Is there a Logic of Imperatives?

Jörg Hansen

Institut für Philosophie
Universität Leipzig
Beethovenstraße 15, D-04107 Leipzig
jhansen@uni-leipzig.de

Abstract. Throughout its history, deontic logic had to face the question whether it is a logic of descriptions or a logic of prescriptions, namely of imperatives. The paper describes how the idea that there is a 'logic of imperatives' first came about, what proposals there have been to explain it and what problems it has had difficulties to solve. The paper argues that the idea of a logic of imperatives rests on a mistaken parallelism between imperative and indicative language and that there is, as a matter of fact, no such logic. However, we can argue about what ought to be done or need not be done according to given imperatives, and appeal to existing imperatives to motivate new ones. Descriptively interpreted deontic logic suffices to explain the reasoning involved.

Keywords: logic of imperatives, deontic logic

1 Introduction

For almost the whole of its over 50 years of existence, authors studying deontic logic and its concepts of obligation, permission and prohibition have been unsure as to what their subject really is. Is it the study of prescriptively used language (imperatives, norms)? Or is it the study of descriptive sentences about norms and the logical relations that hold between them? And if the second description is true, is not deontic logic then a kind of 'ersatz theory' that only mirrors what goes on in the realm of norms, a theory that may, if properly devised, result only in dull isomorphisms of the 'real' relations that hold between the norms themselves? Troubled by this prospect, many authors have tried to answer the question in the first way and upheld the point of view that deontic logic is the study of prescriptions, namely of imperatives, and their logical relations. In this paper I describe how the idea that there is such a thing as a 'logic of imperatives' came first into being, how authors have tried to explain this idea and devise formal theories for this logic, as well as the main problems that such proposals have run into. I finally argue that the idea of a logic of imperatives rests on a mistaken conception of a total parallelism between imperative and indicative language and that there are in fact, in ordinary language, no argument forms that resemble 'imperative inferences'. So there is also no place for a formal theory for such a logic. When we appeal to imperatives in arguments, we use deontic logic to describe what is obligatory or needs not obtain according to the imperatives that we have used or accepted before, and we motivate new imperatives by what is obligatory according to other imperatives that have been or should be accepted. Deontic logic, in its descriptive interpretation, is thus the (only) logic of normative concepts like obligation.

2 Beginnings: Poincaré's Proposal

Can imperatives, i.e. sentences in the imperative mood, be part of logical inferences? Henri Poincaré considered this question in his 1913 essay "La Morale et la Science" [87]. He begins by observing that if the premisses are all indicatives, then so will be the conclusion, hence for an imperative conclusion at least one premiss in the imperative mood is required, and so science alone cannot establish standards of morality. However, just as steam can be put to use in different machinery, science may also be used for moral reasoning:

> "It [the moral sentiment] will give us the major premiss of our inference which, as it happens, is in the imperative mood. At its side, science will put the minor premiss which will be in the indicative mood. From these a conclusion can be drawn that is in the imperative mood."

Poincaré seems to have in mind Aristotelian syllogisms of the following kind:

Hang all dwellers of the Nottingham Forest!
All members of Robin's band dwell in the Nottingham Forest.
Therefore: Hang all members of Robin's band!

Poincaré then proceeds to give a second example of an inference with an imperative conclusion:

> "One can imagine inferences which are of the following type: do this, but now if one does not do that, one cannot do this, so do that. And such reasoning is not outside the field of science."

The following is an example of the suggested inference:

Open the door!
The door cannot be opened unless it is first unlocked.
Therefore: Unlock the door!

So an order to do one thing includes orders to do all that is necessary to satisfy the primary command. Poincaré's proposals raise questions: by exchanging in his first example the syllogism *barbara* for *camestres* we obtain:

Hang all dwellers of the Nottingham Forest!
No member of Robin's band is hanged.
Therefore: No dweller of the Nottingham Forest shall be a member of Robin's band!

But it did not seem as if the speaker, e.g. the Sheriff of Nottinghamshire, was creating rules for band membership. The second type of inference is problematic when there are no legal means to fulfill an imperative (cf. Foot [21] p. 384):

Sustain your aged parents!
I can only sustain my aged parents if I rob somebody.
Therefore: Rob somebody!

So if commanding means also commanding all necessary acts, then even forbidden acts may be included. To improve matters, the second clause in Poincarés scheme might be changed to 'this can only *legally* be brought about by doing that'. But this introduces a normative element into a premiss that Poincaré assumed to be established by science alone.

While all this suggests that imperative inferences might require some additions and modifications, the most difficult question has turned out to be what makes them *inferences*. The problems attached to this question go under the name of 'Jørgensen's Dilemma'.

3 Jørgensen's Dilemma

Logic's concern is with the soundness of arguments, or inferences. These consist of sentences that represent the 'premisses' and usually one sentence that forms the 'conclusion'. The argument is then called 'sound', 'valid' or 'logical', if it is not possible that all of its premisses are true but the conclusion false. The premisses are then said to 'entail' the conclusion which thus 'follows' from them.[1] Expressions in the imperative mood are not, in any usual sense, true or false. Therefore they are incapable of functioning as premisses or conclusions in logical inferences. However, people maintain that the opposite is true and that there are inferences that have conclusions in the imperative mood and premisses of which at least one is likewise in the imperative mood (cf. Poincaré's examples above). This is a puzzling situation, which was first pointed out by Jørgen Jørgensen in [48]:

> "So we have the following puzzle: According to a generally accepted definition of logical inference only sentences which are capable of being true or false can function as premisses or conclusions in an inference; nevertheless it seems evident that a conclusion in the imperative mood may be drawn from two premises one of which or both of which are in the imperative mood. How is this puzzle to be dealt with?"

To find this puzzle, called 'Jørgensen's Dilemma', perplexing, one must accept that imperatives cannot be meaningfully termed true or false. This seems to be the philosophical consensus, it can point to Aristotle's definition of an assertion as a grammatical entity that can be true or false, in distinction to other

[1] For such textbook definition cf. Mates [71] p.5, Lemmon [64] p.1, Hodges [44] p.55.

grammatical entities like requests that are neither true or false (*De interpretatione* 17 a 4). Nevertheless, a way out of the dilemma may consist in giving up just this claim. Most prominently, Kalinowski [49], [52] has argued that in the case of expressions of moral or legal norms, the attitude of the 'ordinary', non-philosophical person is to treat these as true or false. E.g. people say that it is true that another person's right to live must be respected, or that slander is prohibited, and people would uphold these truths even if particular legislators did not enact such norms, or proclaimed otherwise. So Kalinowski concludes that legal or moral norms can be part of logical inferences. I think that these considerations confuse truth with the notion of a legal or moral norm's validity: the 'external' recognition of a norm as valid in a certain society. For the present discussion it suffices that Kalinowski himself restricts his view to legal and moral norms and does not claim that 'imperatives in the strict sense' can be said to be true or false, and in fact writes that they are not true or false.[2] But it is these that we are concerned with.

On another view, any imperative can be equivalently replaced by a first person expressions like 'I command you to ...', 'I order you to ...', 'I request of you that you ...' or 'I want you to ...'.[3] But expressions like "I command you to ..." cannot only be used to command, but also to assert that I do in fact command so-and-so. Therefore Sigwart [94] claims that each imperative includes the statement that the speaker wills the act which he commands. While he adds that nevertheless the import of what is said by the imperative is not the communication of a truth but a "summons to do this, to leave that undone", Ledig [61] argues that because imperatives include such assertions, one must consequently apply the terms of truth and falsity to an imperative, and that it will be true unless it is e.g. stated for fun. Likewise, Kamp [54] points out that from the viewpoint of the addressee, it really makes no difference if an expressions like "I command you to ..." is understood as a command or as an assertion, since if the utterance is appropriate (i.e. made in earnest in the appropriate conditions) then it *must be true* and so the practical consequences are the same. Similarly, Opałek & Woleñski [79] call a normative statement *qua* performative true if it is effective. However, it is hard to see how these considerations can solve Jørgensen's dilemma. If any imperative that is meant seriously is true, then (i) any imperative either follows from any other 'true' imperative or is not meant seriously, and (ii) all imperatives follow from one made for fun. If imperatives are ambiguous and include an assertion, then the fact that this assertion may be used in an indicative inference does not mean that the indicative conclusion is likewise ambiguous. A sign post bearing the words 'Frankfurt/Main' conveys the information that if I follow the indicated direction I will eventually arrive at Frankfurt. I may also infer that there exists a place called Frankfurt and that a geographical entity called Main, namely

[2] Cf. Kalinowski [50] p. 36 and [52] p. 107.

[3] In the case of English grammar, it has been claimed that the exclamation mark, or rather the characteristic intonation that it replaces in written language, is only the remainder of such explicit performative lead-ins, cf. Harris [39] pp. 391–392.

a river, exists in its proximity. Knowing that sign posts tend not to include redundant information, I can also conclude that there must be another place that is also called Frankfurt (namely Frankfurt/Oder), and that this place does not lie near the Main, because otherwise the extra information would not have been discriminating. But the sign post does not point into the direction of this other city. Similarly the information that is obtained by interpreting an imperative utterance may be used to infer some other information. But this is not inferring imperatives.

The fact that imperatives are traditionally not considered to be true or false finds its explanation in the different intentions in which imperatives and indicatives are used. The main use of indicatives is to convey what the speaker believes the world to be like. If it is so, then the sentence is called 'true', if not, then it is called 'false' and the recipient might point out that I should perhaps change my beliefs. By use of an imperative I tell the addressee what I want to be done. If the addressee does what is demanded, the action may be qualified as 'right', or satisfactory with respect to the command, and if not, then the behavior of the addressee is in some sense 'wrong' and I will perhaps remind the agent of his or her obligation. So truth and falsity are the qualities of descriptions when things are or are not as they have been described, while 'right' or 'wrong' are the qualities of acts that are or are not in accordance with what has been prescribed. Descriptions and prescriptions have a different 'direction of fit', and true/false are the terms used to express the match/mismatch on the language side in case of a descriptive use of language, and right/wrong are the terms employed for the match/mismatch on the world side in case of a prescriptive use.[4] Therefore it is a confusion of language, and indicates a misunderstanding of the intention in which the sentence has been uttered, if imperatives are termed true or false.

Accordingly, the most effort regarding Jørgensen's dilemma has been spent on developing alternative definitions for 'imperative inferences', rather than arguing for the application of the terms of truth and falsity to imperatives – unless one is already convinced by the dilemma that such things as inferences with imperatives are at all impossible (e.g. Keene [56]).

4 Dubislav's Trick and Related Theories

To deal with his own 'dilemma', Jørgensen [48] endorsed a proposal by Walter Dubislav [18] to transfer the 'usual definitions' of inferences between indicatives to imperatives 'by analogy'. Dubislav gives the following example:

> Though shalt not kill.
> Therefore: Cain shalt not kill Abel.

[4] This explanation of why the terms of truth and falsity are not applicable to normative uses of language originates with Anscombe [7] §32. Independent accounts can be found in Kenny [58] p. 68 and Peczenik [81], [82] who speaks of the norm as a 'qualifying utterance'. The dual terms right/wrong are used as corresponding qualifications e.g. by Engliš [19] and Kelsen [57] p. 132.

Here, he argues, the analogue of the following 'ordinary' inference is applied:

> No human being kills any other human being.
> Cain and Abel are human beings.
> Therefore: Cain does not kill Abel.

Dubislav observes that to each imperative belongs a descriptive sentence that describes the state of affairs that obtains if the subjects of the imperative realize what the commanding authority demands. The formalisms of descriptive inferences are then transferred to imperatives by what he calls a 'trick' (*Kunstgriff*): imagine the state that the commanding authority desires realized, describe it, from this description infer some other descriptive sentence, which is then again interpreted as describing a state the authority wants to see realized. He then proposes the following *convention* (DC) on the meaning of imperative inference (also see the next figure):

> **(DC)** "An imperative F is called derivable from an imperative E if the descriptive sentence belonging to F is derivable with the usual methods from the descriptive sentence belonging to E, whereby identity of the commanding authority is assumed."

