliefs, obligations, intentions and desires lead to agent behavior? How does the agent's behavior influence these attitudes?

Between level 2 and 3 is the micro-macro dichotomy. How does behavior of agents lead to emergent system properties like social norms (i.e. norms of the society), reputation of agents, conventions, social or game theoretic equilibria? How does the emergence of these properties influence the behavior of the agent?

2.2. Agents, violations and deontic logic

It is an open question when and how deontic logic can or should be used as a computational logic for multi agent systems in agent architectures or in agent programming, specification and verification languages. The motivation of the previous section to use norms in multi agent systems is that norms can bridge the autonomous agent level and the multi agent system level, an instance of the so-called micro-macro dichotomy. Another motivation to represent norms in a multi agent system is the fact that in the social sciences norms have been identified as crucial tools for important issues in multi agent systems such as e.g. coordination, cooperation (norms of reciprocity in evolutionary game theory [5]) and trust (within reason [29]). We think the answer to this question *when* deontic logic should be used in multi agent systems can be found in the history of deontic logic in computer science.

We believe that deontic logic should be used in agent logics and architectures when norms can be violated and the agents have to reason explicitly about these violations and their consequences [37,36]. In the eighties deontic logic – which thus far had been used by philosophers to analyze the structure of the normative use of language – had a revival, when it was discovered by computer scientists that this logic can be used for the formal specification and verification of a wide variety of topics in computer science (for an overview and further references see [80]). In all these applications, norms can be violated and the system has to reason explicitly about these violations and their consequences. For example, deontic logic can be used in qualitative decision theory to represent the agent's goals [51,10,38], in computer security to analyze and specify security issues for electronic commerce [24], and in planning and scheduling problems to formally specify soft constraints. The advantage of soft constraints is that norm violations do not create an inconsistency in the formal specification, in contrast to violations of hard constraints.

In this paper we therefore study the violability of norms, and in particular how to proceed once a norm has been violated. Violability of norms is relevant in multi agent systems for the following three reasons.

Deliberation. Most applications can do with simple or reactive agents, some more advanced applications may need complex agents with some deliberation, for example BDI agents [11,54], and some futuristic applications may need to do some norm-based social deliberation. That is, futuristic multi agent systems may contain normative or social agents [14].

Autonomy. If the agents cannot violate the norms, then these agents are not truly autonomous, cf. [17,19,20]. Given the importance of the concept of autonomy in multi agent systems, it may be expected that agents should be able to break their norms. In electronic commerce this type of autonomy is of great practical relevance, because experiments in the contract net protocol have shown that systems are more efficient when commitments can be broken (i.e. when there is leveled commitment [57]). Moreover, here it is important how to proceed once a norm has been violated. For example, if the delivery time is over due the responsible agent might be obliged to pay the extra transport and warehousing costs that result from the delay. If the violation is not too serious, or was not intended by the violating party, then the contracting parties usually do not want to consider this as a breach of contracts, but simply as a disruption in the execution of the contract that has to be repaired.

Control. Norms together with delegation can control the behavior of a multi agent system. Consider an agent that needs cooperation of other agents to realize its goals. It can either negotiate with the other agents, or, if there is the necessary organizational social structure, it can command them. The latter is often the most efficient. However, in general the agent does not only delegates some actions, it also delegates some of the planning problem. That is, it delegates a sub-goal when it does not know exactly what the other agent's capabilities are, when it does not know how the sub-goal it delegates can be achieved, and when it does not know whether the sub-goal is achievable.[2] If norms cannot be violated or agents cannot reason about violations, then it is difficult to control this kind of social interaction.

In the following two subsections we discuss the role of norms in agent architectures and in agent specification and verification languages, focusing on agent decision making.

[2] In some applications an agent A cannot delegate a (sub-)goal to another agent B unless A has reasons to believe that B has capabilities to achieve the goal. However, this does not hold for all applications. This is in accordance with examples from human organizations where, for example, tasks are distributed (or delegated) top-down, e.g. from the boss to the head of the departments to the employees. Here, obviously, the boss may not be aware of the capabilities of all his employees.

2.3. Normative agent architecture: BOID architecture

Recently several agent architectures have been proposed which contain explicit representations of norms, such as for example Dignum *et.al.*'s extension of BDI architectures with obligations [19,20], and Broersen *et.al.*'s extension of Thomason's BDP [63] with obligations in the BOID architecture [12]. In this section we briefly discuss some elements of the latter to illustrate the motivation to introduce norms, why norms can be violated, and why it is important to know how to proceed once a norm has been violated.

The BOID architecture is an agent system that can be decomposed into logical sub-systems – for amongst others beliefs, obligations, intentions and desires – and a coherence strategy to coordinate or control these sub-systems. The motivation of this decomposition is to reduce the complexity of the overall system by a divide and conquer strategy, a well known approach of component-based software engineering. The BOID decision model is based on an agent that is overloaded with input, and typically lives in a complex and noisy environment. Most cognitive decision models are based on an attempt to reach goals, satisfy desires, fulfill obligations etc, for example on goal-based planning or on variants of decision theory like qualitative decision theory [51,10], on cognitive models like belief-desire-intention or BDI models [11,16,54],[3] or on social concepts like obligations and norms, as in deontic action programs [23]. In the BOID architecture the agent's main problem is not finding out the way to reach his goals, satisfy his desires or fulfill his obligations, but which of these impulses he should follow given his beliefs and intentions. That is, his main problem is to resolve his conflicts.[4] This conflict resolution mechanism is the BOID architecture's coherence strategy to coordinate or control its sub-systems of beliefs, obligations, intentions and desires.

[3] For completeness we mention some differences between this BOID architecture and the well known Rao and Georgeff's BDI approach, see e.g. [54,55]. These differences are derived from Thomason's BDP logic, which is based on two elements: the BD component generates goals from beliefs and desires, and the P component finds plans for these goals. First, in BDI logics decisions are restrained by previous intentions. Second, in BDI logics planning and scheduling have not been incorporated, whereas it is a crucial part in BDP. Third, whereas BDI only discusses absolute beliefs and desires, BDP focuses in particular on conditional ones. Fourth, whereas BDI logics are based on modal logic, Thomason's BD logic is based on Reiter's normal default logic.

[4] This may be compared with psychological theories in which thoughts fight for attention.

In the BOID architecture *internal conflicts*, originating within each component for beliefs, obligations, intentions or desires, can be distinguished from *external conflicts*, originating between these components. Internal conflicts can be distinguished into four unary subtypes (B ; O ; I; D). A typical example of conflicting obligations is: *It is obligatory to be honest. It is obligatory to be polite. If I am honest, then it would be impolite.* External conflicts can be distinguished into six binary conflict subtypes (BO ; BI ; BD ; OI; OD ; ID), and four ternary conflict types (BOI; BOD; BID; OID) and one quadruplicate conflict type (BOID). In [12] the followings benchmark examples are given.

Conflict BO: *It is obligatory to see my mother-in-law this weekend. But I think I have no time to go.*

Conflict OI: *It is obligatory to see my mother-in-law this weekend. But, I already have a plan to go to a conference. If I go to the conference, I cannot go to see my mother-in-law.*

Conflict OD: *It is only permitted to smoke in a smoking area. I desire to smoke in my office. However, my office is a non-smoking area.*

Conflict BOI: *It is obligatory that if I smoke, I smoke in a smoking area. I intend to smoke. However, I know that there is no smoking area here.*

Conflict BOD: *It is obligatory that if I smoke, I smoke in a smoking area. I desire to smoke. However, I know that there is no smoking area here.*

Conflict OID: *I intend to go to a conference. I desire to stay in a luxury hotel. However, it is obligatory for me that if I go to the conference, I do not stay in a luxury hotel.*

Conflict BOID: *I intend to go to a conference. It is obligatory for me not to spend too much money for the conference. Namely, either I should pay for a cheap flight ticket and stay in a better hotel, or I should pay for an expensive flight ticket and stay in a budget hotel. I desire to stay in a better hotel. But, I know that the secretary has booked an expensive flight ticket for me.*

Summarizing, the BOID architecture is a component-based architecture in which a coherence strategy resolves conflicts between components for beliefs, obligations, intentions and desires. The list of benchmark examples illustrates that in a BOID architecture norms will often be violated. Moreover, if a norm is violated, then it is important to know how to proceed, because other beliefs, obligations, intentions and desires may be invoked. Consequently norms have to be represented explicitly, for which deontic logic can be used.

2.4. Agent specification and verification languages: BOID logic

Agent logics that describe cognitive agents at the highest abstraction level are used for two different tasks, namely [59]:

- to specify agents in an abstract manner or, alternatively,
- to verify agents implemented in some other programming language.[5]

For example, consider an intelligent computer system that is asked to make travel arrangements for a scientist to go to a conference, including for example booking a flight and reserving a hotel room. Belief-Desire-Intention or BDI logics can specify such a system, or verify an existing one, by formally stating that the system believes that direct flights are not available, it wants to book a cheap flight, and it intends to book a hotel room close to the conference center. The success of this intentional stance, as Dennett [18] calls it, can be explained by the fact that humans are used to ascribe mental states to artificial systems, and it thus facilitates human-computer interaction. However, it is not clear how existing BDI logics can be used to specify that if the system desires a cheap room and it desires a room close to the conference center, then in absence of cheap close rooms it prefers a cheap one over a close one (conflicting objectives), that it desires a cheap indirect flight but it has been obliged to order a direct flight (obligations overruling desires), or that it has delegated the obligation to arrange the flight to another computer system (communication). One of the missing elements to formalize these specifications is the concept of *obligation*. We call the extension of BDI logic with obligations BOID logic.