The convention is illustrated by the next figure (where I write $!A$ for an imperative to which 'belongs' the descriptive sentence A):[5]

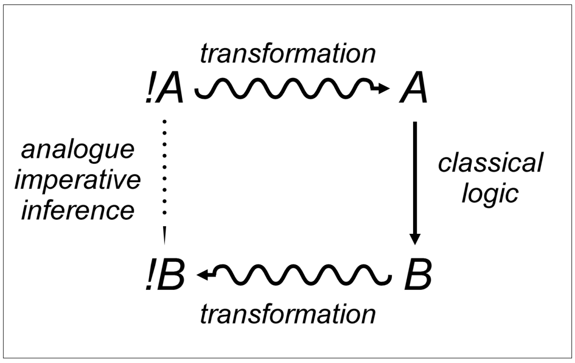


Fig. 1. Dubislav's convention (DC).

[5] Mally [70] p. 12 seems to have introduced the symbolism $!A$, which was then employed by Hofstadter & McKinsey [45] for the imperative that demands that A be the case. However, Mally intended $!A$ to be interpreted theoretically, as an assertion or assumption that 'A ought to be', which we now call a deontic proposition and formalize by OA.

Dubislav's convention does not cover inferences with more than one imperative premiss, though he mentions this possibility.[6] For such inferences, (DC) can be modified as follows:

> **(DCM)** An imperative F is called derivable from the imperatives $E_1, ..., E_n$ if the descriptive sentences belonging to F is derivable with the usual methods from the descriptive sentences belonging to $E_1, ..., E_n$, identity of the commanding authority presupposed.

Dubislav then proceeds to inferences in which the imperative premiss is accompanied by another one in the indicative mood, and where the conclusion is again an imperative, for which he extends his convention:

> **(DEC)** "An imperative F is called derivable in the extended sense from an imperative E if the descriptive sentence belonging to F is at least jointly derivable from the descriptive sentence belonging to E and true descriptive sentences that are consistent with the first."

This extended convention (DEC) may again be modified to facilitate inferences with more than one imperative premiss to produce the following modified extended convention

> **(DECM)** An imperative F is called derivable in the extended sense from imperatives $E_1, ..., E_n$ if the descriptive sentence belonging to F is at least jointly derivable from the descriptive sentence belonging to $E_1, ..., E_n$ and true descriptive sentences that are consistent with these.

Jørgensen [48] endorsed Dubislav's proposal as one way to deal with his dilemma and states that it seems clear to him that any imperative sentence has an indicative parallel-sentence which describes the contents of the command or wish. Jørgensen suggests that an imperative consists of an *imperative factor* and an *indicative factor*, where the first indicates *that* something is commanded, and the second *what* is commanded. The indicative factor can then be separated from the imperative and formulated in indicative sentences describing the action, change or state of affairs which is commanded. Applying the rules to these latter sentences we can thus indirectly apply the rules of logic to the imperative sentences to make their entailments explicit. Jørgensen's concept was in turn further refined by R. M. Hare [36], according to whom an imperative sentence and an indicative sentence 'correspond' if they have the same 'descriptor', but different 'dictors', where what is described by the descriptor is what would be the case if the sentence is true or the command obeyed, and the dictor is what does the saying or commanding.[7] This is still not much different from Jørgensen's

[6] Cf. Dubislav's use of the plural when stating that "an inference from demand-sentences will now be formally facilitated by the following convention", and Dubislav's summary, in which he stresses that no demand-sentence can be derived from premisses that do not contain *at least one* demand-sentence.

[7] This distinction was anticipated by Ledig [62], who wrote that norms and descriptions have an isolable imaginary content (*isolierbarer Vorstellungsinhalt*). Hare's [37] later terminology of a 'neustic' and 'phrastic' mirrors his earlier distinction.

analysis, but Hare finds it is misleading to speak of an 'indirect', 'parallel' or 'analogous' application of logic. Instead, in Hare's view imperatives are logical in the same way as indicatives; he argues that "most inferences are inferences from descriptor to descriptor and we could add whichever set of dictors we pleased". Since most logical reasoning is done with descriptors only - this Hare calls the 'principle of the dictive indifference of logic', there is no special need for a logic of imperatives. Rather, all logic is recast as a logic of descriptors, where if the descriptors of the premisses describe a state of affairs, then the descriptor of the conclusion describes, at least partially, the same state of affairs.

It is immediate that neither Jorgensen's nor Hare's account make a material difference for what imperative inferences should be accepted. Hare's concept of a 'dictor' that operates on a 'descriptor' poses problems: grammatically, it is hard to see how dictors can be removed from sentences, or exchanged in them, so that the remainder or the new composite is a meaningful expression.[8] So there may be reasons not to follow Hare's analysis, but – closer to Dubislav's original concept – speak of a thematically parallel sentence in the indicative mood. However, this way or another, the idea of a descriptive sentence that parallels an imperative or of an imperative's indicative factor has become the most successful part of the Dubislav-Jørgensen-Hare analysis. Ross [92] calls this element the 'theme of demand', a state the realization of which is requested by the demand, and proposed that to any imperative corresponds an ordinary indicative sentence which contains a description of the imperative's theme of demand. Frey [22] p. 440 uses the term *Erfüllungsaussage* for the parallel sentence that indicates if the imperative is satisfied or violated, which Rescher [88] p. 52 calls a 'command termination statement' and Keene [56] the 'actualization' of the imperative. Geach [24] states that for every imperative there is a future-tense statement whose 'coming true' is identical with the fulfilment of the imperative, Sosa [96], [98] speaks of the 'propositional core' of an imperative and Hanson [35] of a state s the commanding agent 'envisages', of which then a description S is used. Von Wright (e.g. [131] p. 269) calls the state the norm 'pronounces' as obligatory or allowed the 'content of a norm'. The seemingly universal consensus is explained by the pragmatic function of imperatives, that is, the regulation of human behavior: if there were no imperative-correlated indicative sentences, it could not be understood *what* ought to be, and neither would it be possible to determine whether the norm is satisfied or violated.[9] This is why even 'anti-reductionist'

[8] Opałek [77] points out that even if the imperative is rephrased as 'I command that ...' or 'it is obligatory that ...', the '...'-part is a Latin *ut*-expression that only due to a peculiarity of English grammar may be confused with an indicative sentence, also cf. Opałek [78] ch. 2, Kalinowski [51] and Rödig [90] for similar criticism. On the other extreme, Leonard [65] has argued that it is the descriptor that is called 'true' and 'false', and so imperatives share these properties with descriptive sentences.

[9] This explanation and the formulation of the principle below is most clearly expressed in Weinberger [122] p. 172. Weinberger uses the term 'coordination' instead of 'correspondence', but this suggests an onto or even one-to-one mapping, and the uncertainty whether tautologies and contradictions can be thus 'coordinated' to imperatives first made Weinberger [109] p. 121 doubtful about (W). Later, he proposes

authors that oppose the idea that imperatives or imperative reasoning can be reduced to indicatives or indicative logic, agree on the following principle:[10]

Principle (W) (Weinberger's Principle).
To each imperative there corresponds a descriptive sentence that is true if the imperative is satisfied and false if it is not-satisfied (violated).

It is clear that an acceptance of (W) does not force us to also accept the Dubislav-Jørgensen-Hare account of imperative inference, which nevertheless has been accepted by a number of authors.[11] (W) can then be used to show that seemingly differing explanations of imperative inferences are in fact equivalent to this account. Thus Rescher [88] defines command inferences in terms of satisfaction in the following way:

> A command inference is valid if there is no possible world in which the premisses are all satisfied and the conclusion fails to be satisfied.

which, using (W), is equivalent to

> A command inference is valid if there is no possible world in which the descriptive sentences corresponding to the premisses are all true and the descriptive sentence corresponding to the conclusion is false.

Using the textbook definition of an argument, this is equivalent to

> A command inference is valid if the descriptive sentences corresponding to the premisses entail the descriptive sentence corresponding to the conclusion.

which in turn is the modified Dubislav convention (DCM).[12]

(W) on the condition that the coordinated indicative is not contradictory ([112] p. 17 and p. 19) and not tautological ([118] p. 229-231), also cf. [116] p. 69.

[10] Besides Weinberger cf. Hamblin [29] pp. 151-152: "Take the exact words of the imperative, and transform them into indicative mood (...) Now the worlds which extensionally satisfy the imperative are just those of which the description is true.", and Moutafakis' theorem T3 ([75] p. 155), which expresses the equivalence of the statements that an imperative is satisfied and that a description of the prescribed action as performed is true.

[11] These include Simon [95], who converts commands to declarative mode "by removing the imperative operators from them", obtaining a theory in which all recipients obey the commands, and then applies the 'ordinary laws of logic' to derive new relations that may be converted back into commands. According to Niiniluoto [76], an "imperative $!p$ entails imperative $!q$ if p entails q". Very close to (DCM) is von Wright's account in [129] pp. 71, 164, where he defines the content of a prescription as 'the prescribed thing', and defines that a command is entailed by a second command or by a set of commands if the content of a command is a consequence of the conjunction of the content of a command with the contents of none or one or several other commands.

[12] Sosa's [96], [98] definition of a 'directive argument' is very similar to Rescher's, but additionally demands that the imperatives that function as premisses are jointly

Using the idea that norms qualify the states of affairs that satisfies or violates them as 'right' and 'wrong', one can define:[13]

> An imperative $!A$ entails an imperative $!B$ if and only if (iff) every state of affairs that is qualified as wrong by $!B$ is qualified as wrong by $!A$.

which can then be translated into

> An imperative $!A$ entails an imperative $!B$ iff every state of affairs that violates $!B$ also violates $!A$.

which using (W) is equivalent to

> An imperative $!A$ entails an imperative $!B$ iff every state of affairs in which B is false also makes A false.

which using classical logic is equivalent to

> An imperative $!A$ entails an imperative $!B$ iff every state of affairs in which $\neg B$ is true also makes $\neg A$ true.

which using *tertium non datur* and *modus tollens* is equivalent to

> An imperative $!A$ entails an imperative $!B$ iff every state of affairs in which A is true also makes B true.

which by definition of entailment equals

> An imperative $!A$ entails an imperative $!B$ iff A entails B.

and this is again Dubislav's convention (DC).

Lemmon [63] defines an entailment relation for imperatives *via* a definition of inconsistency of a set of imperatives and indicatives, where such a set is called inconsistent if it cannot be the case that all indicatives are true and all imperatives obeyed.[14] Lemmon's entailment relation is then defined as follows:

satisfiable in order to cope with normative conflicts. Sosa adds a second condition, demanding that if the imperative conclusion is violated, at least one imperative premiss must be violated, in order to also cope with conditional imperatives. This equals Keuth's [59] condition B_1 that it must be logically impossible to violate the conclusion without violating a premiss. Obviously, the second condition makes no difference for categorical imperatives.

[13] The definition is similar to the one used by Peczenik [81], [82] for forbidding norms and the quality 'forbidden'. Kamp [53] uses an analogous definition for permissions, where one permission entails another if the second makes only such courses of actions permissible that were already so before.

[14] Lemmon expresses reservations regarding his definition, but only because he thinks that it does not sufficiently restrict imperative conclusions to statements about future actions. A definition similar to Lemmon's seems to be intended by Philipps [86] p. 364 who defines: 'to do p is forbidden!' is true iff the indicative 'someone does p' is incompatible with the class of valid prescriptions, where 'compatible' means that if the indicative is true, then at least one prescription is violated.

> An imperative $!A$ is entailed by a set indicatives and imperatives if this set together with $!\neg A$ is inconsistent.

where $!\neg A$ is the imperative that is satisfied if and only if A is false. Using (W) we obtain:

> An imperative $!A$ is entailed by a set indicatives and imperatives if the set of indicatives together with the set of descriptive sentences corresponding to the imperatives together with $\neg A$ is inconsistent.

which, using classical logic, is equivalent to

> An imperative $!A$ is entailed by a set indicatives and imperatives if the set of indicatives together with the set of descriptive sentences corresponding to the imperatives entails A.

and this is Dubislav's modified extended convention (DECM). This shows that, with (W), Dubislav's proposal is equivalent, or at least very similar, to a number of other proposals how imperative inferences are possible in the face of Jørgensen's dilemma.

5 Explanations of Imperative Inferences

Dubislav's 'trick' provides a formal method that explains *how* inferences between imperatives can be defined without having to assign them truth values. What has not been made clear is *what* is achieved by the method, i.e. why we should think that this is what it means to 'infer' an imperative from some other imperative or a set of imperatives and indicatives, or formally, what is means that some imperative inference scheme

(ImpInf) $!A$
$\therefore\ !B$

is valid. Dubislav's trick can easily be applied to e.g. sentences of the form 'I doubt that ...'. Then from 'I doubt that he is staying at his sister's place in San Francisco' follows 'I doubt that he is staying in San Francisco', which, though we can derive 'he is staying in San Francisco' from 'he is staying at this sister's place in San Francisco', seems wrong: I might not doubt that he is staying in San Francisco, but doubt very much that he is staying with his sister. So why should Dubislav's trick work for imperatives if it would not for other expressions?

5.1 Logic of Satisfaction

On one interpretation, which has been called the 'logic of satisfaction', the scheme (ImpInf) is understood as stating that if the imperative sentence $!A$ is satisfied, then it must be that the imperative sentence $!B$ is also satisfied. This interpretation is usually attributed to Hofstadter & McKinsey [45], whose formalization of the scheme (ImpInf) would be $!A >!B$, which is derivable in their axiom system whenever $A \to B$ is classically derivable. It is immediate that (ImpInf) must be valid on this interpretation whenever the 'ordinary' argument

$$\begin{array}{l} A \\ \therefore\ B \end{array}$$

is valid for the descriptive sentences A and B: If A classically implies B, then it must be that if $!A$ is satisfied, then A is true and also B is true and hence $!B$ is satisfied. Thus Dubislav's 'trick' receives its semantic justification. But if (ImpInf) is interpreted in this way, then it seems one should also accept the following scheme:

$$\begin{array}{l} !A \\ \therefore\ A \end{array}$$

According to our (informal) convention, $!A$ represents an imperative sentence that is satisfied iff A is true. So it must be that if $!A$ is satisfied, then A is true and so the above scheme is valid. But on the look of it, this scheme seems to state that from an imperative that demands A it can be inferred that A is the case, which is nonsense. And this misunderstanding reveals that when we spoke of the possibility of an inference in which the premisses and the conclusion are imperatives, it seems that we talk about *inferring an imperative* from some other imperatives, and not about reasoning whether or not the imperatives in question are satisfied. So though the inferences of a 'logic of satisfaction' are valid in the interpretation in which they were intended, thus interpreted inferences seem not to be what we want from a logic of imperatives. For these reasons, Ross [92] p. 61 and also Hare [38] doubted that a logic of satisfaction is what one has in mind in the case of practical inferences.[15]

5.2 Logic of Existence

When we speak of inferring one imperative from some other imperative, this could mean that the existence of an imperative is logically deduced from the existence of some other imperative.

What is meant by saying that an imperative 'exists'? First, it could be the existence of an utterance of some sentence in the imperative mood by some commanding agent towards some commanded subject.[16] Second, one might demand that the utterance, as a performative use of the imperative sentence, was effective and established an *imperativum*, i. e. a 'command', 'demand', 'request'

[15] Kanger [55] p. 49 and Føllesdal & Hilpinen [20] p. 7 criticize Hofstadter & McKinsey for making $!A$ and A 'equivalent', which is somewhat unfair since the intended interpretation of their formulas (in terms of satisfaction) is not presented.

[16] This existence is what Frey [22], along with an additional property of 'justification', infers in imperative inferences: "If the imperatives that appear in the premisses exist and are justified, then also the imperatives derived from these exist and are justified" (p. 465). Frey's 'justification' means that what is demanded is 'good' regarding some aim of the commanding agent, called 'axiological validity' in Ziembiński [136].

or the like. For this it may be required that the commanding agent had the will to command (and did not use the words for fun) as well as some authority (power to punish or reward) over the addressee.[17] Third, for an order by legal authorities in this capacity to come into 'legal existence' it may be required that the authority was competent to utter it according to the legal rules of some normative system that confers such competence, and similar for bodies that are constituted not by law but by other rules like a firm or Robin's band.[18]

Yet however much the concept of existence is thus refined, it seems to require the presence of actual facts: a (still alive?) speaker, a linguistic entity like an utterance and circumstances of speaking, a certain attitude of the speaker towards the act of speaking, a backing of the speaker by force or an authority conferred by existing and/or valid rules, etc. But it is difficult to see how logic can stipulate such an existence. This is illustrated in the following example by Aleksander Peczenik:[19]

> "The premiss 'love your neighbour' may be regarded as describing the fact that the authority – Jesus – has in fact said 'love your neighbour.' The imperative existed because it was uttered by Jesus. But the conclusion, for example, 'love Mr. X' does not describe anything which in fact has been said by Jesus."

Here, the intended argument from 'love your neighbour' to 'love Mr. X' is not accepted because the commanding agent 'did not actually say' what appears in the conclusion, and so unlike the premiss the conclusion did never 'exist' as a fact. An imperative sentence that has not been expressed was not received and cannot be understood by its addressee as a command or legal order. Thus the required 'existence' of the imperative, or 'validity' of the command seem to be the analogues not of truth of a descriptive sentence, but of 'stated' and 'asserted'. Yet indicative logic does not force anyone to state or assert anything, even if some other descriptive sentence was used in a way that expresses ones commitment to it. It only explains what people mean when they use an indicative sentence in order to assert some fact, by saying what other sentences must be true if the stated sentence is true.[20] Because the imperative sentence in the conclusion may not exist as an utterance, Hamblin [29] p.89 warns against speaking of

[17] Cf. von Wright [129] p. 120-126. This is Ziembiński's [136] 'thetic validity'. Lemmon [63] seems to have this notion of 'validity' or 'existence' in mind when he demands that the entailment of imperatives must be defined in terms of what imperatives are *in force* at a given time.

[18] Bulygin [13] uses the term 'systemic validity'. According to Weinberger [115], [118] p. 259, this validity takes the place of truth as the 'hereditary trait' (*Erbeigenschaft*) that is transferred from the premisses to the normative conclusion in inferences with normative sentences (*Normsätze*).

[19] Quotation from a letter by A. Peczenik to R. Walter, printed in [108] p. 395

[20] Kenny [58] first pointed out that 'valid' and 'invalid', interpreted as meaning 'commanded' and 'not commanded', are not the analogues of 'true' and 'false', but of 'stated' and 'not stated'.

inferences between imperatives. Because the ‘telling part’ (or attitude) of the speaker cannot be inferred, the possibility of command inferences was denied by Sellars [93] p. 239-240, and for the same reason, Lemmon’s [63] attempt to define imperative inferences via the notion of an imperative’s being ‘in force’ was rejected by Sosa [98] p. 61 who argues that such notions involve attitudes by the authority or its subject that cannot be inferred. Von Kutschera [60] argues that a (used) imperative is an action, actions do not follow from actions, so there is no logic of imperatives. In Alchourrón & Bulygin’s [3] ‘expressive conception of norms’, the existence of a norm is seen as dependent on empirical facts and the possibility of a logic of norms is consequently denied as there are nor logical relations between facts. For similar reasons, Philipp [83], [84] denied both the possibility of a logic of imperatives and of norms.

5.3 Logic of Metaphysical Existence

To get around this difficulty one may consider to interpret ‘existence’ not with respect to natural facts, but with respect to some ideal ‘world of ought’ or an assumed ‘normative system’ that is closed under consequences, where the closure operation may be understood e.g. in the sense of derivability by use of Dubislav’s convention. So if the agent’s use of the imperative mood has resulted in the existence of a command in the ‘world of ought’, or, due to the agent’s legal competence as e.g. a police officer, created an order that now belongs to the normative system, then all the ‘consequences’ of the command that can be derived by an appropriate method exist in this world or system as well.[21] What thus has ideal existence is not the imperative sentence as a spatial and temporal phenomenon or as a grammatically correct or meaningful combination of words. Rather, it is what the use of an imperative sentence expresses or accomplishes – a command, request etc. Then it must be that not only commands, requests etc. ‘exist’ in this sense that in fact *have been* expressed by a performative use

[21] The ‘world of ought’ terminology originates with Walter [107], who is however following Kelsen [57] p.195 in that an individual norm does not exist before the general norm was applied by a judge, so orders that can ‘only be deduced’ do not exist in Walter’s ‘world of ought’. The idea to explain logical relations between ‘norm sentences’ (like imperative sentences) in terms of their existence in a ‘system of norms’ that is closed under consequences is that of Stenius [99]. In Opalek & Wolenski [80], norms are non-linguistic entities expressed by (descriptively interpreted) deontic statements, and normative systems consist not only of norms that have been expressed by a normative authority, but also of the consequences of these ‘basic obligations’. In Alchourrón & Bulygin’s ‘hyletic’ variant of a conception of norms [5], ‘implicitly promulgated’ norms have ‘existence’ in a logically closed normative system, and descriptively interpreted (deontic) norm propositions are then “propositions about the existence of norms (in that system)”. Holländer [46] promotes the idea of a ‘deontically perfect world’ where norms exist that obey logical principles, like that conflicts are excluded. Kelsen [57] pp. 187–188 rejects the idea of an ‘ideal existence’ of norms because there is no ‘ideal’ act of will that creates them, and rejects the whole idea of a logic of norms.

of a sentence in the imperative mood, but also some that only *can* be expressed. For if all that 'exists' in the 'normative system' already exists as a result of a pragmatic use of language, then there would be no need to let the 'normative system' e.g. be closed under a consequence operation. That what we can express by using language (commands, requests, assertions etc.) has some existence, possibly independent from any pragmatic origin[22], in some ideal 'world of ought', is a difficult concept that possibly creates more problems than it solves.[23] But it is even difficult to see that it solves the problem of entailment between imperatives. For to say that some ideal object created by use of an imperative implies the existence of some other ideal object in some 'world of ought' or normative system, is not to say that an imperative implies another imperative. The existence of a forest might imply the existence of a tree, but to say that 'the forest implies the tree' is making a categorical mistake. The first is an indicative argument, which can be formalized in the usual way:

(1) $\exists x : \mathit{Forest}(x)$
$\therefore\ \exists y : \mathit{Tree}(y)$

The argument is analytical when the words 'forest' and 'tree' have their usual meaning and for all that understand this meaning and thus know that there cannot be forests without trees. By starting to talk about the (ideal) existence of commands it seems that we silently changed (ImpInf) into

(2) the command given by α to x by the use of the imperative sentence $!A$
Therefore: the command given by α to x by the use of the imperative sentence $!B$

which appears confused. This is because the argument form is not used as it is usually used, and now we do not know what to make of it. We are used to filling in the blanks of the argument form

(3) $\underline{\hspace{5cm}}$
$\therefore\underline{\hspace{4.6cm}}$

with sentences. Whether also imperative sentences can be meaningfully used to fill in the blanks is the open question. However, there is no pre-established usage of filling in the blanks with names of objects, as in

(4) a
$\therefore\ b$

where a means a forest and b means a tree. At most, this is a mistaken way to try to express (1). Similarly, the scheme (2) must be corrected into (5):

[22] Cf. Stenius [99] according to whom all normative systems include a norm that demands a tautology.

[23] Note that the topic of this discussion has not suddenly become the ontological status of notoriously difficult concepts of practical philosophy and jurisprudence, like moral obligations, natural law, human rights, laws of custom etc. Our concern are still ordinary sentences in the imperative mood, addressed e.g. to a husband, secretary, student, child or dog (cf. Ziemba [135] p. 386).

(5) There exists a command given by α to x by the use of the imperative sentence $!A$.
Therefore: There exists a command given by α to x by the use of the imperative sentence $!B$.