Verification and specification languages which have been developed thus far are only concerned with agent decision making, but not with the micro macro

[5] The verification task assumes that the behavior of agent systems may be explained in terms of cognitive concepts, regardless whether they are built using these concepts. An analogous assumption in psychology would be that human behavior can be explained in terms of psychological concepts though humans do not use these concepts either consciously or subconsciously, and no regions in the brain can be found which are assigned to reason about them. For rational agents it has even been *proven* that formal systems do not have to be descriptive, in the following sense. Classical decision theory identified with Savage's foundations of statistics [58] is based on the idea that if an agent acts according to certain (innocent looking) postulates, then its behavior can be understood *as if* it maximizes its expected utility based on a subjective probability distribution and utility function. Unfortunately, such so-called representation theorems are still missing for cognitive agents.

dichotomy. That is, in Figure 1 these languages are only concerned with the relation between level 1 and level 2 but not with the relation between level 2 and level 3. The reason is that the former has to be better understood before a formal analysis of the latter can be made. Moreover, it is at this moment not clear whether and how any emergent property can be used for specification and verification tasks.

2.5. The logical difference between obligations and desires

As agent logics have studied desires in some detail, in particular in the context of BDI logics, the question may be raised in which *logical* aspects obligations differ from desires. A problem of this comparison is that in the traditional logical (philosophical) literature is has been the other way around, the concept of obligation has been studied intensively (in deontic logic to formalize e.g. legal and moral reasoning) whereas the logic of desires has been mostly ignored (with exception maybe the logic of decision [33] and the logic of preference [76]). The following differences are usually mentioned:

- Desires are internal, originating from the agent itself, whereas obligations come from the outside of the agent, and are usually imposed on him (see [38] for a discussion). Consequence:
  - ∗ The dynamics of desires is different from the dynamics of obligations [75].
- Desires are studied in their relation to intended actions, whereas obligations are studied in relation to each other (see [38,74,40] for a discussion).
- Desires can conflict (due to for example conflicting objectives, see [39]) but obligations cannot. Consequences:
  - ∗ In case of violation obligations must be revised (the issue of contrary-to-duty reasoning).
  - ∗ Obligations are studied in relation with exceptions, and different types of defeasibility have been studied (e.g. obligations under uncertainty or prima facie obligations, see also Section 7).

Several links have been established between the logic of desires and the logic of obligations, such as for example obligations and desires both have been formalized with utilitarian semantics [34,40]. However, the exact relation between these two logics is still an open problem.

3. Philosophical problems with practical implications

In the previous section we have given and discussed two arguments why norms are interesting objects of study for multi agent systems. The first argument is that norms bridge cognitive and social concepts, that is, they bridge the autonomous agent level and the multi agent system level. Norms of individual agents play a role in the agent's decision-making, and norms in a multi agent system are crucial tools for coordination, cooperation and trust. The second argument is that norms which can be violated play an important role in deliberation and autonomy of an agent, and control of a multi agent system. This discussion raises for example the question of how norms play these different roles, or how they can be used in these different circumstances. In this paper we restrict ourselves to what deontic logicians consider to be the hardest question of reasoning about norms: how to proceed when a norm is violated?

There is a huge philosophical literature on deontic logic, but there are hardly any up-to-date surveys, there is no consensus or commonly accepted standard theory, nor a commonly accepted list of open problems or benchmark examples (but see [44,69]), most traditional logics lack concepts like agents, time, and actions to represent multi agent systems, most proposals in artificial intelligence and computer science are relatively complicated (see e.g. [51,10,70]), and most proposed formalisms seem to have counterintuitive properties. This at least partly explains why deontic logic has not been used in multi agent systems like for example logics of beliefs, desires and intentions.

In our opinion, the main reason why deontic logic is not widely used in multi agent systems is that a practical deontic logic not only has to deal with violations but also with exceptions. In other words, we not only need a traditional deontic logic, but we need a *defeasible* deontic logic. Defeasible deontic logics are usually complicated, although a simple traditional approach is *nearly* good enough to formalize defeasible obligations. In this section we discuss this traditional deontic logic, where it goes wrong, and we discuss some problems which have been discussed in a more in depth analysis of violability.

In Section 3.1 we introduce Hansson's deontic logic DSDL3, and we discuss how it can be used in multi agent systems. In Section 3.2 we argue why this logic cannot be used as a defeasible deontic logic in the sense that it does not formalize exceptions of norms. In Section 3.3 and 3.4 we informally introduce our analysis of reasoning about violations, based on contexts and independence.

3.1. Dyadic standard deontic logic 3 (DSDL3)

So-called ‘Standard’ Deontic Logic (SDL) is a normal modal system that is closed under the inference rules modus ponens $\frac{\alpha, \alpha \to \beta}{\beta}$ and necessitation $\frac{\vdash \alpha}{\vdash \bigcirc \alpha}$ and contains the propositional tautologies and the axioms $K : \bigcirc(\alpha \to \beta) \to (\bigcirc\alpha \to \bigcirc\beta)$ and $D : \neg(\bigcirc\alpha \wedge \bigcirc\neg\alpha)$. However, the representation of conditional obligations by either $p \to \bigcirc q$ or $\bigcirc(p \to q)$ has been criticized, because both representations cannot satisfactorily formalize contrary-to-duty reasoning. For example, the following formalization of Forrester's paradox [25] in SDL is inconsistent, whereas intuitively it is consistent (see [68,73] for a detailed discussion):

- ‘Smith should not murder Jones’ $\bigcirc\neg k$,
- ‘if he murders him, then he should do it gently’ $k \to \bigcirc g$,
- ‘Smith murders Jones’ k and
- ‘gentle murder implies murder’ $\vdash g \to k$.

Hansson [27] proposed a simple, but popular dyadic deontic logic which he called Dyadic Standard Deontic Logic 3 (DSDL3). DSDL3 was developed to formalize contrary-to-duty (CTD) reasoning as it occurs in Forrester's and Chisholm's notorious deontic paradoxes [15,25,52,53,71]. The obligation for α given γ is true, which we write as $\bigcirc_\gamma \alpha$ (where Hansson writes $\bigcirc(\alpha \mid \gamma)$) if the preferred γ worlds are α worlds. For example, it can consistently represent Forrester's paradox by ‘Smith should not murder Jones’ $\bigcirc\neg k$, but ‘if he murders him, then he should do it gently’ $\bigcirc_k g$ with $\vdash g \to k$. Note that the logic does not contain factual (non-modal) sentences or nested modalities.

Definition 1 ($\models_{DSDL3}$).[27] Let $M = \langle W, \leq, V\rangle$ be a preference model with W a set of worlds, $\leq : W \times W$ a reflexive, transitive, and totally connected relation, and V a valuation function. We have $M, w \models \bigcirc_\gamma \alpha$ iff there are no γ worlds, or there exists a γ world w' such that for all worlds $w'' \leq w'$ we have $M, w'' \models \gamma \to \alpha$. Permission is defined by $P_\gamma \alpha =_{def} \neg \bigcirc_\gamma \neg\alpha$. We write $\bigcirc\alpha$ for $\bigcirc_\top \alpha$, etc.

In the truth conditions of $\bigcirc_\gamma \alpha$ there is no reference to the so-called actual world w, such that we have $M, w \models \bigcirc_\gamma \alpha$ if and only if $M \models \bigcirc_\gamma \alpha$, where as usual $M \models \phi$ iff for all worlds w we have $M, w \models \phi$. Hansson's original definition prohibited infinite descending chains, such that minimal worlds always exist and the definition can be slightly simpler (as noted and discussed by Lewis [41]). In deontic logic and in non-monotonic logic there is a lot of work on generalizations of Hansson's logic, for example with facts and with nested conditionals, see e.g. [9,10,79]. DSDL3 was first axiomatized by Spohn.

Definition 2 ($\vdash_{DSDL3}$).[60, p.240] Let $\vdash$ be the inference relation of propositional logic. DSDL3 is axiomatized by a correct and complete set of propositional axioms and rules, extended with:

- $\bigcirc_\alpha \alpha$
- $\neg \bigcirc_\gamma \bot$ if $\not\vdash \gamma \leftrightarrow \bot$
- $\bigcirc_\gamma(\alpha \wedge \alpha') \leftrightarrow \bigcirc_\gamma \alpha \wedge \bigcirc_\gamma \alpha'$
- $\bigcirc_\gamma \alpha \leftrightarrow \bigcirc_{\gamma'} \alpha'$ if $\vdash \alpha \leftrightarrow \alpha'$ and $\vdash \gamma \leftrightarrow \gamma'$
- $P_\gamma \alpha \rightarrow (\bigcirc_{\gamma \wedge \alpha} \alpha' \leftrightarrow \bigcirc_\gamma(\alpha \rightarrow \alpha'))$

Concepts of proof, provability or theoremhood, refutability, inference etc. are defined for DSDL3 in the usual fashion.

If we want to use DSDL3 in a multi agent system, then it has to be extended in several ways. First, obligations must be labeled with normative system, agent or group identifiers. For example, $\bigcirc_q^{ns,a}(p)$ may be read as 'under normative system ns, agent a is obliged to p if q', $\bigcirc_q^{ns,\{a,b\}}(p)$ may be read as 'under normative system ns, the group of agents $\{a, b\}$ is obliged to p if q', $\bigcirc_q^{ns,a,b}(p)$ may be read as 'under normative system ns, agent a is obliged toward agent b to p if q', etc. These kinds of identifications of agents and normative systems has always been implicitly assumed in the deontic logic literature, see e.g. [1]. It is usually not made explicit, because it does not seem to raise any interesting *logical* questions. There may of course be practical issues involved, for example how an obligation of a group is distributed over the members of the group.

Second, we have to distinguish between ought-to-be and ought-to-do obligations. Pearl [51] observes that in contrast to most decision-theoretic methods, in deontic logic it is usually assumed that an agent can see to any proposition p. Boutilier [10], who proposes to use DSDL3 to express conditional preferences in his logic of qualitative decision theory, proposes to partition the propositional atoms into so-called controllable and uncontrollable propositions (called decision variables and parameters by Lang [38]). An example of a controllable proposition is 'you have your umbrella with you' and an example of an uncontrollable proposition is 'it is raining'. Likewise, in the multi agent case each agent has a set of controllable atoms.