Now we have arrived at something that resembles an argument – though we have not yet shown when such a scheme constitutes valid arguments. But this argument is one that uses only descriptive sentences that can be true or false. It is not a case of (ImpInf), i.e. not an argument where imperative sentences function as premisses or conclusion.[24] So by appealing to the notion of existence we obtain at most an inference relation between sentences that describe the existence of certain commands, but not a logic *of* commands.

5.4 Formalistic Approaches

Maybe we should have gone looking for the explanation of an entailment relation between imperatives not in some relation between entities in a 'world of ought', but in the meaning of the sentence that the speaker uses. Someone who says that 'either Barack Obama or Hillary Clinton will be the 44th president of the U.S., but it won't be Obama', thereby expresses the belief that Hillary Clinton will be the 44th U.S. president. The speaker did not say explicitly that it will be Clinton (these were not the words used), but the speaker can be said to have implicitly, or tacitly, expressed this opinion. Therefore if a person asserts something, she may be taken to 'implicitly' assert something else. Likewise, when a person commands something, she may also be 'commanding something by implication'. Consider the following example employed by Hare [38]:

Go via Coldstream or Berwick!
Don't go via Coldstream!
Therefore: Go via Berwick!.

Here, an officer who must go from London to Edinburgh is ordered 'go via Coldstream or Berwick,' and – a bit later – is given the order 'don't go via Coldstream'. Both commands have been taken to imply that the officer is (now) ordered to go via Berwick, and that the inference is therefore valid. For the question *what* commands should be considered to have been 'implicitly commanded' by explicitly given commands, one may then point to Dubislav's conventions or similar rules – e.g. (DCM) clearly accounts for the above inference.[25] To search

[24] That the 'world of ought' approach thus only provides arguments with descriptive sentences is accepted by Walter [107], for he turns to identify imperatives with descriptions of the 'world-of-ought'-existence of a command that is created by the use of an imperative – consequently such sentences can be true or false and therefore part of logical inferences. Thus imperative logic is reduced to indicative logic, where the difficult part is now the verification of some descriptive sentences.

[25] The term of 'commanding something by implication' was introduced by Geach [25]. Alchourrón [2] speaks of the consequences of what is prescribed as 'indirectly pre-

for an explanation of imperative inferences by interpreting explicitly given commands seems a much better idea than to find it in otherworldly relations between metaphysical objects. Unfortunately, it seems also a circular idea: according to it one can infer a command from another command if by giving the latter command one implicitly commands the former. Or: we may use Dubislav's convention to infer imperatives from other imperatives because it provides the imperative sentences that are implicitly commanded when the first imperative sentences are used for commanding.

An interesting response to this reproach of circularity is to say that by giving rules for inferring one imperative from another, one *did all* that is required to explain the meaning of imperative inferences. It is this view that seems to lie at the root of Dubislav's proposal to "formally facilitate inferences" from demand-sentences through a 'convention' (*Übereinkunft*) or 'trick' (*Kunstgriff*). In fact, in his main work *Die Definition* [17], Dubislav gives a similarly 'formalistic' characterization of propositional logic (and later predicate logic). There, Dubislav starts with a 'pure, game-like calculus' that is played with 'pieces' and signs ('$\neg$', '$\vee$', brackets) by first arranging some into initial positions and then replacing pieces and re-arranging pieces into new game positions according to the rules of the game. This game then *becomes* the calculus of propositional logic by interpreting its elements as indicated: the 'pieces' as propositions, the signs as 'not', 'or', and brackets, the 'initial positions' as axiomatic basis, the game rules as the usual rules of substitution and *modus ponens*, and the achievable game positions as derivable formulas. This characterization of propositional logic is meant by Dubislav as an exposition of Boole's [12] idea that the "validity of the processes of analysis does not depend upon the interpretation of the symbols which are employed, but solely upon the laws of their combination". In Dubislav's view, which he calls 'the formalistic theory', this description of logic functions as a mould for all scientific theory: a theory is constituted by a pure calculus (of formulas and rules), combined with a fixed interpretation. Observational sentences are captured in formulas that can be used alongside axioms or derivable formulas of the system to derive other formulas within the calculus. Then the assignment of these derived, or better: 'calculated' formulas is reversed, i.e. they are translated back into observational sentences. If these are regularly true, then the observational sentences are 'explained' by the theory. If a calculated obser-

scribed', Alchourrón & Bulygin [3] write that e.g. if teacher commands that all pupils should leave the class-room, he also implicitly commands that John (who is one of the pupils) should leave the class-room, even if the teacher is not aware of the fact that John is there, and in [5] they view the 'deductive consequences' of norms as 'implicitly promulgated', where the deduction process is equivalent to the modified Dubislav convention (DCM). Hare [38] and Rescher [88] both propose to define command inferences in terms of 'implicitly given commands' – analogously, Rescher's 'assertion logic' [89] is concerned with assertions that a speaker is 'implicitly committed to' in virtue of overtly made assertions. It was shown above that Rescher's explanation of imperative inferences is equivalent to the modified extended convention (DECM).

vational sentence turns out to be false, then the theory is erroneous. Thus it also becomes possible to decide between competing, non-isomorphic theories.

The usefulness of the Dubislav's formalistic approach for the problem of imperative logic is immediate. In fact, Dubislav's own proposal in [18] satisfies all requirements in [17] for being a theory of imperative inference: there are entities that may function as premisses and conclusions, namely imperative sentences. There is an interpretation that assigns each imperative sentence a formula, namely that of the indicative 'parallel sentence' in the calculus of 'ordinary logic'. There is a calculus, namely 'ordinary logic', that tells us what formulas can be derived from the formulas assigned to the imperative sentences that function as premisses. And finally, this assignment is reversible to provide derived imperative sentences. Other authors – taking their cues from Tarski's [103] syntactical definition of consequence relations and deductive systems,[26] Tarski's [104] definition of truth,[27] Gentzen's [26] idea that to define a symbol is to give rules for its introduction and elimination,[28] and Wittgenstein's dictum that the meaning of a word is its use ([127] §43) – have similarly argued that instead of searching in vain for analogues of truth values, it suffices for an explanation of imperative inferences to give formal rules for obtaining imperatives from other imperatives.

If this 'formalistic' approach to the logic of imperatives is accepted, we are still not finished yet. If the assignment of formulas, calculations and back-translations of derived formulas are to be more than a game, there must be some way to judge the adequateness of the theory, and be it only to decide between competing proposals.[29] In analogy to Dubislav's general approach, where a the-

[26] Cf. Alchourrón & Bulygin [5] who employ a formal consequence relation to explain what norms are 'implicitly promulgated' by a set of norms.

[27] Both Rödig [91] and Yoshino [133] appeal to Tarski and argue that meaningful operations with prescriptions are made possible by supposing that normative attributes like 'obligatory' or 'punishable' may be applied to actions. Rödig draws attention to the problem of objective verifiability and therefore truth of such statements. But he circumvents the problem by assuming that meta-language truth conditions *can be* given, which is sufficient to handle normative attributes as normal predicates in the object language. Rödig and Yoshino then use these predicates to formalize e.g. a norm that says that helping in an emergency situation is obligatory as $\forall acts$: $In\_emergency(act) \wedge Helping(act) \rightarrow Obligatory(act)$. The puzzling thing is that if this really is a prescription (norm), i.e. it makes so far unregulated acts of helping in cases of emergency obligatory, then for no such act the 'truth' of the part $Obligatory(act)$ can be established before the 'truth' of the whole is established. This at least differs from Tarski's compositional truth definition.

[28] Cf. Alchourrón & Martino [6] who provide a calculus with an 'introduction rule' for a prescriptively interpreted O-operator, where their rule corresponds to the modified Dubislav convention (DCM) plus a requirement of joint satisfiability.

[29] It seems consensus that there must be some 'test' of adequacy. Weinberger [111] writes that one must test a rule for the logical manipulation of norm sentences for its adequacy for the area of normative thought, and Sosa [97] speaks of a 'control of commonsense' that is necessary because otherwise there would simply be no end to the possible "logics".

ory is only an explanation of phenomena if its calculated observational sentences are regularly true, one should require of any proposed 'logic of imperatives' that the imperative it 'derives' from other sentences are normally – not 'true' of course, but accepted as 'implicit' in other sentences that are used as premisses. This resembles what is called the 'soundness' of a calculus: if the calculus allows 'false' (unacceptable) conclusions to be drawn from 'true' (accepted) premisses, then it must be discarded as 'unsound'.[30] I now turn to the question of adequacy in this sense.

6 Ross's Paradox and Weinberger's Variant

Shortly after Jørgensen's dilemma and Dubislav's workaround for a logic of imperatives had been described, Alf Ross re-considered inference schemes in 'the most simple form', where a 'new' imperative is inferred from one imperative premiss, i.e. where the scheme used is that of Dubislav's convention (DC). The following is an instance of such a scheme:

$$\begin{array}{l} !A \\ \therefore\ !(A \vee B) \end{array}$$

Here, $!A$ means (as now usual) an imperative sentence that is satisfied if and only if the descriptive sentence A is true, and $!(A \vee B)$ is an imperative that is satisfied if and only if either A or B are true. It is immediate that the second imperative can be inferred from the first sentence $!A$ by Dubislav's convention. Fine, said Ross, let $!A$ be interpreted as the imperative 'post the letter', so we can infer from the imperative 'post the letter' the imperative 'post the letter or burn it' $!(A \vee B)$. So

(1) Post the letter!
Therefore: Post the letter or burn it!

is a valid imperative inference. Ross himself points out that his paradox is not paradoxical if this 'validity' of an imperative inference is understood in the sense of a logic of satisfaction. If the letter is posted and the imperative $!A$ satisfied, then the imperative $!(A \vee B)$ will likewise be satisfied – this is no more paradoxical than that $A \vee B$ can be inferred from A. But if the meaning of 'imperative inference' refers to anything like the 'validity' or 'existence' of an imperative, then Ross claims that his inference is not only *not* immediately felt to be evident, but rather evidently false.

Why does Ross's example of an imperative inference seem paradoxical? In particular, regarding the 'formalistic theory' of imperative inference given in the

[30] The other possibility, that the calculus does not provide *all* the inferences from premisses that are acceptable (usually called 'completeness'), is less harmful and can be dealt with by e.g. refining it. For a similar definition of adequacy cf. Chellas [16] p. 4, where however the terminology is vice versa.

last section, why should it be paradoxical to say that if one uses the imperative $!A$ for commanding, then one 'implicitly' also commands $!(A \vee B)$? One explanation has been that that by using a disjunctive imperative, i.e. an imperative sentence that like $!(A \vee B)$ is satisfied if some state of affairs or some other state of affairs holds, the authority has left it to the subject how to satisfy her command. Suppose Romeo hands a letter to Mercutio with the words 'Post this letter or burn it, but relieve me from deciding its fate and mine', would his friend not be free to do as he pleases? Analyzing this 'freedom', it has been argued that giving a command entails an 'imperative permission' or implicitly authorizes to carry out the actions required to satisfy the command.[31] So the imperative 'post this letter or burn it' would contain the permission 'I hereby permit you to post the letter or burn it'. Now explicit disjunctive permissions are often understood in a 'strong' sense that grants both disjuncts: when someone says 'help yourself to a cup of coffee or a cup of tea', then the guest is permitted to help herself to coffee *and* also permitted to help herself to tea (though possibly not both). So one obtains the following chain:

(2) Post the letter!
Therefore: Post the letter or burn it!
Therefore: You may post the letter or burn the letter!
Therefore: You may burn the letter!

But it seems counterintuitive to say that by ordering a letter posted one permitted it to be burned.[32] To avoid this result, one may argue that it is not the first step in (2) that is problematic, but the second, i.e. we should not be allowed to infer a strong permission from an imperative. Yet nothing seems wrong with the following piece of Mercutio's reasoning:

(3) Romeo asked me to post the letter or burn it.
Therefore: I may post the letter or burn the letter, as I wish.
Therefore: I may burn the letter.

The reason why the inference from the first line of (3) to the second line seems not objectionable, while the similar inference from the second line in (2) to its third line appears somehow wrong, may lie in the fact that the imperative to 'post the letter or burn it' that is used in the reasoning is only implicit, i.e. derived, while Mercutio's reasoning was about an imperative that was explicitly used by Romeo. So one could modify one's view on the second step in (2) by saying that one is only allowed to infer a strong permission to do what is commanded if this command is not itself derived. I return to such a distinction between 'explicitly

[31] Cf. Chellas [16] p. 19 for the term 'imperative permission' and Keene [56] for the 'implicit authorization'.

[32] The idea to explain the counterintuitive nature of Ross's paradox using the also, or even more, counterintuitive inference to 'you may post the letter or burn it' was von Wright's in [130] pp. 21–22, also cf. von Wright [132] pp. 121–122 and Hintikka [43].

given' premisses and 'implicitly given' imperatives in a moment. But consider the example from the last section, where an officer was commanded to go via Coldstream or Berwick, and (a little later) told not to go via Coldstream, where both commands were viewed as implying the command to go via Berwick. The proposed modification would still allow us to make the following inference:

(4) I was commanded to go via Coldstream or Berwick.
Therefore: I may go via Coldstream or Berwick, as I wish.
Therefore: I may go via Coldstream.

So it seems the authority contradicted herself when ordering (a little later) *not* to go via Coldstream, i.e. first a choice between the two routes was granted, and later this choice was retracted, or rather: the second command modified the original command.[33] Whenever a command contradicts, cancels or modifies another command, the conflict may be absorbed e.g. by application of the rule of *lex posteriori*, which says that as a rule authorities should be considered competent to modify their own orders. But the puzzling thing is that the example was originally presented as a smooth application of imperative logic as facilitated by Dubislav's convention (DCM). Nothing made it appear as if there is some contradiction or modification involved and that more is used or required than just a flat application of the rules.[34]

An answer to these problems could be to give up the idea of strong imperative permission altogether: without it, the agent cannot reason that burning the letter is permitted. But while strong permission might be seen as problematic,[35] it is not clear why it should be altogether discarded. In particular, nothing seemed wrong with assuming strong permission in the case of Romeo's request (3). But whatever view is taken on strong permission, there is another point that makes Ross's paradox seem counterintuitive without appealing to some 'implied' permission: Imagine that, having been given the command 'Go via Coldstream or Berwick', the agent finds the road Coldstream blocked. Then the following reasoning seems logical:

(5) I was commanded to go via Coldstream or Berwick.
I cannot go via Coldstream.
Therefore: I should go via Berwick.

[33] According to Hare [38] it is just a conversational implicature that gets canceled. But it seems that by saying "go via Berwick or Coldstream" the authority *really* leaves it to the agent which route she wants to take – and later retracts this choice –, while someone who says e.g. "the tickets are upstairs or in the car", and later adds "they are not in the car" only made it *seem* as if the tickets could be in either location. If the order was only given "further orders pending', as Hare also argues, then the first order was not complete, because it left the agent unable to determine how to fulfill it. It is as if the authority had said in the middle of a sentence: "hang on, I'm not finished yet, I'll be right back."

[34] This was the point in Williams's [126] criticism of Hare's [38] scheme.

[35] Cf. Stenius [100]: "Free choice permission is too strong a concept to be useful."

It seems the kind of deliberation that one would expect of reasonable agents. Likewise, Mercutio, having been asked by Romeo to 'post the letter or burn it', might be found reasoning in the following way:

(6) Romeo asked me to post the letter or burn it.
For fear of Tybalt's revenge, I cannot bring myself to post the letter.
Therefore: I should burn the letter.

One might dispute whether Mercutio's fear is really on a par with a road blocked e.g. by a landslide. But if we suppose it is, then Mercutio's reasoning seems as impeccable as that of the officer. Now return to Ross's paradox: here the agent was ordered to 'post the letter'. Implicit in this imperative, so we are told by Dubislav's convention (DC), is the imperative 'post the letter or burn it'. Imagine that the agent is not able to post the letter for some cause (the postal workers are on strike and the mail bins have been locked up). So the agent could reason in the following way:

(7) I was (implicitly) ordered to post the letter or burn it.
I cannot post the letter.
Therefore: I should burn the letter.

But this reasoning is absurd. Just because the agent cannot fulfill her obligation to post the letter, this does not mean that she is obliged to do something that was never mentioned, and in fact could be anything: the words 'burn the letter' could be replaced e.g. by 'go to the zoo', 'kill a passer-by' or 'love your neighbor' and the inference would be just as valid – if it is valid.[36]

Now the agent, in reasoning in the above settings, used indicative statements about natural facts – like that something cannot be done – to reason about the imperatives 'go via Berwick or via Coldstream', 'post the letter' or 'post the letter or burn it'. But inferences that mix imperatives and indicatives are notoriously troublesome and should perhaps be avoided. As MacKay [66] points out, both of the following 'inferences'

Go fly a kite!
You are going to drop dead.
Therefore: Drop dead!

You are going to fly a kite.
Drop dead!
Therefore: Go fly a kite!

are validated by Dubislav's extended convention (DEC), where both inferences seem plainly invalid, and so perhaps (DEC) should not be accepted. Yet consider again the case of the officer. Imagine that it was not the original authority that

[36] This is Weinberger's [112], [113] explanation of why Ross's paradox poses a problem.

issued the command not to go via Coldstream, but someone else, like the officer's husband (who in the past had some bad experience on this road). Since there is some discretion in the authority's order, there is no reason why the officer should not give in to her husband's request, and so the following reasoning of the officer seems correct:

(8) I was commanded to go via Coldstream or Berwick.
My husband asked me not to go via Coldstream.
Therefore: I should go via Berwick.

Similarly, we can imagine Mercutio to reason in the following fashion:

(9) Romeo asked me to post the letter or burn it.
Tybalt threatened me not to post any of Romeo's letters.
Therefore: I should burn the letter.

But then the following reasoning of the agent to whom Ross's imperative 'post the letter' was addressed must be likewise correct:

(10) I was (implicitly) ordered to post the letter or burn it.
I have been asked not to post the letter.
Therefore: I should burn the letter.

There is some discretion in the (implicit) order to 'post the letter or burn it', so why should the agent not take an additional request into account? The only difference between the (9) and (10) is that in (9) the reasoning appeals to an explicitly used imperative, whereas in (10) it starts by considering an order that was only 'implicit' in the use of some imperative. So maybe what was wrong was that derived imperatives were used, without paying enough attention to the fact that the derived imperatives are only 'part of a system', that the 'explicitly' used imperatives have not ceased to exist, that imperatives that are only derived do not 'exist' on quite the same level as explicit imperatives, or that the agent is somehow expected to make use of the logically strongest information that is available.[37] So we are back at the proposal that a difference must be made between 'explicitly used' imperatives, and imperatives that 'only derive' from explicit imperatives. But to require that reasoning with imperatives starts with

[37] Rödig ([91] p. 184–185) points out that by deriving the norm to 'post the letter or burn it', the original order to 'post the letter' does not 'cease to exist', and that it is the conjunction of both norms that must be satisfied. That the entailed norms do not 'exist' in quite the same way as explicit norms is the idea of von Wright e.g. in [131] and [132] p. 114 and p. 122. According to Stenius [100], the use of 'post the letter or burn it' carries the tacit information that a stronger regulation like 'post the letter' does not 'belong to the codex'. For the idea that using a weaker sentence 'post the letter or burn it' violates a conversational presupposition cf. Hintikka [43]. Also cf. Hamblin [29] p. 88: " 'implicit imperatives' may be different from the real thing, and we should be wary of loading them up with the full range of imperative properties."

'explicit' imperatives, and must not start with imperatives that are only inferred, reveals an unusual, non-classical meaning of 'imperative inference'. For classically, logical inferences may very well be conducted by proving first that some assumptions have some desired conclusion, and then show that the assumptions follow from an accepted set of premisses. This is facilitated by the transitivity of classical consequence: if $A \in Cn(B)$ and $B \in Cn(C)$ then $A \in Cn(C)$ ('consequences of the consequences are also consequences'), or the monotonicity rule: if $A \in Cn(X)$ then $A \in Cn(X \cup Y)$ (what follows from some axioms also follows from a larger set of axioms).

Ross's paradox seems to demonstrate that given the imperative inferences provided e.g. by Dubislav's convention, it becomes necessary to distinguish between the imperatives that are explicitly given and the imperatives that are inferred: agents can use the former for their reasoning, but not always the latter, or not the latter by themselves, which makes reasoning with imperatives somehow non-classical. And so there may yet be another way to get around the difficulties: perhaps Ross's example is not really a case of an imperative inference. Perhaps (1) is simply invalid. It is obvious that the scheme is an application of Dubislav's convention, so (DC) must be modified. One way to do that is to let the logic that is used for the right hand side inference in figure 2.1 not classical logic (propositional or predicate logic), but some other logic that does not allow one to infer $A \vee B$ from A. Let e.g. Dubislav's convention be reinterpreted in terms of *relevant* implication, i.e. the scheme used is now the one in the next figure. There are several ways to define relevant implication, and I will not go into them. But standardly we cannot derive $A \vee B$ from A, and so Ross's paradox is solved.[38]

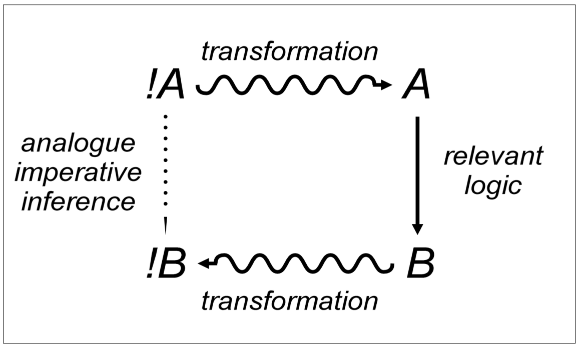


Fig. 2. Dubislav's convention with relevance.

[38] Weingartner & Schurz [125] and Weingartner [124] tailored their 'R-consequence' explicitly to eliminate Ross's paradox (for deontic logic, not imperative logic). It is stronger than other relevance logics in that not only the inference of $A \vee B$ from A is blocked, but also of $A \vee B$ from $(A \vee B) \wedge A$ or $(A \vee B) \wedge \neg B$. The price is

However, there is a variant of Ross's Paradox, also presented by Ross [92], that remains valid even on such a 'relevant' reinterpretation of (DC): the inference from $!(A \wedge B)$ to $!A$, where $A \wedge B$ means the sentence that is the conjunction of A and B. Consider the following arguments, they constitute 'Weinberger's Paradox' or the 'Paradox of the Window':[39]

(11) Close the window and play the piano!
Therefore: Close the window!

(12) Close the window and play the piano!
Therefore: Play the piano!

Suppose that α wants x to practice the piano, but neighbors have already complained about the disturbance and even called the police on a previous occasion. So α does not want x to play the piano while the window is open. Closing the window will reduce the noise so much that the neighbors are left with nothing to complain about. Suppose then that α sends x to play the piano, using the words 'close the window and play the piano'. A little bit later, the following discussion ensues between α and β:

α: I told him to play the piano, but I didn't hear him doing it all afternoon.
β: Well, at least he closed the window.
α: Why should he do that?

Here, the positive view on x's behavior by β is not accepted by α. Closing the window by itself is meaningless. It might even be unwanted in general – it blocks out fresh air – if it weren't for the sake of piano practice. But backed with the inference (11), β can continue in the following way:

β: You ordered him to close the window, that's what he did, so he did something right, didn't he?

Now consider the following, alternative dialogue:

α: He practised the Khachaturian with the window wide open. What shall we tell the police this time?
β: It was you that told him to play the piano.
α: But I didn't. He was also to close the window.

β's reproach for x's playing the piano is not accepted by α, because while playing the piano, x did not as he had been requested. However, backed with the inference (12), β could reply in this way:

a strange consequence relation: not only monotonicity fails, but also reflexivity, i.e. $X \subseteq Cn(X)$ is not valid.

[39] The origin of the example is unclear. The name 'Paradox of the Window' is used e.g. by Stranzinger [101] and Weinberger [119].