With these first two extensions of agents identifiers and for each agent sets of controllable propositions, DSDL3 can not only formalize what an agent has to do when it violates an obligation, i.e. how to recover from its violations, but also what it has to do when some other agent violates an obligation. For example, when p is a proposition controlled by agent a and q is a proposition controlled by agent b, then $\bigcirc^{ns,a}(p) \wedge \bigcirc_{\neg p}^{ns,b}(q)$ expresses that agent a has to see to it that p but if it does not do so, then agent b has to see to it that q. In other words, it is not agent a who has to recover from its own violation, but this is left to agent b. In the other examples of this paper the normative system, agents and the distinction between controllable and uncontrollable propositions are clear from their context, and to keep the formalism as simple as possible we leave all of them implicit.

Third, a method has to be developed how to detach absolute obligations from dyadic ones. Boutilier discusses that DSDL3 does not satisfy the following principle of *factual detachment.*

$$O_\gamma \alpha \wedge \gamma \rightarrow O\alpha$$

Instead of using one of the more complicated proposals in the literature, Boutilier proposes a simple solution based on the observation that to determine preferences based on certain actual facts, we consider only the *most ideal* worlds satisfying those facts, rather than *all* worlds satisfying those facts. Let KB be a knowledge base containing statements of conditional preference and actual facts. Given that the facts actually obtain, the ideal situations are those most preferred worlds satisfying KB. Boutilier thus simply asks for those propositions α such that

$$\models_{DSDL3} \bigcirc_{KB}(\alpha)$$

Boutilier's approach can be traced back at least to [77], where von Wright discusses whether monadic obligations $\bigcirc\alpha$ should be represented by either $\bigcirc(\alpha|\top)$ or by $\bigcirc(\alpha \,|\, S)$, where S stands for the actual circumstances. Boutilier further observes that this is precisely the preliminary scheme for conditional default reasoning he proposed in earlier work [9], and that this mechanism has to be further extended to deal with irrelevant factual information. In this paper we also use this simple solution, and we discuss the issue of irrelevance later in relation to so-called independence statements.

3.2. The confusion between violability and exceptionality

Recently several authors have suggested or even explicitly argued, following [1], that $\bigcirc_\gamma \alpha$ of DSDL3 can be interpreted as a defeasible obligation in the following sense. The formulas $\bigcirc\neg f \wedge \bigcirc_a f$ and $\bigcirc\neg f \wedge P_a f$ imply that a are exceptional circumstances in which the obligation for $\neg f$ is defeated or overridden by the more specific obligation for f. For example, in this logic we can consistently represent specificity structures like 'you should not eat with your fingers, but if you are served asparagus you are permitted to do so' [31]. The motivation for this suggestion is as follows.

- $\bigcirc\neg f \wedge P_a f$ as well as $\bigcirc\neg f \wedge \bigcirc_a f$ are consistent;
- The logic of obligation lacks strengthening of the antecedent, e.g. $\bigcirc\neg f$ does not imply $\bigcirc_a \neg f$;
- Structural similarities between $\bigcirc_a f$ and 'a defeasibly implies f' $a \Rightarrow f$, defined as 'the most normal a worlds are f worlds.' Given preference models $M = \langle W, \leq_\Rightarrow, V\rangle$ with $\leq_\Rightarrow$ defined just like $\leq$, we have $M, w \models a \Rightarrow f$ iff $\forall w' \in W : M, w' \models \neg a$ or $\exists w' \in W : M, w' \models a$ and $\forall w'' \leq w' : M, w'' \models a \rightarrow f$.

We think this argument is mistaken for the following reason (see e.g. [43,53] for similar counterarguments). All minimal worlds of a model M such that $M \models \bigcirc\neg f \wedge P_a f$ satisfy $\neg a$. This represents that a is false in the ideal worlds, and a is thus a violation. This is counterintuitive, since there is no prohibition for a. The formula $\bigcirc\neg f \wedge P_a f$ implies $\bigcirc\neg a$ which is contrary to intuition; all that should (maybe) be derivable is 'normally not a' $\top \Rightarrow \neg a$.

For example, consider the asparagus example. The two sentences intuitively do not imply that it is forbidden to eat asparagus, only (maybe) that eating asparagus is exceptional. However, it is not implied that a is exceptional. There is not even a representation of normality or any other representation of uncertainty in the semantics. The same argument holds for $\bigcirc\neg f \wedge \bigcirc_a f$, since it implies $\bigcirc\neg f \wedge P_a f$. In the next subsection we give more examples of this distinction.

A defeasible obligation 'a defeasibly implies $\bigcirc f$' can better be formalized by $a \Rightarrow \bigcirc f$.[6] From $\top \Rightarrow \bigcirc\neg f$ and $a \Rightarrow \bigcirc f$ we cannot derive $\bigcirc\neg a$ but we can

[6] For example using SDL, if $M = \langle W, \leq_\Rightarrow, R, V\rangle$ with accessibility relation $R : W \times W$, $M, w \models \bigcirc\alpha$ iff $\forall w' \in R(w) M, w' \models \alpha$, and thus $M, w \models \gamma \Rightarrow \bigcirc(\alpha)$ iff $\exists w' \in W : M, w' \models \gamma$ and $\forall w'' \leq w' : M, w'' \models \gamma \rightarrow \bigcirc\alpha$. Of course we can also build a semantics $M = \langle W, \leq, \leq_\Rightarrow, V\rangle$ that contains two orders: one for obligation and one for normality $\Rightarrow$.

derive $\top \Rightarrow \neg a$, as desired. This formalization also illustrates how to interpret the motivations that lead to the mistaken interpretation.

- $\bigcirc\neg f \wedge P_a f$ as well as $\bigcirc\neg f \wedge \bigcirc_a f$ are consistent, because it has to be consistent if in addition we have $\bigcirc(\neg a)$.
- Strengthening of the antecedent is invalid, i.e. $\bigcirc_a \neg f$ should not always be derivable from $O\neg f$, because a can be a violation and in a violation state an ideal obligation may no longer be in force.
  For another well-known example to illustrate this point, consider 'you should go to the assistance of your neighbors and tell them that you will come' $\bigcirc(a \wedge t)$, which implies $\bigcirc a$ and $\bigcirc t$. We should not derive $\bigcirc_{\neg a} t$ from $\bigcirc t$, because once the obligation to go to the assistance has been violated you also do not have to tell them that you will come. In fact, we consistently may have that if your do not go, then you are forbidden to tell that you are coming, $\bigcirc_{\neg a} \neg t$, like in Chisholm's paradox [15].
- Finally, the structural similarity between conditional obligation $\bigcirc_a f$ and the defeasible conditional $a \Rightarrow f$ is simply caused by the analogous semantic definition, and does not hint at any deeper similarity.

The relation between violability and defeasibility is a complex issue, which has been discussed in the more philosophic oriented literature such as [52,66]. One of the sources of this complexity is that there are different types of exceptionality, due to for example prima facie obligations and obligations under uncertainty (see Section 7). A more detailed discussion of this issue of out of the scope of this paper, in which we are motivated by more practical considerations. What is important here is that we can often recognize whether a certain case is a violation or an exception by simple structural properties [66,72].

3.3. Violability without exceptionality: violation contexts

The crucial problem of violability is that a violation changes the context of normative reasoning. New obligations may arise as a result of the violation, and old obligations may no longer be in force. Consequently, obligations are context-dependent. This is the case for the agent's own violations as well as violations of other agents.

The so-called *violation context* of an obligation distinguishes between ideal and varying sub-ideal contexts to formalize contrary-to-duty reasoning as it occurs in the contrary-to-duty paradoxes. In these paradoxes obligations referring to the ideal context conflict with obligations referring to the sub-ideal context. If the violation context is left implicit, then the paradoxes result in a counterintuitive inconsistency. We are interested in the relation between violation contexts and independence. They can be used to analyze the following example of Prakken and Sergot [53], that motivates much of their study in contextual deontic reasoning.

Example 1 (Cottage housing). Consider the following sentences.

(1) There must be no dog.
(2) If there is a dog, then there must be a sign.
(3) There must be no sign.

Prakken and Sergot argue that (2) and (3) are inconsistent[7] if (2) is read as an (unconditional) contextual obligation and not as an ordinary conditional obligation, because the contrary-to-duty obligation (2) regulating the violation of (1) does not respect the primary obligation (3). Contrary-to-duty obligations are intended to regulate norm violation, but they cannot just ignore all other norms. In the absence of prioritisation we may conclude that on the one hand there must be a sign if there is a dog, but on the other hand there must not. They continue to consider the following variant.[8]

(1a) There must be no dog.
(2a) If there is a dog, it must be a poodle.
(3a) There must be no poodle.

The first two sentences are a version of Forrester's gentle murderer [25] ('Smith should not murder Jones,' but 'if he murders him, then he should do it gently'), because poodles are dogs. Since it is generally accepted that Forrester's paradox is

[7] Prakken and Sergot further explain that when they say that a set of regulations or set of (contextual) obligations such as (1)-(3) is inconsistent, they mean that the set is logically inconsistent with some further assumptions, in this case that it is logically possible that there is a dog [53, footnote 3].

[8] Since Forrester's paradox is better analyzed when different types of killing are replaced by different colors [26], Prakken and Sergot also consider the following alphabetic variant: (1b) The door must be painted red; (2b) If the door is not painted red, it must be left unpainted; (3b) The door must not be left unpainted. Again they want to say that (1b)-(3b) are inconsistent, even though (1b)-(2b) are consistent, and (1b) implies (3b).

consistent, intuitively (1a) and (2a) should therefore be consistent too. Moreover, Prakken and Sergot suggest that (2a) and (3a) are inconsistent in precisely the same way that (2) and (3) are. (2a) regulates the violation of (1a) but it does not respect another primary obligation, (3a). The new feature of the poodle example, of course, is that (3a) is implied by (1a). At first sight, it seems that we have to abandon consequential closure, because this is the only way to make (1a)-(3a) logically stronger than (1a)-(2a) (such that we can add constraints that make only the latter set inconsistent). However, if a normgiver forbids having dogs he surely also implicitly forbids having any particular kind of dog. Consequential closure should not be disregarded when determining what is obligatory.

Prakken and Sergot conclude that 'the difference between what appears consistent and what appears inconsistent seems to depend critically on what is stated explicitly.' Moreover, they sketch the development of an entailment relation $\Gamma \models A$ in which designated *explicit obligations* in premises Γ will be given special status.