β: You ordered him to play the piano, that's what he did, so don't try to wiggle out of your responsibilities.

In these dialogues, α's position seems natural, while β's reaction is strange and uncomprehensible. But given the inference schemes (11) and (12), β is right: from α's command, the imperatives 'close the window' and 'play the piano' can be inferred – so we are told by Dubislav's convention (DC), and Dubislav's convention restricted by relevant implication. Moreover, these derived imperatives are used by β as imperatives are meant to be used, namely compared with reality, and reality accordingly qualified as 'right' or 'wrong'. So x did something right by closing the window, as x satisfied an (implicit) imperative, and similarly when playing the piano without closing the window. But intuitively, closing the window by itself produces nothing good, and playing the piano with the window open seems a clear violation of obligations and not satisfactory in any way.[40]

The 'window paradox' seems to arise whenever the states of affairs mentioned in the imperative are only conjunctively desired by an authority. That for this reason we cannot detach conjuncts in wishes, i.e. we cannot conclude from 'she wishes for a and b' that 'she wishes for a' was pointed out by Menger [72] for the case of complementary goods, e.g. when a is 'a cigarette' and b is 'a match', for one may not wish either one of the goods by itself. Ross [92] points out that the same difficulty arises for imperatives, e.g. when the imperative is to 'write a letter and post it'. Other examples have included the imperatives 'take the parachute and jump', 'pay the bill and file it' or 'fill up the boiler with water and heat it'.[41] Goble [27] showed that even a seemingly innocuous obligation to 'sing and dance at Gene's party' may be planted in a setting that makes it impossible to speak of fulfilling the obligation when only one act, singing or dancing, is performed. To determine whether an imperative is 'separable' or 'inseparable', i.e. whether doing A alone produces something 'right' with respect to an imperative $!(A \wedge B)$ or not, it is necessary to examine the intentions and wishes of the authority that used the imperative, it is not a matter of logic.[42]

To solve these difficulties, Kenny [58] proposes a logic of 'satisfactoriness'. This logic uses a set of propositions to represent the wishes of the authority under considerations. A *fiat* (an impersonal imperative like 'let there be light') is called satisfactory if and only if whenever the *fiat* is satisfied then every proposition

[40] It gives the paradox a further twist if we imagine that playing the piano with the window open is explicitly forbidden. For by Dubislav's convention (DC), the imperative $!\neg(A \wedge \neg B)$ ('don't play the piano while the window is open'), is derivable from the imperative $!(A \wedge B)$ ('close the window and play the piano'). But it seems that the additional prohibition is best formalized as a conditional imperative (in Hofstadter & McKinsey's [45] formalism: $\neg B \Rightarrow !\neg A$). Conditional imperatives pose other problems outside the current topic. In any case, one would still have to say that playing the piano with the window not closed was satisfactory with regard to some (derived) imperative.

[41] Cf. Hare [38], Weinberger [110] and [123]. These difficulties led Weinberger to reject the validity of an inference from $!(A \wedge B)$ to $!A$ in his publications since [110].

[42] The terminology here is that of Hamblin [29] p. 184.

in the set of wishes true. Finally an inference of one *fiat* from another *fiat* is defined as follows:

$!B$ may be inferred from $!A$ in the logic of satisfactoriness
if and only if
if $!A$ is satisfactory then $!B$ is satisfactory.

It is clear that the troublesome inferences (1), (11) and (12) are invalidated by this logic of satisfactoriness: when posting the letter is satisfactory for the wishes of the authority, then burning the letter need not be so. Likewise, if closing the window and playing the piano is satisfactory with respect to all wishes, then playing the piano alone does not guarantee that the wishes of the authority are also satisfied. But Kenny's approach gives rise to other paradoxes: in the logic of satisfactoriness we can e.g. derive:[43]

(13) Open the door!
Therefore: Open the door and wear a tie today!

The inference is clearly absurd and so Kenny's logic does not help us to solve the paradoxes.

In Ross's paradox, the imperative to 'post the letter or burn it' was 'inferred' from the imperative to 'post the letter', thus forcing one to acknowledge that some (though only inferred) imperative is satisfied by burning the letter. In the 'window paradox' we could 'infer' the imperative 'play the piano' from the imperative 'close the window and play the piano', thus forcing us to acknowledge that an (inferred) imperative is satisfied when the piano is played with the window wide open. In both cases, we would much rather say that *no* imperative was satisfied by burning the letter that was meant to be posted, and by playing the piano with the window open when it should have been closed. This, I think, is the main cause why Ross's paradox and the window paradox give rise to counterintuitive feelings, or are 'paradoxical'. So we should not be allowed to infer such imperatives. So Dubislav's convention is not an apt theory to explain how an imperative may be derived from another one. And so we are back at square one: all theories, including the formalistic approach, have so far failed to explain what it means to infer an imperative from some other imperative in spite of Jørgensen's Dilemma.

7 Ordinary Language Arguments

Maybe it is not really the case that all options have run out to redefine Dubislav's scheme in a way so that it avoids Ross's paradox and Weinberger's variant of it. Maybe we have to replace the classical logic that appears in his scheme by yet another logic, or develop such a logic.[44] But it is hard to see what kind

[43] A similar counterexample was given by Gombay [28], also cf. Sosa [97].

[44] Cf. Keene [56]: "What we wanted here is a logic of *actions*, in which a well-defined concept of *inclusion* plays a leading role."

of logic this could be, since most logics, including other non-monotonic logics, will permit us to either infer $!(A \lor B)$ from $!A$ or $!A$ from $!(A \land B)$, and so at least one of the two paradoxes will arise. So I think, after all these troublesome attempts to define a 'logic of imperatives', it is worthwhile to take another look at Poincaré's proposal that originally started the controversy.

Poincaré's only explicit example of an inference with an imperative conclusion has the following form:

(1) Do this!
This cannot be done without that.
Therefore: Do that!

The following is an instance of this scheme:

(2) Drive me to the airport!
To get to the airport, one must drive in a northerly direction.
Therefore: Drive me in a northerly direction!

In which setting could these sentences be used? Suppose I have entered a taxi and used the above sentences. But some confusion could arise. The driver could reply: "So what do you want me to do, drive you to the airport or just drive north?" The driver needs a direction. Ordering her to go to the airport alone is sufficient for this, and the behavior expected of a passenger entering a taxi. Using two imperatives where each contains an instruction of where to go is unexpected and confusing.

So suppose I have just used the sentence 'drive me to the airport'. A little later I realize that we seem not to be going north, and I say to my partner:

> "Is she hijacking us? I ordered her to go to the airport, and the airport lies to the north. So she ought to be driving us in a northerly direction. But she is not."

This reasoning seems flawless. Yet it does not involve sentences in the imperative mood and so cannot be an example of an imperative inference. But maybe this would be a good time to say to the driver:

(3) I ordered you to go to the airport.
To get to the airport, one must drive in a northerly direction.
Therefore: Drive me in a northerly direction!

But here the two sentences that function as premisses are both descriptive. Since Poincaré explained that an imperative cannot be derived from indicative premisses alone (and there is no reason not to follow him), this cannot be an imperative inference, and there must be something more involved than the drawing of a logical conclusion. One such other function of the 'therefore' appearing at the front of the last sentence of (3) is not to reason, but to motivate, as in:

(4) The car is broken.
Therefore: Take the bus into town!

Here the speaker is motivating the imperative to take the bus by explaining that driving into town is impossible, since the car is broken. So similarly, what seems to have happened in (3) is that I motivate my (new) imperative 'drive me in a northerly direction' by an already given command and an assumed fact.

Consider again the proposed inference (2). Just like indicative inferences are explained by the fact that someone who accepts (or: assents to) the premisses must also accept the conclusion, Hare [37] has argued that an imperative inference is one where someone who assents to all imperative premisses must also assent to the imperative conclusion:

> "A sentence *p* entails a sentence *q* if and only if the fact that a person assents to *p* but dissents from *q* is a sufficient criterion for saying that he has misunderstood one or other of the sentences. (...) A person who assented to this command ['Take all the boxes to the station'], and also to the statement 'This is one of the boxes' and yet refused to assent to the command 'Take this to the station' could only do so if he had misunderstood one of these three sentences."

But what does it mean that a person 'assents' to a command? Suppose John's mother tells him 'John, clear the table and do the washing up', and John's little brother echoes: 'John, do the washing up'. If John 'assents' to his mother's order, does he also have to 'assent' to an order by his brother, whom he might not accept as an authority? Perhaps the analysis assumes identity in the person who uses the commands. Suppose then it was not John's mother but some officer who used the imperative, and John is not obliged as son, but as this officer's orderly. The second command is also used by the officer, maybe a little later. But suppose that John is only obliged to the officer if the commanding is done in a certain fashion, e.g. when the officer is standing up, or when the officer is not drunk, and that when the second imperative was used the officer was, as a matter of fact, not standing up or already had more than her fill. Or suppose that John is not an orderly, but some djinni, and the officer is the person who rubbed the lamp, but that, when she used the first imperative, this already was the last of the three wishes that she had been granted. Does John, in these cases, have to 'assent' to the second command? It seems that such an interpretation of 'assent' would have to get involved into reasoning about whether the act of using an imperative 'really creates' a command. But it did not seem as if such reasoning is involved in Hare's proposal.

So let then the word 'assent' be understood in its weakest possible interpretation. A person could hardly be said to assent to a command given to her if she did not satisfy or to try to satisfy it. Returning to the situation where I have asked the taxi driver to take me to the airport, when the taxi driver assents to this request, she will start driving me to where she thinks the airport lies, i.e. start to satisfy, or try to satisfy, my request. If the taxi driver agrees that the airport lies in a northerly direction, she will, in obeying my request, eventually drive in what she thinks is a northerly direction. So she might be said to additionally satisfy, or try to satisfy, a request to drive me in a northerly direction, had such a request been made. But did I request the taxi driver to drive me north? I might be absolutely sure that the airport is to the north, but I would still blame the taxi driver for not going where I requested if, opposite to what I believed, the airport is in fact to the south-west of my starting point and the driver still went north. So I did not utter such a request, and would not even imply such a request, lest I be charged by the driver for going there instead of the airport. All we can say is that the taxi driver would also be satisfying, and so seemingly assenting to, a purely hypothetical request to drive me in a northerly direction, if she satisfies the request to drive me to the airport and the airport does in fact lie in a northerly direction. But this is again not a logic that infers one imperative from some set of other imperatives and/or indicatives, but the logic of satisfaction as explained in sec. 4.1.[45]

It would be nice to have some 'real life' examples, cases of 'ordinary' reasoning with imperative premisses and an imperative conclusion, i.e. instances of

(ImpInf) $!A$
$\therefore\ !B$

where $!A$ and $!B$ are sentences in the imperative mood, and where the use of the inference – not the imperatives – is either accepted in some ordinary discourse, or opposed (and the person who uses it blamed for being 'unreasonable' or 'illogical').

Use of indicative arguments in everyday discourse often occurs in singular sentences, like

(5) Unemployment is rising, so there are not enough jobs created.
(6) She has got an 'A' in English, so she achieved top-marks in at least one subject area.
(7) I have read all of Vladimir Nabokov's novels, so I have read *Pnin*.

Here two descriptive sentences are linked with the adverb 'so' (similar adverbs would be 'therefore' or 'hence'). (5) seems analytical if one understands 'enough' to be elliptical for 'enough to make up job-losses elsewhere'. (6) is analytical if one knows that 'A' is a top-grade and that English is one of several high-school subjects. (7) is made into a logical argument by the assumed background knowledge that *Pnin* is a novel by Nabokov. It is often not easy to distinguish such indicative arguments from sentences that present reasons, motives or are otherwise explanatory, for these also use the form of descriptive sentences that are concatenated by an adverb like 'so' or 'therefore', as in the following examples:

(8) I couldn't get the car started, therefore I took the bus.
(9) I wanted to make friends with her, therefore I asked her if she would go shopping with me.

[45] This resembles the criticism by Keene [56] of Hare's proposal.