3.4. Violation contexts and deontic independence

In the analysis presented in this paper, the crucial distinction between (1a)-(2a) and (1a)-(3a) in Example 1 is an implicit independence assumption. We show that the three obligations are inconsistent only if the independence assumption is added. Prakken and Sergot's conclusion simply means that the explicit manner of presentation suggests that (3a) depends on (1a) if it is derived from (1a), whereas it is independent of (1a) when it is given explicitly. By making the violation contexts and independence assumptions explicit, the paradox disappears.

Example 2 (Cottage housing, continued). Consider only the first two obligations, i.e. 'there must be no dog ' and 'if there is a dog, it must be a poodle,' together with the fact that there is a pitbull. The cottage owner can either violate no rule, violate only the first one or violate both, because he cannot violate only the second one. Consequently, there are three violation contexts: the ideal context in which no obligation is violated and thus representing $\mathsf{there\ is\ no\ dog}$, the sub-ideal context $\mathsf{there\ is\ a\ dog}$ and the sub-sub-ideal context $\mathsf{there\ is\ a\ dog\ that\ is\ not\ a\ poodle}$. There is an obligation *'there must be no dog'* referring to the ideal context, and there is an obligation *'there must be a poodle'* referring to the sub-ideal context. Obviously, both obligations are violated by the pitbull. Selling the dog as well as exchanging it for a poodle are improvements of the present state of affairs. However, in the general reasoning context its owner should sell it; only in the reasoning context in which it is not sold he should exchange it. Whether the dog must be sold or exchanged depends on the context referred to, because different reasoning contexts make different violation contexts optimal.

The two derived obligations *'there must be no dog'* and *'there must be a poodle'* are contradictory. In a logic that does not distinguish between contexts, like SDL, the facts together with the two sentences of the regulations are inconsistent. DSDL3 has been developed to analyze contrary-to-duty reasoning and can be interpreted as the extension of SDL with an explicit context. However, DSDL3 with this interpretation can no longer formalize ordinary conditional obligations. We therefore write $\bigcirc_\gamma(\alpha|\beta)$ for 'α ought to be (done) if β is (done) in the context γ,' in which we represent the reasoning context γ by a propositional sentence.

To formalize the regulations without introducing the background knowledge that poodles and pitbulls are dogs, Prakken and Sergot represent the poodle by $p \wedge d$ and the pitbull by $i \wedge d$, where d stands for dog. If the reasoning context is the ideal context then dogs are forbidden ($\bigcirc\neg d$) and if the reasoning context is the sub-ideal context then there must be a poodle $\bigcirc_d(p \wedge d)$. Moreover, if there is a pitbull then the obligations are both still in force $\bigcirc(\neg d\,|\,i \wedge d)$ and $\bigcirc_d(p \wedge d\,|\,i \wedge d)$. Contextual obligations can formalize hypothetical reasoning, because the latter two obligations can be read as 'I have a pitbull and I should sell it, but if I will not sell it (hypothesis) then I should at least exchange it for a poodle.'

4. Contextual Deontic Logic

In this section we introduce a contextual deontic logic. We use the phrase 'Contextual Deontic Logic' (CDL) to refer to the logic that contains formulas of the type $\bigcirc_\gamma(\alpha|\beta)$, $F_\gamma(\alpha|\beta)$ and $P_\gamma(\alpha|\beta)$. If we have $\bigcirc\alpha$ and $\neg\bigcirc(\alpha|\beta)$, then we say that β represents exceptional circumstances in which the obligation for α is defeated. Conversely, if we have $\bigcirc\alpha$ and $\neg\bigcirc_\beta\,\alpha$ then it does *not* mean that the obligation for α is defeated by exceptional circumstances β. It means that if we assume β then we refer to another violation context that no longer prefers α. Thus β is a violation.

We keep the semantics as simple as possible. The dynamic preference semantics proposed below generalizes Hansson's preference semantics in Definition 1. The basic idea is that, when evaluating the truth condition of the obligation for α, the condition β determines the relevant preference ordering, and the context γ determines the relevant part of the preference ordering. Now, the obligation for α is true if α is true in the preferred states of this sub-ordering.[9]

Definition 3 ($\models_{CDL}$). Let $M = \langle W, \leq_\alpha, V\rangle$ with W a set of worlds, $\leq_\alpha$ a set (for each propositional formula) of transitive, reflexive and totally connected accessibility relations, and V the valuation function. We assume $\leq_\alpha = \leq_\alpha$ if $\vdash \alpha \leftrightarrow \alpha'$, where $\vdash$ stands for propositional derivibility. We have $M, w \models \bigcirc_\gamma(\alpha|\beta)$ iff there are no γ worlds, or there exists a γ world w' such that for all worlds $w'' \leq_\beta w'$ we have $M, w'' \models \gamma \rightarrow \alpha$ (all preferred γ worlds of the β ordering satisfy α). Prohibition and (weak) permission are defined by $F_\gamma(\alpha \,|\, \beta) =_{def} \bigcirc_\gamma(\neg\alpha \,|\, \beta)$ and $P_\gamma(\alpha \mid \beta) =_{def} \neg \bigcirc_\gamma (\neg\alpha \mid \beta)$. We write $\bigcirc_\gamma\alpha$ for $\bigcirc_\gamma(\alpha \mid \top)$, $\bigcirc(\alpha \mid \beta)$ for $\bigcirc_\top(\alpha|\beta)$, $\bigcirc\alpha$ for $\bigcirc_\top(\alpha|\top)$. Similarly for F and P.

Note that $\bigcirc_\gamma$ only orders γ worlds, whereas $\bigcirc(\alpha|\beta)$ does *not* only order β worlds. This difference between context and condition is related to the fact that the context is assumed to be fixed or unalterable whereas the condition is not, as explained at the end of this section. Definition 4 is analogous to Definition 2.

Definition 4 ($\vdash_{CDL}$). Let $\vdash$ be the inference relation of propositional logic. CDL is axiomatized by a correct and complete set of propositional axioms and rules, extended with:

- $\bigcirc_\alpha(\alpha|\beta)$
- $\neg \bigcirc_\gamma (\bot|\beta)$ if $\not\vdash \gamma \leftrightarrow \bot$ and $\not\vdash \beta \leftrightarrow \bot$
- $\bigcirc_\gamma(\alpha \wedge \alpha'|\beta) \leftrightarrow \bigcirc_\gamma(\alpha|\beta) \wedge \bigcirc_\gamma(\alpha'|\beta)$
- $\bigcirc_\gamma(\alpha|\beta) \leftrightarrow \bigcirc_{\gamma'}(\alpha'|\beta')$ if $\vdash \alpha \leftrightarrow \alpha'$ and $\vdash \gamma \leftrightarrow \gamma'$ and $\vdash \beta \leftrightarrow \beta'$
- $P_\gamma(\alpha|\beta) \rightarrow (\bigcirc_{\gamma\wedge\alpha}(\alpha'|\beta) \leftrightarrow \bigcirc_\gamma(\alpha \rightarrow \alpha'|\beta))$

The examples we discussed thus far all focus on the inference pattern strengthening of the antecedent, which for $\bigcirc_\gamma(\alpha \,|\, \beta)$ in CDL we can split into

[9] In contextual deontic logic we cannot represent the problematic conditions that occur later than the consequents, as the obligation in Chisholm's notorious paradox 'if the man goes to the assistance, then he should tell them before that he will come' [15]. We thus 'solve' the backwards version of the paradox by restricting the language, cf. [67].

strengthening of the context γ and strengthening of the condition β. The latter axiom illustrates that for the contexts we have conditional strengthening, because $\bigcirc_{\gamma\wedge\gamma'}(\alpha)$ can be derived from $\bigcirc_{\gamma}(\alpha)$ if $P_{\gamma}(\gamma')$. Strengthening of the context is thus blocked if we have $F_{\gamma}(\gamma')$. A further inspection of the axioms reveals that for the condition there is no strengthening at all.

Again we have to consider the absence of factual detachment, which is now represented by the absence of the following two principles.

$$O_{\gamma}\alpha \wedge \gamma \rightarrow O_{\gamma}\alpha$$

$$O(\alpha|\beta) \wedge \beta \rightarrow O_{\gamma}\alpha$$

We use a solution in the spirit of Boutilier's simple approach discussed in Section 3.1, based on Lang's extension of Boutilier's framework [38,40]. Lang distinguishes between background knowledge (like before represented by KB) and contingent knowledge or facts, represented by F. We now simply ask for those propositions α such that

$$\models_{CDL} \bigcirc_{KB}(\alpha|F)$$

The extension of DSDL3 to CDL has been straightforward from a technical point of view, but it poses two new challenges when we use CDL to represent examples in it: do we represent knowledge by KB or by F, and do we represent norms 'α ought to be (done) if β' by $\bigcirc_{\beta}\alpha$, by $\bigcirc(\alpha\,|\,\beta)$ or even by $\bigcirc_{\beta}(\alpha\,|\,\beta)$? Lang's work as well as the discussion in the previous section of the benchmark examples of deontic logic suggest the following guidelines.

First, a sentence is represented as background knowledge if it is assumed to be 'fixed' or 'unalterable', otherwise it is contingent knowledge. This notion has already been discussed by Hansson [27] when he argued that the point of dyadic deontic logics of the kind he introduced (to formalize contrary-to-duty reasoning) is that the antecedent is assumed to be fixed and the consequent expresses 'how to make the best of the sad circumstances.' In Loewer and Belzer's so-called dyadic Deontic Detachment (or 3D) this intuition is formalized by a modal operator $U\alpha$, to be read as 'α is unalterable.' It is also (either implicitly or explicitly) used in many more recent approaches in deontic logic, see Section 8 on related research.

We can distinguish two kinds of sentences: sentences which cannot change after they have become true, and sentences which still can be changed. Sentences of the first type are *usually* (but not always!) best represented by background knowledge. For example, a sentence 'John is murdered' is best represented by background knowledge if we are interested in what must be done afterwards, since once he is murdered it cannot be undone again. However, even such sentences may be represented by contingent knowledge, if for example we are interested in the question whether this murder is a violation.