(10) There were holes in the roof, so birds had come in and were roosting in the rafters.

(8) explains why today the speaker used the bus. Since the bus need not have been the only means to get into town, or the speaker may have stayed at home, the hearer cannot just conclude the second part from the first. (9) presents the psychological motive why the speaker asked the third person to go shopping with her. Other people might have been motivated differently by the desire to make friends with the third person. In (10), a natural event is explained by a certain state of affairs. Again, this is not a logical argument: the birds could also have not flown in, or flown in but not nested in the ceiling. Now the adverbs 'so' and'therefore' can also be used to meaningfully link imperatives. Consider the following examples:

(11) Stop the rise of unemployment, so see to it that more jobs are created!
(12) Make your guests comfortable, so introduce your guests to each other!
(13) Don't let vermin into you house, therefore patch up the roof!
(14) Read all of Nabokov's novels, so read *Pnin*!

(11) might be encountered in some political debate. At first it appears to be a good argument, but then doubts arise: is the speaker really appealing to logic, or is she just complementing her first imperative by a second, more specific one, as when we say: "Go there! Go there now!"? And one could also stop the rise of unemployment by e.g. prohibiting companies to dismiss their workers, or making it more difficult for them (maybe (5) was not so analytical after all). Then (11) would seem to be rather a case of a motivating use of 'so': the imperative to see to it that more jobs are created is motivated by a primary order to stop unemployment. Likewise, in (12), the advice to introduce guests to each other is rationalized by the more general aim to make guests comfortable. It is hard to see what could be analytical here: to ease tensions, the host may equally encourage the guests to *guess* each others names, or serve them plenty of alcohol, or maybe the guests are easygoing and do not really require any effort on the host's part to make themselves at home. Similarly, in (13) the more readily accepted advice to keep vermin out of the house is used as a rationale to make the addressee accept the drudgery of having to patch up the roof. The most promising candidate for an appeal to analyticity seems to be (14), i.e. that the imperative to read all of Nabokov's novels includes the imperative to read *Pnin*, given the background knowledge that *Pnin* is a novel by Nabokov. Note that when making the background knowledge explicit, it becomes a case of Dubislav's extended convention (DEC). Such a sentence may be used e.g. by a teacher of a literature course when addressing her students. But again we cannot rule out that this is just a case of complementing an imperative by a second, more specific one, as we sometimes do to get things done.

Adherents of Dubislav's convention (DC) must also accept the following argument:

(15) Aim for an ‘A’ in English, so aim for top-marks in at least one subject area!

But it seems dubious what reason the speaker could have for adding the ‘so’ part. Just aiming for top-marks in some subject area is clearly not what the speaker wants the addressee to do. More meaningful would be the converse,

(15a) Aim for top marks in at least one subject area, so aim for an ‘A’ in English!

where the advice to aim for ‘A’ in English is rationalized by the wish to have the student achieve top-marks somewhere. But since the student could not know from the first imperative that it was the subject of English that the speaker wanted her to achieve top marks in, this would – like (12) and (13) – rather be a ‘motivating so’, and not a use of ‘so’ that appeals to a logical capability.

Matters are further complicated by the fact that expressions of the following kind can also be meaningfully employed:

(16) The car isn’t working properly, so take the bus!
(17) I forgot my keys, therefore leave your key under the mat!
(18) Gill is your best friend, so invite her to your party!

In all three sentences, the first part is descriptive and the second is in the imperative mood. We have already noted in the case of (3) that such arguments exist, but for anyone who agrees to Poincaré’s thesis that imperative conclusions do not follow from an indicative premisses it is clear that (16) – (18) cannot represent valid arguments. (16) seems again a case where the ‘so’ is used to motivate the advice that is expressed by the imperative. The ‘so’ does not express a logical relation, for sometimes it is better to use a car that stutters than a coach that won’t take one back. In (17) the indicative gives a reason why the speaker wants her request to be followed. According to Hamblin [29], such reason-providing indicatives are often attached to advice-expressing imperatives, yet here the imperative might also be an order (e.g. of a parent). For the same reason the speaker might have ordered the agent to hand over her key, and not to leave it under the mat, and so what is expressed is again not a logical relation. (18) seems also like presenting a motive for inviting Gill to the party (she is the addressee’s best friend), but here things might be a bit more complicated – the expression could be elliptical for:

(18.a) Invite your best friends to the party, Gill is your best friend, so invite her to your party!

This is very similar to what Dubislav considered a valid argument, namely his inference from ‘thou shalt not kill’ to ‘Cain shall not kill Abel’. But then, (18) might also be elliptical for

(18.b) Gill is your best friend, one invites one's best friends to one's parties, so invite her to your party!

where the second part (which is not in imperative mood) appeals to the existence of a rule that the speaker might consider binding, or binding for the addressee. So this is rather a case of reason-giving, and not of a logical inference: the speaker motivates her imperative by asking the speaker to conform to some preexisting rule. And so it appears possible that also in (18.a) the first imperative served only as a rationale for the second imperative.

To tell the uses of 'therefore's' and 'so's' that are motivating, reason-giving or explanatory in a non-logical sense, apart from those that separate the premisses from the conclusion in an argument that is intended to be a logical one, we can use the following trick: instead of 'therefore' or 'so', use a clause like "... It follows logically from this that ..." to separate the sentences. The new phrase makes the appeal to a logical capability explicit. Where the original adverbs 'so' and 'therefore' were used to indicate a (claimed) logical inference, the new formulations

(5.a) Unemployment rates are rising. It follows logically from this that not enough jobs are created.

(6.a) She has got an 'A' in English. It follows logically from this that she achieved top-marks in at least one subject area.

(7.a) I have read all of Vladimir Nabokov's novels. It follows logically from this that yes, I have read *Pnin.*

appear only to be changes in expression. The speaker, just as before, appeals to a shared understanding of words, concepts and background knowledge, to make the second sentence seem to be expressing nothing new, but only a logical consequence from what has been said before. Note that it does not matter whether the arguments are, in fact, analytical. People sometimes think they use valid arguments when they are not. But the rephrased sentences make it clear that the speaker *intends* the sentences to be just that. And the new formulations seem not to change the meaning of the original sentences whenever a 'logical' use of the adverbs 'so' and 'therefore' was really intended. By contrast, when the first part was used to give some background information, a reason, explanation or motive, the rephrased expressions appear odd:

(8.a) I couldn't get the car started. It follows logically from this that I took the bus.

(9.a) I wanted to make friends with her. It follows logically from this that I asked her if she would go shopping with me.

(10.a) There were holes in the roof. It follows logically from this that birds had come in and were roosting in the rafters.

The phrase 'it follows logically from this' makes again an appeal to some shared understanding of used words, concepts and background. But here, this background knowledged obviously does not allow one to 'conclude' the second sentence from the first. The listener could not have known from the first sentences in these examples that the speaker took the bus, asked someone to go out shopping or has birds nesting in the roof of her house. So claiming, as the rephrased sentences do, that the second part can be concluded from the first, makes the sentences seem irritating, weird and false, while the earlier sentences appeared quite harmless.

Now consider what happens if such a method is used on imperative. So far, (14) seemed the best candidate for a sentence that 'appeals to logic', so I will concentrate on this example. First note that

(14.a) Read all of Nabokov's novels. It follows logically from this that read *Pnin*!

is not grammatical, so instead of the 'that' e.g. a colon (corresponding to a pause in oral language) must be used, as in the following expression:

(14.b) Read all of Nabokov's novels. It follows logically from this: read *Pnin*!

But here, the part that follows the colon seems strangely detached. Is this a command, i.e. is the speaker, using the expression following the colon, still giving a command? Or is the emphasis on the part before the colon, and so the purpose of the second sentence is merely to tell (truly or falsely) that some consequence relation holds? The impression that this is a strange use of words increases if we add the subject of the request:

(14.c) John, read all of Nabokov's novels. It follows logically from this: John, read *Pnin*!

Here, the phrase 'it follows logically from this' makes it appear as if the speaker was not giving commands to John at all. It seems what the speaker really does is talking about logical relations between sentences – maybe it is a logician presenting an example of an imperative inference. So perhaps we should try out another phrase:

(14.d) John, read all of Nabokov's novels. We can conclude from this: John, read *Pnin*!

Yet this expression also has a false ring: who is doing the commanding of the 'conclusion' – the speaker? Or the 'we' that is to do the concluding? Do the speaker and the listeners all join into giving John the command? Apparently it was wrong to use the first person plural, and so we might want to change the sentence into:

(14.e) John, read all of Nabokov's novels. I conclude from this: John, read *Pnin*!

But this seems to be the worst alternative so far. Is the speaker concluding the last sentence? Or is the speaker commanding it? And if so, then why is the speaker saying that she is concluding it? The performative acts of concluding and commanding seem to collide, whereas the acts of stating and concluding seemed to go hand in hand. But we have yet another phrase to try out:

(14.f) Read all of Nabokov's novels. So you can conclude for yourself: read *Pnin*!

Though this is perhaps a less common phrase to signal logical arguments, the new sentence seems to be the most successful so far. But it appears necessary that the 'you' is the person to whom both commands are addressed. So let us make the addressees explicit. Of the following sentences

(14.g) John, read all of Nabokov's novels. So John, you can conclude for yourself: read *Pnin*!

(14.g) John, read all of Nabokov's novels. So Mary, you can conclude for yourself: read *Pnin*!

(14.i) John, read all of Nabokov's novels. So Mary, you can conclude for yourself: John, read *Pnin*!

only the first seems somehow acceptable. In (14.g) it appears as if Mary is asked to read the book, but this can hardly be 'concluded' from a command not directed at Mary. (14.i) makes it seem as if Mary is asked to give a command to John (and not just to draw a conclusion). Moreover, if the addressee is expressly included in the inferred command, then also 14.f), which seemed so promising at first, looks strange:

(14.j) John, read all of Nabokov's novels. So you can conclude for yourself: John, read *Pnin*!

It seems that in (14.f) and (14.g) the speaker has not just asked the addressee of the first command to 'draw a conclusion', but in this process to 'give himself' the command expressed by the second sentence, i.e. to 'tell himself to read *Pnin*'. When the addressee is made explicit in the 'inferred' command, it looks as if the addressee is additionally asked to use his own first name when telling himself to read *Pnin* – which is a weird thing to ask of anybody. And this points at another problem of (14.f) and (14.g): if the person who commands 'read all of Nabokov's novels' (the teacher) and the person who commands 'read *Pnin*' (John himself) are not identical, how can the second imperative be inferred from the first?

By contrast, all of the above phrases can be employed for 'deontic sentences' (non-imperative sentences that do not prescribe, but describe what ought to be done) without difficulty:

(19) You ought to read all of Nabokov's novels, therefore you ought to read *Pnin*.

(19.a) John ought to read all of Nabokov's novels, therefore John ought to read *Pnin.*
(19.b) John ought to read all of Nabokov's novels. It follows logically from this that John ought to read *Pnin.*
(19.c) John ought to read all of Nabokov's novels. We can conclude from this that John ought to read *Pnin.*
(19.d) John ought to read all of Nabokov's novels. I conclude from this that John ought to read *Pnin.*
(19.e) John ought to read all of Nabokov's novels. You can conclude for yourself that John ought to read *Pnin.*
(19.f) John ought to read all of Nabokov's novels. Mary, you can conclude for yourself that John ought to read *Pnin.*

All these sentences seem grammatical, meaningful and not confusing. We might even view the inferences they express as sound, but this is not the question here. Yet as we have seen, all attempts to use the phrases that link these sentences, normally used to indicate logical arguments in indicative discourses, to link imperatives to indicate 'imperative inferences', result in expressions that seem somehow confused and wrong. When used to link imperatives, they mix up the roles of commanding, command-receiving, and drawing conclusions. And since the method to use such clauses to distinguish appeals to logic from e.g. motivating uses of 'therefore's and 'so's, fails to produce sentences that do not appear strange or confused in the case of imperatives, perhaps it did so because these adverbs really are not used to indicate a claimed analyticity when linking imperatives. A motivating use of the adverb 'so' suffices to explain why the sentence (14) seemed meaningful: the teacher, perhaps asked by John whether he also has to read *Pnin*, motivates the more specific imperative to read this book by prefixing to it the general requirement to read all of Nabokov's novels, thus making it clear that *Pnin* is in fact one of the books that John has to read. (14.f) appears comparatively less strange than the other reformulations because to ask John to 'give himself' the imperative to read *Pnin* may be a (roundabout) way to make sure he actually reads it. To understand (18) we do not need to determine whether the speaker refers to an explicit command to 'invite one's friends', or a social custom to do so, because what is in any case implicit in (18) is an appeal to a preexisting obligation to motivate the agent to do what the speaker wants her to do. It also explains why (15) seemed so strangely pointless: the reason for using the less specific imperative to achieve *some* top marks is not sufficiently explained by prefixing to it a more specific imperative to achieve top marks in English.[46] And so it seems that all of the imperative arguments (11)–(18) are really cases of reason-giving and motivation, and the 'so's and 'therefore's' used in these expressions that like Poincaré's '*donc*', or the '*also*'s,

[46] Note that the same strangeness does not necessarily arise for deontic logic. The dean of one faculty may say to another: "Our students are obliged to have an 'A' in English, so yes, ours are – like yours – obliged to achieve top marks in at least one subject are."