Sentences of the second type are best represented by contingent knowledge, except in cases of contrary-to-duty reasoning. For example, if the sentence 'there is a dog' is considered to be fixed, then it is assumed that situations in which there is no dog are no longer taken into consideration, despite the fact that it is still physically possible to for example sell the dog.

Second, whether a norm 'α ought to be (done) if β' is represented by $\bigcirc_\beta \alpha$, by $\bigcirc(\alpha|\beta)$ or by $\bigcirc_\beta(\alpha|\beta)$ follows from this distinction. Most norms should be represented by $\bigcirc(\alpha|\beta)$, except when they involve contrary-to-duty reasoning. If α is only obliged if we assume that β is true and cannot be altered anymore, then it must be represented by $\bigcirc_\beta \alpha$. For example, if β represents that there is a dog and α that there is a poodle, and the obligation to have a poodle if there is a dog is only meant to make the best out of the sad circumstances instead of reaching for the ideal situation, then it must be represented by $\bigcirc_\beta \alpha$.

An issue which we leave open is which further restrictions on the logic may be imposed. For example, it seems that in most situations we envision we could demand that F implies KB, and that $\bigcirc_\beta \alpha$ implies $\bigcirc_\beta(\alpha|\beta)$, and maybe even that $\bigcirc_\beta(\alpha|\beta)$ implies $\bigcirc_\beta \alpha$ (the meaning of these two expressions seem to be very close). We left these issues open since we do not want to exclude any possible unforeseen applications of the logic.

We end with a warning. The deontic logic literature discusses many complicated examples, one of them we discuss in the following section, for which it is very difficult to find the best or even a satisfactory representation. This seems to be inherent to normative, moral and legal reasoning. One consequence is that we cannot give guidelines which always solve the problem of the knowledge engineer, and another consequence is that to see whether a representation is good enough, we have to consider the logical consequences of the representation. In this paper we do not further discuss this issue, but we turn to an additional tool which can help to represent examples: independence assumptions.

5. Deontic Independence

In this section we formally define independence, and we show that the deontic paradoxes discussed thus far do not arise when the violation contexts and independence assumptions are made explicit. The basic idea is as follows. Thus far, we used three sets of sentences to represent examples: a set of deontic sentences, a set of background knowledge sentences KB and a set of contingent knowledge sentences F. In this section we further partition the set of deontic sentences in a set of norms N, a set of factual independence statements FI, and a set of deontic independence statements DI.

Qualitative independence has been used to deal with the irrelevance problem of default reasoning. For example, 'red birds fly' $r \wedge b \Rightarrow f$ can be derived from 'birds fly' $b \Rightarrow f$ if 'red' r is independent of 'flying' f [22]. In contextual deontic reasoning there are two types of qualitative independence which can be used to deal with irrelevance problems. *Factual independence* is independence with respect to the condition, and *deontic independence* is independence with respect to the context. The following definition formalizes these notions.

Definition 5 (Normative specification). A *normative specification* is a tuple $NS = \langle N, KB, F, FI, DI \rangle$ where N is a set of CDL sentences, KB and F sets of propositional sentences, and FI and DI sets of CDL sentences of the form of respectively:

$$FI : \bigcirc_\gamma(\alpha|\beta) \leftrightarrow \bigcirc_\gamma(\alpha|\beta \wedge \beta')$$

$$DI : \bigcirc_\gamma(\alpha|\beta) \leftrightarrow \bigcirc_{\gamma\wedge\gamma'}(\alpha|\beta)$$

or their negations (for respectively factual and deontic dependence). We say respectively that β' is factually independent with respect to the obligation for α if β in the context γ and that γ' is deontically independent with respect to the obligation for α if β in the context γ.[10] We also say that $\bigcirc_\gamma(\alpha \mid \beta)$ and β' are factually independent, or that $\bigcirc_\gamma(\alpha \mid \beta)$ and $\bigcirc_\gamma(\alpha \mid \beta \wedge \beta')$ are factually independent, etc.

Following [22], we call $\bigcirc_\gamma(\alpha|\beta)$ the *a priori* obligation, and $\bigcirc_\gamma(\alpha|\beta \wedge \beta')$ the *de facto* obligation. We can discriminate between nine different situations,

[10] Prakken and Sergot [52] call the formula $O_{A\wedge B}C \rightarrow \bigcirc_A C$ up, and the formula $O_A C \rightarrow \bigcirc_{A\wedge B} C$ is called down.

because a priori and de facto α is either obliged, forbidden, or it and its negation are both permitted. Factual and deontic independence are powerful concepts, in the sense that obligations can be derived from obligations and permissions from permissions in case of independence, and obligations can be derived from permissions and permissions from obligations in case of dependence. The following continuation of Example 2 shows that factual independence plays a similar role in defeasible deontic reasoning as independence in default reasoning.

Example 3 (Cottage housing, continued).
Let $NS = \langle \{Fd\}, \emptyset, \emptyset, \{Fd \leftrightarrow F(d \mid i \wedge d), \neg(Fd \leftrightarrow F(d \mid b))\}, \emptyset\rangle$. The CDL prohibition 'dogs are forbidden' Fd is factually independent of the existence of a pitbull, and from this independence relation we can derive the prohibition 'if there is a pitbull, then dogs are forbidden' $F(d|i \wedge d)$. Moreover, the prohibition is factually dependent on the needs of a blind owner, from which we can derive the permission *'if the owner is blind, then it is permitted that there is a dog'* $P(d|b)$.[11]

As discussed in Section 3.1, we have to distinguish between decision variables and parameters or events (controllable and uncontrollable variables), because it does not make sense to oblige someone to see to it that a parameter has a desirable value [10,64,38]. In this example we take blindness as a parameter, because we do not want to be able to infer that it is obligatory or forbidden that the owner is blind.

The following example illustrates how deontic independence can be used to analyze the Cottage Housing Regulations in Example 1.

Example 4 (Cottage housing, continued).
$NS_1 = \langle N_1, \emptyset, \emptyset, \emptyset, DI_1\rangle$, $NS_2 = \langle N_2, \emptyset, \emptyset, \emptyset, DI_2\rangle$,
$N_1 = \{Fd, \bigcirc_d(p \wedge d)\}$, $DI_1 = \{\neg(F(p \wedge d) \leftrightarrow F_d(p \wedge d))\}$,
$N_2 = \{Fd, \bigcirc_d(p \wedge d), F(p \wedge d)\}$, and $DI_2 = \{F(p \wedge d) \leftrightarrow F_d(p \wedge d)\}$.
N_1 and N_2 are equivalent since Fd implies $F(p \wedge d)$. A typical CDL model of these sets consists of three equivalence classes (of $\{(w, w') \mid w \leq w'$ and $w' \leq w\}$), in

[11] It may seem strange to *derive* $P(d \mid b)$ from $F(d)$ and $\neg(F(d) \leftrightarrow F(d \mid b))$, because factual dependence is defined in terms of existence of the permission. However, factual dependence statements can also be derived from other knowledge, such as for example graphs. This issue is discussed in the following section.

order of preference the $\neg d$ worlds, the $p \wedge d$ worlds and the $\neg p \wedge d$ worlds. However, we have $N_1 \cup DI_1 \models P_d(p \wedge d)$ and $N_2 \cup DI_2 \models F_d(p \wedge d)$, because in NS_1 the prohibition of poodles deontically depends on dog, whereas they are deontically independent in NS_2. The latter prohibition is inconsistent with $\bigcirc_d(p \wedge d)$, and $N_2 \cup DI_2$ is therefore inconsistent. We thus explained Prakken and Sergot's problem in Example 1. Moreover, the deontic independence relation is suggested by the explicit manner of presentation, because $F(p \wedge d)$ is deontically independent of d only if it is given explicitly (in N_2).

Deontic independence can also be used to analyze the Reykjavik Scenario [7,64]. The example is taken from a discussion by Makinson [44]. The explicit obligations are that neither Reagan nor Gorbatchov should be told the secret, if Reagan is told the secret then Gorbatchov should be told, and if Gorbatchov is told then Reagan should be told. Intuitively, under the condition that Gorbatchov is told, we would like to be able to conclude that Reagan should be told, without also deriving its negation. On the other hand, if the first item is broken down into two parts, saying separately that Reagan should not be told and that Gorbatchov should not be told, so that there are four explicit obligations, then the conclusion seems intuitively more ambiguous, depending on whether promulgations are prioritized according to, say, specificity of their antecedents. In the absence of prioritisation we seem authorized to conclude both that Reagan should be told and that he should not; with prioritisation, only the former.

Example 5 (Reykjavik Scenario).
$NS_1 = \langle N_1, \emptyset, \emptyset, \emptyset, DI_1 \rangle$, $NS_2 = \langle N_2, \emptyset, \emptyset, \emptyset, DI_2 \rangle$,
$N_1 = \{\bigcirc(\neg r \wedge \neg g), \bigcirc_r g, \bigcirc_g r\}$, $DI_1 = \{\neg(\bigcirc \neg r \leftrightarrow \bigcirc_g \neg r)\}$,
$N_2 = \{\bigcirc(\neg r) \wedge \bigcirc(\neg g), \bigcirc_r g, \bigcirc_g r\}$, and $DI_2 = \{\bigcirc \neg r \leftrightarrow \bigcirc_g \neg r\}$.
N_1 and N_2 are equivalent since $\bigcirc(\neg r \wedge \neg g)$ implies $\bigcirc(\neg r) \wedge \bigcirc(\neg g)$ and vice versa. A typical CDL model consists of three equivalence classes, containing in order of preference the $\neg r \wedge \neg g$ worlds, the $r \wedge g$ worlds and the $r \leftrightarrow \neg g$ worlds. However, $N_1 \cup DI_1 \models P_g \neg r$ and $N_2 \cup DI_2 \models \bigcirc_g \neg r$, because in NS_1 the obligation $\bigcirc \neg r$ is deontically dependent on $\neg g$, and in NS_2 they are independent (as suggested by the explicit manner of presentation in N_1 and N_2). The set $N_2 \cup DI_2$ is inconsistent, because the premise $\bigcirc_g r$ and the derived $\bigcirc_g \neg r$ are contradictory. Only with the independence assumption are the sets inconsistent.

The latter example suggests that also the Reykjavik paradox does not arise when violation contexts and independence assumptions are given explicitly. However, for a full analysis of this paradox we also have to consider so-called defeasibility interpretations of it, because the real problem of the Reykjavik paradox is how to interpret the sentences when more specific obligations override more general ones [47,64].

Example 6 (Reykjavik Scenario, continued).
$NS_1 = \langle N, \emptyset, \emptyset, FI_1, \emptyset\rangle$, $NS_2 = \langle N, \emptyset, \emptyset, FI_2, \emptyset\rangle$,
$N = \{\bigcirc(\neg r), \bigcirc(\neg g), \bigcirc(g|r), \bigcirc(r|g)\}$,
$FI_1 = \{\neg(\bigcirc(\neg r) \leftrightarrow \bigcirc(\neg r|g))\}$, and $FI_2 = \{\bigcirc(\neg r) \leftrightarrow \bigcirc(\neg r|g)\}$.
In NS_1 the obligation for $\bigcirc(\neg r)$ is factually dependent on g, the obligation is overridden and the set is consistent. In NS_2 they are factually independent, and the set is inconsistent.

The Reykjavik example further clarifies the question whether an obligation 'α is obligatory if β' should be formalized by $\bigcirc_\beta \alpha$, $\bigcirc(\alpha\,|\,\beta)$ or even $\bigcirc_\beta(\alpha\,|\,\beta)$, because the Reykjavik paradox illustrates that natural language is ambiguous for complicated examples that contain violations and exceptions. Consequently, which of these formalizations of the obligation is intended by the speaker cannot be derived from the structure of natural language, but it can only be determined from its intended behavior. Summarizing, an obligation is deontically independent if it refers to the same violation context, and factually independent if it is not overridden. Deontic independence and factual independence are thus quite different. One aspect of this difference is that different types of independence relations can be inferred from the explicit representation, as discussed in the following two sections.

6. Manner of presentation

Explicitly given violation contexts and independence assumptions may be useful to analyze deontic paradoxes, but they do not directly lead to a practical approach to reasoning about obligations. For example, let $NS = \langle N, KB, F, FI, DI\rangle$, $N = \{\bigcirc(p)\}$, and $DI = \{\bigcirc(p) \leftrightarrow \bigcirc_q(p)\}$. We may as well represent this example more concisely by $N = \{\bigcirc(p) \wedge \bigcirc_q(p)\}$ and $DI = \emptyset$. From the perspective of knowledge representation, we do not seem to have gained much with the introduction of FI and DI. Clearly, in practical reasoning we need another way to represent the independence relations FI and DI. The typical example of using independence relations is to derive them from a graphical representation like a Bayesian network [50]. Another well-known approach in non-monotonic reasoning and reasoning about uncertainty is to infer the independence relations FI and DI from the explicit manner of presentation of the obligations in N (as also suggested by Prakken and Sergot [53]). This has been called a syntactic approach, because semantically identical sets may derive different conclusions due to their syntactic presentation. In this section we use a variant of the syntactic approach. For simplicity we assume that the N only contains contextual obligations, i.e. no disjunctions and negations of modal operators, and no permissions or prohibitions.

We study the explicit manner of presentation by introducing a procedure to derive the set DI from N. Definition 6 says that each $\bigcirc_\gamma(\alpha \mid \beta)$ in N induces the independence relations that $\bigcirc_\gamma(\alpha|\beta)$ is deontically independent of any fact, unless it violates 'ought implies can.' That is, if $\bigcirc_\gamma(\alpha|\beta)$ is explicitly represented, then we have that it is deontically independent of γ' iff $\alpha \wedge \gamma \wedge \gamma'$ is consistent.

Definition 6. Let N be a set of contextual obligations of CDL. The set of derived deontic independence statements of N is:

$$DI(N) = \{\bigcirc_\gamma(\alpha|\beta) \leftrightarrow \bigcirc_{\gamma\wedge\gamma'}(\alpha|\beta) \mid \bigcirc_\gamma(\alpha|\beta) \in N, \alpha \wedge \gamma \wedge \gamma' \text{ is consistent}\}$$

The derivation of the independence assumptions is illustrated by our two running examples.

Example 7 (Cottage housing, continued). Consider the two sets of CDL formulas $N_1 = \{\bigcirc(\neg d), \bigcirc_d(p \wedge d)\}$ and $N_2 = \{\bigcirc(\neg d), \bigcirc_d(p \wedge d), \bigcirc(\neg(p \wedge d))\}$. N_1 and N_2 are equivalent, but $DI(N_1)$ and $DI(N_2)$ are not. Moreover, $N_1 \cup DI(N_1)$ is consistent whereas $N_2 \cup DI(N_2)$ is inconsistent.

Example 8 (Reykjavik Scenario). Consider the two sets of CDL formulas $N_1 = \{\bigcirc(\neg r \wedge \neg g), \bigcirc_r g, \bigcirc_g r\}$ and $N_2 = \{\bigcirc(\neg r), \bigcirc(\neg g), \bigcirc_r g, \bigcirc_g r\}$. N_1 and N_2 are equivalent, but $DI(N_1)$ and $DI(N_2)$ are not. Moreover, $N_1 \cup DI(N_1)$ is consistent whereas $N_2 \cup DI(N_2)$ is inconsistent.

The following example illustrates that the notion of deontic independence formalized in Definition 6 is very strong, in the sense that many independence relations are derived from the explicit representation.

Example 9. Let $N = \{\bigcirc p, \bigcirc q\}$. Due to the theorem $\bigcirc_\gamma \alpha \leftrightarrow \bigcirc_\gamma(\alpha \wedge \gamma)$ characteristic of the preference-based approach [68], $N \cup DI(N)$ not only derives $\bigcirc_{\neg(p\wedge q)} p$ and $\bigcirc_{\neg(p\wedge q)} q$, but also the contradictory $\bigcirc_{\neg(p\wedge q)}(p \wedge \neg q)$ and $\bigcirc_{\neg(p\wedge q)}(\neg p \wedge q)$. Thus, in CDL extended with Definition 6 these two obligations of N can only be represented in a consistent way by $N' = \{\bigcirc(p \wedge q)\}$.

The example also illustrates that if we want to combine independent obligations with the strong independence assumptions of Definition 6, then we have to weaken the logic such that $\bigcirc_\gamma \alpha \leftrightarrow \bigcirc_\gamma(\alpha \wedge \gamma)$ no longer holds. For example, in [45,46,65] a logic has been proposed that does not imply this formula, but that still validates the intuitive derivation of 'if a then $\bigcirc y$' from 'if a then $\bigcirc x$' and 'if $a \wedge x$ then $\bigcirc y$.'

7. Two types of defeasible deontic logic

In this section we present more examples which give an intuitive motivation and explanation of the formalization, focusing on factual independence. Before we start, we notice that there are many different ways to formalize exceptionality. In this section we only illustrate how two different notions of defeasible obligations – obligations under uncertainty and prima facie obligations – can both be formalized in CDL. We have discussed in earlier work [66,72] that the distinction between obligations under uncertainty and prima facie obligations is that when an obligation under uncertainty is overridden it is completely cancelled, whereas a prima facie obligation is only overshadowed. Consequently, if the overriding obligation is itself overridden, then with obligations under uncertainty there is no reinstatement, but with prima obligations there is.

In the following definition of obligations under uncertainty that does not satisfy reinstatement we use a simple definition of overridden [30,31,64].

Definition 7 (Obligations under uncertainty). Let N be a set of contextual obligations of CDL. The set of derived factual independence statements of N is:

$$FI(N) = \{\bigcirc_\gamma(\alpha|\beta) \leftrightarrow \bigcirc_\gamma(\alpha|\beta \wedge \phi) \mid \bigcirc_\gamma(\alpha|\beta) \in N \text{ is not overridden for } \beta \wedge \phi\}$$

where $\bigcirc_\gamma(\alpha|\beta) \in N$ is overridden for ϕ if there is a $\bigcirc_\gamma(\alpha'|\beta') \in N$ such that:

1. $\phi \vdash \beta$, $\phi \vdash \beta'$, $\beta' \vdash \beta$, $\beta \not\vdash \beta'$, where $\vdash$ stands for propositional derivibility, and
2. $\phi \wedge \alpha \wedge \alpha'$ is inconsistent, and
3. $\phi \wedge \alpha'$ is consistent.

To introduce reinstatement we make the definition recursive. This definition is well-founded, because an obligation can only be overridden by a more specific obligation and therefore there cannot be cycles.

Definition 8 (Prima facie obligations). Let N be a set of contextual obligations of CDL. The set of derived factual independence statements of N is:

$$FI(N) = \{\bigcirc_\gamma(\alpha|\beta) \leftrightarrow \bigcirc_\gamma(\alpha|\beta \wedge \phi) \mid \bigcirc_\gamma(\alpha|\beta) \in N \text{ is not rec. overridden for } \beta \wedge \phi\}$$

where $\bigcirc_\gamma(\alpha|\beta) \in N$ is recursively overridden for ϕ if it is overridden for ϕ by an obligation that is itself not overridden for ϕ.

The wealth of approaches in non-monotonic reasoning and reasoning about uncertainty suggests that there are many other approaches to derive independence assumptions. A further discussion is out of the scope of this paper, because here we only illustrate that different approaches can be formalized in CDL. We end with some closing remarks.

- CDL does not impose any restrictions on the condition. Such restrictions can be added depending on the application of the logic. For example, for reasoning by cases one could add the following disjunction rule.

$$\bigcirc_\gamma(\alpha|\beta_1) \wedge \bigcirc_\gamma(\alpha|\beta_2) \rightarrow \bigcirc_\gamma(\alpha|\beta_1 \vee \beta_2)$$

- An approach to strengthening of the antecedent popular in reasoning about uncertainty is to select an in some sense least informed model. In CDL we could take the least informed model of the ones satisfying a set of obligations with derived deontic independence assumptions given in Definition 6.
- The following example is well-known from dyadic deontic logic [78,68] and illustrates why we cannot derive factual independence analogous to deontic independence in Definition 6.

Example 10 (Window [78]). The diamond 'the window should be closed if it rains' $\bigcirc(c \mid r)$ and 'the window should be open if the sun shines' $\bigcirc(\neg c|s)$ should not derive the contradictory $\bigcirc(c|r \wedge s)$ and $\bigcirc(\neg c|r \wedge s)$.

- Example 10 illustrates that the factual independence relations implied by the explicit representation are much weaker than the deontic independence relations due to possible conflicts. As a consequence of these conflicts, in more specific circumstances there are only *prima facie* obligations [56,4,49,72]. For example, there are only prima facie obligations to close and to open the window when it rains and the sun shines. In a preference-based approach this can be obtained by replacing the connected ordering by a possibly non-connected ordering. There is a prima facie obligation for α if there is an equivalence class of preferred worlds that satisfies α.

In developing new derivations of independence assumptions from an explicit set of norms, the interaction between the two definitions (in our case Definitions 6 and 7/8) has to be taken into account. For example, the derivation of $\bigcirc_\gamma(\alpha|\beta)$ from $\bigcirc\alpha$ can be obtained in two ways, respectively via $\bigcirc_\gamma\alpha$ and via $\bigcirc(\alpha|\beta)$, i.e. via first using deontic independence or via first using factual independence. This also illustrates one of the ways in which CDL can be used to study the interaction between violability and defeasibility.

8. Related Research

8.1. Normative agents

Originally the concept 'agent' was used in economics as a general description of an actor, a decision maker. According to classical decision theory such a decision maker is rational if it maximizes its expected utility [42,58]; such an economic agent is now usually called a rational or an ideal agent. The theory has been extended for multi agent systems in game theory [42] and social choice theory [3]. In artificial intelligence the concept 'agent' is used for computationally or resource-bounded decision makers, and such an agent is usually called an autonomous or an intelligent agent. It is closely related to the notion of cognitive agent as discussed in psychology, cognitive science [21], and philosophy [11]. Norms have been discussed in relation to rational, autonomous as well as cognitive agents.

Classical decision theory and game theory [42] suggest that norms are of practical use when the agents are cooperative (for example increasing stability in so-called coordination games), and they are necessary when the agents are competitive (for example the norms of reciprocity in prisoner's dilemmas and evolutionary game theory [5]). Our favorite example is the following two agent game called the thousand pennies game [28]. We explain it in some detail, because it nicely illustrates the distinction between rational, cognitive and normative agents. When the game starts, there are thousand pennies on the table. Agent 1 can either take two pennies after which the game is over, or take one penny in which case it is agent 2's turn. It can also take two pennies after which the game is over or take one penny after which it is again agent 1's turn. Etcetera. Agents maximize the number of pennies they get (it is a so-called non-zero sum game) and assume the other agents do likewise.

- Rationality dictates the first agent to take two pennies immediately, following a backward induction argument. If there are only two pennies left, then agent 1 will surely take both. Therefore, if there are three pennies left, agent 2 will take two. Etcetera. This is called a paradox, because decision makers are thought to be able to 'recognize' that they can do much better. Several solutions have been proposed to break the paradox, in particular the following two.
- The limited or bounded rationality analysis of BDI logics argues that the agent takes one instead of two, thus breaking the backward induction paradox, because agents cannot reason thousand steps deep. However, the problem of this solution is that when agents get smarter, they trust less (they can reason more steps backwards) and therefore get worse results. In human reasoning this 'trust within reason' problem [29] has been observed as merely 'philosophical' since humans do no get much smarter, but for artificial agents it is not! It is well known that each 18 months the computing power is doubled (known as Moore's law), and it thus may be expected that limited agents get smarter, they trust less and obtain worse results.
- An alternative way to break the backward induction argument is to introduce norms in the agent's decision process [29]. The typical property of such norms is that they contain information about the past. The first agent can take one penny instead of two, even when it is smart enough to reason thousand steps deep. This agent will get a higher pay-off when it encounters other social

agents.

Cognitive science suggests that an agent in a multi agent system is a *social* agent [14], in the sense that agents that are part of a group make decisions in a different way than agents that are alone. This is inspired by theories developed in sociology, where the fact that man is a social animal is thought to have several consequences on his behavior. Norms play an important role in these social theories.

8.2. Independence

In this paper we have shown how independence assumptions can be used in normative reasoning considered as context-dependent reasoning. We are not aware of any other notions of independence in deontic logic, but contexts have been used before. For example, Åqvist's optimality classes [2] can be interpreted as contextual parameters. Moreover, Jones and Pörn's bimodal logic [35] and more recent proposals of Carmo and Jones [13] are also based on different degrees of ideality and the idea of some facts being settled (like our context of reasoning). Context and temporal reasoning have been discussed in [62] and context and defeasible reasoning in [71]. In the semantics we proposed in this paper the violation contexts are represented implicitly by the preference relation $\leq$. It could be represented more explicitly in several different ways. For example, we can explicitly represent the violations of the violation contexts, or we can explicitly represent the equivalence classes of the relation $\{(w, w') \mid w \leq w' \text{ and } w' \leq w\}$ by a set of constants, like e.g. Åqvist's optimality classes [2]. Moreover, we can explicitly represent the norm context or the prescriptive obligations in the semantics, and deduce the preference relation, violation contexts or optimality classes from them. We did not introduce explicit contexts here, because we want to stay as close as possible to the well known traditional approaches, we want to keep our semantics as simple as possible and it is not necessary to introduce notions of deontic independence.

Independence has been studied in other areas. Gärdenfors studied independence inspired by the axiom of probability theory that for disjoint sets A and B we have that the probability of $A \cup B$ is the sum of the probabilities of A and B, i.e. $Pr(A \cup B) = Pr(A) + Pr(B)$ if $A \cap B = \emptyset$. Moreover, independence relations between sets of variables is an important concept to reduce complexity in probability and utility functions in respectively Bayesian networks (reasoning about uncertainty and default reasoning, and in epistemic reasoning) and decision-theoretic planning (reasoning about goals and desires), such that for example for independent x and y we have $p(x, y) = p(x)p(y)$ and $u(x, y) = u(x) + u(y)$. In normative reasoning, such independence relations between variables could be used for example to derive for independent p and q the permission $P(p \wedge q)$ from Pp and Pq. Moreover, context has been studied in several areas. Recent ideas of formalizing context in the AI literature are for example McCarthy's formalism based on an ist-operator.

An alternative to independence assumptions is the repair approach to conditional knowledge bases [8]. The idea is to take advantage of the main features of nonmonotonic inference with respect to rule addition. Contrary to classical logic, it is legitimate to expect that undesirable conclusions can be blocked by adding new rules (and thus not only by deleting them). Moreover, it is also conjectured that reasoning on basis of a ranking is expressive and flexible enough to capture any 'reasonable' expected behavior of a normative base. It is an open problem how this repair approach can be used to formalize contrary-to-duty reasoning too.

Independence is closely related to irrelevance. However, our work is not related to relevant implication [48] or relevant deontic logic [61].

9. Conclusion

In this paper we considered the relevance of philosophical discussions for the representation of norms in multi agent systems. We have given and discussed two arguments why norms are interesting objects of study for multi agent systems. The first argument is that norms bridge cognitive and social concepts, that is, they bridge the autonomous agent level and the multi agent system level. Norms of individual agents play a role in the agent's decision-making, and norms in a multi agent system are crucial tools for coordination, cooperation and trust. The second argument is that norms which can be violated play an important role in deliberation and autonomy of an agent, and control of a multi agent system. This raises many subjects for further research, such as for example the question of how norms play these different roles, or how they can be used in these different circumstances.

An important question for deontic logic applications studied in this paper is how to proceed once a norm has been violated. We proposed a practical approach for reasoning about norms and obligations based on violation contexts and independence assumptions. The problem how to represent and reason about norms is split into two parts. The first part is concerned with a correct representation of norms and uses a weak deontic logic and violation contexts. We have $\bigcirc\alpha$ without $\bigcirc(\alpha|\beta)$ if β expresses exceptional circumstances, and we have $\bigcirc\alpha$ without $\bigcirc_\gamma\alpha$ if γ expresses a violation. The obligation is factually dependent if it is overridden, and deontically dependent if it is violated. The second part is concerned with an *efficient* representation of norms and derives independence relations from other sources, like the explicit representation of norms or a graphical representation. We have also shown how violation contexts and independence assumptions can be used to represent deontic paradoxes like the cottage housing regulations and the Reykjavik paradox.

A subject for further research is the representation of independence assumptions by other representations (like the graphical representation of independence by independence networks [6]). Moreover, the approach introduced in this paper can be compared to the repair approach to conditional knowledge based mentioned in the previous section. Finally subjects of further research are the relevance of our study for philosophical discussions on the interference between factual and deontic independence, the relation between the two different types of independence and the different types of defeasibility that have been observed in [66], and finally the relation between contextual deontic reasoning and defeasible [71] and temporal reasoning [62].

Acknowledgment

Thanks to the anonymous referees of this journal for their insightful and useful comments on an earlier version of this paper.

References

[1] C.E. Alchourrón. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 43–84. John Wiley & Sons, 1993.

[2] L. Åqvist. Systematic frame constants in defeasible deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, pages 59–77. Kluwer, 1997.

[3] K.J. Arrow. *Social Choice and Individual Values*. Wiley, New York, 2nd edition, 1963.

[4] N. Asher and D. Bonevac. Prima facie obligation. *Studia Logica*, 57:19–45, 1996.

[5] R. Axelrod. *The evolution of cooperation.* Basic books, 1984.

[6] F. Bacchus and A.J. Grove. Utility independence in a qualitative decision theory. In *Proceedings of KR'96*, pages 542–552, 1996.

[7] M. Belzer. A logic of deliberation. In *Proceedings of the AAAI'86*, pages 38–43, 1986.

[8] S. Benferhat, D. Dubois, and H. Prade. Practical handling of exception-tainted rules and independence information in possibilistic logic. *Applied Intelligence*, 9:101–127, 1998.

[9] C. Boutilier. Conditional logics of normality: a modal approach. *Artificial Intelligence*, 68:87–154, 1994.

[10] C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the KR'94*, pages 75–86, 1994.

[11] M. E. Bratman. *Intention, plans, and practical reason.* Harvard University Press, Cambridge Mass, 1987.

[12] J. Broersen, M. Dastani, Z. Huang, J. Hulstijn, and L. van der Torre. The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the Fifth International Conference on Autonomous Agents (AA2001)*, pages 9–16, 2001.

[13] J. Carmo and A.J.I. Jones. A new approach to contrary-to-duty obligations. In D. Nute, editor, *Defeasible Deontic Logic*, pages 317–344. Kluwer, 1997.

[14] C. Castelfranchi. Modelling social action for AI agents. *Artificial Intelligence*, 103:157–182, 1998.

[15] R.M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.

[16] P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[17] R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm acceptance. In *Intelligent Agents V; Proceedings of the ATAL'98*, LNAI 1555, pages 319–333. Springer, 1998.

[18] D. Dennett. *The intentional stance.* MIT Press / Bradford Books, Cambridge, MA, 1987.

[19] F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7:69–79, 1999.

[20] F. Dignum, D. Morley, E. Sonenberg, and L. Cavedon. Towards socially sophisticated BDI agents. In *Proceedings of the fourth International Conference on MultiAgent Systems (ICMAS-2000)*, pages 111–118, 2000.

[21] F. Dretske. *Explaining Behavior: Reasons in a world of causes.* MIT Press, 1988.

[22] D. Dubois, L. Farinas del Cerro, A. Herzig, and H. Prade. Qualitative relevance and independence: a roadmap. In *Proceedings of the IJCAI'97*, pages 62–67, 1997.

[23] T. Eiter, V.S. Subrahmanian, and G. Pick. Heterogeneous active agents, I: Semantics. *Artificial Intelligence*, 108:179–255, 1999.

[24] B. Sadighi Firozabadi and L. van der Torre. Towards a formal analysis of control systems. In *Proceedings of the ECAI'98*, pages 317–318, 1998.

[25] J.W. Forrester. Gentle murder, or the adverbial Samaritan. *Journal of Philosophy*, 81:193–197, 1984.

[26] L. Goble. Murder most gentle: the paradox deepens. *Philosophical Studies*, 64:217–227, 1991.

[27] B. Hansson. An analysis of some deontic logics. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, pages 121–147. D. Reidel Publishing Company, Dor-

drecht, Holland, 1971.

[28] M. Hollis. Penny pinching and backward induction. *Journal of Philosophy*, 88:473–88, 1991.

[29] M. Hollis. *Trust Within Reason*. Cambridge University Press, 1998.

[30] J.F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23:35–65, 1994.

[31] J.F. Horty. Nonmonotonic foundations for deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, pages 17–44. Kluwer, 1997.

[32] J.F. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

[33] R. Jeffrey. *The Logic of Decision*. University of Chicago Press, 2nd edition, 1983.

[34] R.E. Jennings. A utilitarian semantics for deontic logic. *Journal of Philosophical Logic*, 3:445–465, 1974.

[35] A.J.I. Jones and I. Pörn. Ideality, sub-ideality and deontic logic. *Synthese*, 65:275–290, 1985.

[36] A.J.I. Jones and M. Sergot. Deontic logic in the representation of law: Towards a methodology. *Artificial Intelligence and Law*, 1:45–64, 1992.

[37] A.J.I. Jones and M. Sergot. On the characterisation of law and computer systems: The normative systems perspective. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science*. John Wiley & Sons, 1993.

[38] J. Lang. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *Proceedings of the ECAI'96*, pages 318–322, 1996.

[39] J. Lang, L. van der Torre, and E. Weydert. Two types of conflicts between desires (and how to resolve them). In *Proceedings AAAI Spring Symposium Workshop on game-theoretic and decision-theoretic approaches to agency (GTDT2001)*, 2001.

[40] J. Lang, L. van der Torre, and E. Weydert. Utilitarian desires. *Autonomous Agents and Multi Agent Systems*, To appear.

[41] D. Lewis. Semantic analysis for dyadic deontic logic. In S. Stunland, editor, *Logical Theory and Semantical Analysis*, pages 1–14. D. Reidel Publishing Company, Dordrecht, Holland, 1974.

[42] R.D. Luce and H. Raiffa. *Games and Decisions*. Wiley, 1957.

[43] D. Makinson. Five faces of minimality. *Studia Logica*, 52:339–379, 1993.

[44] D. Makinson. On a fundamental problem of deontic logic. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*, pages 29–54. IOS Press, 1999.

[45] D. Makinson and L. van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[46] D. Makinson and L. van der Torre. Constraints for input/output logics. *Journal of Philosophical Logic*, 30:155–185, 2001.

[47] L.T. McCarty. Defeasible deontic reasoning. *Fundamenta Informaticae*, 21:125–148, 1994.

[48] L.T. McCarty. Modalities over actions: 1. model theory. In *Proceedings of the KR'94*, pages 437–448, 1994.

[49] M. Morreau. *Prima Facie* and seeming duties. *Studia Logica*, 57:47–71, 1996.

[50] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, 1988.

[51] J. Pearl. From conditional oughts to qualitative decision theory. In *Proceedings of the UAI'93*, pages 12–20, 1993.

[52] H. Prakken and M.J. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57:91–115, 1996.

[53] H. Prakken and M.J. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In D. Nute, editor, *Defeasible Deontic Logic*, pages 223–262. Kluwer, 1997.

[54] A. Rao and M. Georgeff. Modeling rational agents within a bdi architecture. In *Proceedings of the International Workshop on Knowledge Representation (KR91)*, pages 473–484, 1991.

[55] A. Rao and M. Georgeff. An abstract architecture for rational agents. In *Proceedings of the International Workshop on Knowledge Representation (KR92)*, pages 439–449, 1992.

[56] D. Ross. *The Right and the Good.* Oxford University Press, 1930.

[57] T. Sandholm and V. Lesser. Leveled commitment contracts and strategic breach. *Games and Economic Behavior*, 35:212–270, 2001.

[58] L. Savage. *The foundations of statistics.* J. Wiley & Sons, New York, 1954.

[59] K. Schild. On the relationship between BDI logics and standard logics of concurrency. *Autonomous agents and multi agent systems*, 3:259–283, 2000.

[60] W. Spohn. An analysis of Hansson's dyadic deontic logic. *Journal of Philosophical Logic*, 4:237–252, 1975.

[61] W. Stelzner. Relevant deontic logic. *Journal of Philosophical Logic*, 21:193–216, 1992.

[62] R. Thomason. Deontic logic as founded on tense logic. In R. Hilpinen, editor, *New Studies in Deontic Logic: Norms, Actions and the Foundations of Ethics*, pages 165–176. D. Reidel, 1981.

[63] R. Thomason. Desires and defaults: A framework for planning with inferred goals. In *Proceedings of KR'00*, pages 702–713, 2000.

[64] L. van der Torre. Violated obligations in a defeasible deontic logic. In *Proceedings of the ECAI'94*, pages 371–375, 1994.

[65] L. van der Torre. The logic of reusable propositional output with the fulfilment constraint. In D. Basin *et.al.*, editor, *Labelled Deduction*, Applied Logic Series, pages 245–266. Kluwer, 2000.

[66] L. van der Torre and Y. Tan. The many faces of defeasibility in defeasible deontic logic. In D. Nute, editor, *Defeasible Deontic Logic*, pages 79–121. Kluwer, 1997. Revised and extended version of L. van der Torre and Y. Tan, Cancelling and overshadowing: two types of defeasibility in defeasible deontic logic. In: *Proceedings of the IJCAI'95*, 1525–1532, 1995.

[67] L. van der Torre and Y. Tan. The temporal analysis of Chisholm's paradox. In *Proceedings of the AAAI'98*, pages 650–655, 1998.

[68] L. van der Torre and Y. Tan. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence*, 27:49–78, 1999.

[69] L. van der Torre and Y. Tan. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law*, 7:51–67, 1999.

[70] L. van der Torre and Y. Tan. Rights, duties and commitments between agents. In *Proceedings of the IJCAI'99*, pages 1239–1244, 1999.

[71] L. van der Torre and Y. Tan. Contextual deontic logic: violation contexts and factual defeasibility. In P. Bonzon, M. Cavalcanti, and R. Nossum, editors, *Formal Aspects in Context*, Applied Logic Series, pages 143–160. Kluwer, 2000.

[72] L. van der Torre and Y. Tan. Dynamic normative reasoning under uncertainty: How to distinguish between obligations under uncertainty and prima facie obligations. In P. Smets and D. Gabbay, editors, *Agents, Reasoning and Dynamics. DRUMS HANDBOOK 6*. Kluwer, 2001. Extended and revised combination of the following two papers: L. van der Torre and Y. Tan, An update semantics for prima facie obligations. In: *Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI'98)*, 38–42, 1998. L. van der Torre and Y. Tan, An Update Semantics for Defeasible Obligations. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*, 631–638, 1999.

[73] L. van der Torre and Y.-H. Tan. Two-phase deontic logic. *Logique et Analyse*, To appear.

[74] L. van der Torre and E. Weydert. Parameters for utilitarian desires in a qualitative decision theory. *Applied Intelligence*, 14:285–301, 2001.

[75] H. Verhagen. *Norm Autonomous Agents*. PhD thesis, Stockholm University, 2000.

[76] G.H. von Wright. *The logic of preference*. Edinburgh University Press, 1963.

[77] G.H. von Wright. *An Essay on Deontic Logic and the General Theory of Action*. North-Holland Publishing Company, Amsterdam, 1968.

[78] G.H. von Wright. A new system of deontic logic. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, pages 105–120. D. Reidel Publishing Company, Dordrecht, Holland, 1971.

[79] E. Weydert. Hyperrational conditional logic. In *Proceedings of the ECAI'92 Workshop on Theoretical Foundations of Knowledge Representation and Reasoning*. Springer Verlag, 1992.

[80] R.J. Wieringa and J.-J.Ch. Meyer. Applications of deontic logic in computer science: A concise overview. In J.-J. Meyer and R. Wieringa, editors, *Deontic Logic in Computer Science*, pages 17–40. John Wiley & Sons, Chichester, England, 1993.