‘*daher*’s and ‘*deshalb*’s of German language, may be used to connect both indicative and imperative sentences, provide only reasons, explanations or motives in the case of imperatives, and do not indicate claims of analyticity.

So I want to dare the hypothesis that there are no examples of imperative inferences, i.e. logical conclusions in the imperative mood, drawn from at least one premiss in the imperative mood, to be found in ordinary language arguments. They only appear in the writings of some philosophers.

8 The Way to Go Forward

If there are, as a matter of fact, in ordinary language, no argument forms that resemble ‘imperative inferences’, then there also is no place for a formal theory for such a logic. Presenting formalizations of such a logic would be writing about what Dubislav [17] called an *Unding* or *chimaera*: a non-thing that exists only as a concept, but no real object falls under the concept.

So did Poincaré commit a mistake? Did he confuse an important insight by Hume [47] on the use of ‘is’ and ‘ought’ – that facts cannot be used to argue that they must be so or that other facts should be made similar to them – with a statement about grammar? Curiously, in his essay, Poincaré [87] never claimed to have discovered the logic of imperatives of which he was heralded as the pioneer. His main argument is that findings of science can influence moral reasoning. He just seems to presume that, like scientific arguments consist of sentences in the indicative mood, moral reasoning is conducted using sentences in the imperative mood. It is true that facts can influence the reasoning of agents about their obligations: Hare’s officer, who upon being commanded to go to Edinburgh via Coldstream or Berwick finds the road via Coldstream blocked, acts quite reasonably by concluding that she now ought to go via Berwick. But this is a reasoning *about* what obligations she has, it is a deontic argument, and not a case of ‘inferring’ imperatives. So Poincaré’s main argument is correct, but the assumed parallelism between sentences in the indicatives and the imperative mood, that they can both feature in logical arguments, does not exist. Our language does not work that way.

There are several ways to go forward from a position of ‘imperativological skepticism’.[47] First, one might continue the ‘logic of imperatives’ as a logic of satisfaction. The logic of satisfaction states which imperatives must also be satisfied if some other imperatives are satisfied, and it may also be used to state

[47] A number of authors have already denied the possibility of a logic of imperatives or norms, like von Wright [132] p. 109: “And now I too, after a long and winding itinerary have come to the same view: logical relations, e.g. of contradiction and entailment, cannot exist between (genuine) norms.”. Above we have already noted that Hamblin [29] p.89, Sellars [93] p. 239-240, von Kutschera [60] and Philipp [83], [84] have expressed scepticism or denied the possibility of a logic of imperatives altogether. For imperatives also cf. Moritz [74], Williams [126], Keene [56], Opałek & Woleński [79]. The term is coined from Weinberger’s [117] term ‘normological skepticism’ which denies logical relations not only between imperatives, but any prescriptive language. The main proponent of normological scepticism is Kelsen [57].

which imperatives will be violated by satisfying other imperatives. We can use the notion of satisfaction to distinguish imperatives that might be seen as redundant in a set of imperatives in the sense that these will also be satisfied if some other, different imperatives are satisfied, or identify subsets of imperatives that cannot be all satisfied and so conflict. By providing these concepts, the logic of satisfaction, though it may appear trivial, remains a meaningful and correct way to talk about imperatives.[48]

Second, imperatives normally express the wish or desire on the part of the person or authority using the imperative that what is commanded is satisfied. But it seems unreasonable to wish for A to be realized, but also for $\neg A$ to be realized, and in this sense two wishes may exclude another. If imperatives express wishes of one particular person, we can then point out to her what wishes may be unreasonable. Likewise it might be desirable to view the norms of a particular society as if they all were the wishes of one person, the 'law giver', and logic may then give advice as to which norms must be revised so that the system is 'reasonable'. This is the position of G. H. von Wright in his late work on normative, and deontic logic, cf. e.g. [131], [132].[49]

Finally, there is deontic logic. In the course of the paper I have portrayed deontic logic as a logic about what is obligatory given a set of imperatives or norms, as opposed to a logic of norms or imperatives. This is indeed the way deontic logic has been explained by von Wright in [128] and represents the main view.[50] But other authors have refused to make a commitment to such a descriptive interpretation of deontic logic. Some have explicitly viewed it as a logic of prescriptions. Castañeda [15] viewed deontic logic as a "modal logic of imperatives and resolutives". Åqvist [8] argued that it should be possible to interpret deontic logic "atheoretically" as a logic of commands, in the sense that OA expresses a command, and not a proposition about a command; occasional oddities like the difficulty of interpreting formulas like $\neg OA$ should be accepted as a small price for a logical theory of commands. Alchourrón [1] identified deontic logic with a "logic of norms", and set out to develop a logic of normative propositions in parallel. Chellas [16] replaced the O-operators of deontic logic by the symbol '!', where $!A$ is to be understood as representing a natural language imperative 'let it be the case that A', and presents a logic for such formulas that is equivalent to standard deontic logic *SDL*. For Bailhache [9], deontic logic is the logic of 'deontic norms' (norms that are created by the use of deontic expressions), where the deontic formulas are then evaluated with respect to accessible ideal worlds as usual. Sometimes it is not even quite clear in which way standard accounts

[48] Cf. C. G. Hempel's [40] remark with regard to Ross's Paradox that a logic of satisfaction should not be so easily rejected.

[49] The idea that commands can be identified with wishes, which in the above sense relate to each other, goes back to Bentham [11] pp. 95–97.

[50] Cf. von Wright [128]: "The system of Deontic Logic, which we are outlining in this paper, studies propositions (and truth-functions of propositions), about the obligatory, permitted, forbidden." Also cf. Føllesdal & Hilpinen [20]: "Deontic sentences (represented by the formulae of deontic logic) describe what is regarded as permitted, obligatory, forbidden etc., in some unspecified normative system."

of deontic logic desire its formulas to be read: Carmo & Jones [14] write that deontic logic is a formal tool needed to 'design normative systems' and, just like Bailhache in [10], throughout call the sentences of deontic logic 'norms' or 'deontic norms' – but the authors then employ possible worlds semantics to evaluate deontic propositions (norms?) as true or false as usual. Føllesdal & Hilpinen [20] speak of deontic propositions as constituting or being 'implied' by a normative system (e.g. pp. 13, 29), or of deontic logic formalizing imperatives (p. 26), and treat the descriptive interpretation of deontic logic only as one that they "shall often resort to" (p. 8).

However, it is quite clear that deontic logic cannot be such a logic, that OA cannot be interpreted as representing imperatives or norms. I have explained in sec. 3 why imperatives, and for that reason, also 'deontic norms', cannot be meaningfully termed 'true' and 'false'. But if norms are neither true nor false, then the Boolean operators occurring in the formulas of deontic logic such as '$OA \wedge OB$', '$PA \vee \neg OA$', '$OA \rightarrow PB$', cannot have their usual, truth functional meanings as 'and', 'or', 'not', 'if ..., then'. So a logic of norms, even if it is – contrary to my view – possible, cannot resemble deontic logic,[51] and the meaning of (sub-)formulas such as 'OA' must be interpreted descriptively. So the descriptive interpretation of the formulas of deontic logic is the only tenable one.

What the confusion about the meaning of deontic logic illustrates is the need to better explain how deontic logic relates to (explicitly given) imperatives and norms. Already Ziemba [134] stated that "the lack of distinction between commands (norms) and deontic propositions, or propositions on commands, is a source of various evil in deontic logic," and David Makinson [67], a quarter of a century later, echoed this statement when he noted at the workshop ΔEON '98 on Deontic Logic in Computer Science in Bologna 1998, that work on deontic logic has been going on as if a distinction between norms (that cannot be called true or false) and normative propositions (that can) has never been heard of. Makinson called for a new start, a reconstruction of deontic logic as a logic concerned with norms, but in accord with the philosophical position that norms are devoid of truth values. Following Makinson's call and an existent 'imperatival tradition' of deontic logic I have, in subsequent papers [30], [31], [32], [33], [34], shown how deontic logic can be reinterpreted as a logic *about* imperatives. The imperatives according to which statements like 'it ought to be that A' are true are explicitly represented in the semantics of deontic logic. We can then define deontic operators which make some the following statements true:

(1) $\{!A\} \models OA$
(2) $\{!A, !B\} \models O(A \wedge B)$
(3) $\{!A\} \models O(A \vee B)$
(4) $\{!(A \wedge B)\} \models OA$
(5) $\{!(A \wedge B), !(\neg A \wedge C)\} \not\models OB$
(6) $\{!(A \wedge B), !(\neg A \wedge C)\} \models O(A \wedge B) \wedge O(\neg A \wedge C)$

[51] Cf. Makinson [67] p. 30. Keuth [59] and Swirydowicz [102] therefore restrict their 'logic of norms' to statements of 'normative entailment' of the form $!A \vdash !B$.

(7) $\{!(A \wedge B), !(\neg A \wedge C)\} \models O((A \wedge B) \vee (\neg A \wedge C))$
(8) $\{!(A \wedge B)\}, \not\models O(A \neg B)$

(1) says that if there is an imperative in the set of imperatives (that are taken as mandatory) according to which A ought to be done, then the statement OA ('it ought to be that A') is true. (2) allows us to agglomerate what several imperatives demand: if according to two imperatives, A and B ought to be done, then $O(A \wedge B)$ is true. (3) and (4) are true for a sense of obligation that lets it suffice for the statement OA to be true that according to an imperative in the sense that A must (also) be (done), in the sense that A is necessary to satisfy the imperative. Such operators give rise to deontic versions of Ross's or Weinberger's paradox, but interpreted in the above way these versions appear harmless: there is no imperative in the set of mandatory imperatives at the front of (3) that doing $A \vee B$ would satisfy, and similarly for (4) and doing A alone. If we want to model a 'strong' sense of obligation that does not make $O(A \vee B)$ or OA true in these cases, then we may do so by using an O-operator for which (3) and (4) are not true. (5) is true of a deontic operator that maintains that in a conflict between imperatives the subject can continue to reason about her obligations. A conflict does not make everything obligatory. According to some proposals, each of what the colliding imperatives demand should be described as obligatory (case 6), and according to others, only what they demand disjunctively is obligatory (case 8). (8) holds for a dyadic deontic operator that recognizes that some norms may no longer be satisfiable in the case the situation described by the antecedent C of the formula $O(A/C)$ is true, and does not make what they demand obligatory any longer.

As illustrated by such statements, the fact that there is no logic of imperatives does not mean that imperatives are not 'handled' in certain ways by agents when reasoning about their obligations: the logic of satisfaction and Weinberger's principle (W) make it possible to define from a set of imperatives e.g. what is necessary to satisfy some imperatives, all imperatives, or maximal subsets of non-conflicting imperatives. We can discount imperatives that are already satisfied or violated in a certain situation, compare different sets of imperatives and say e.g. that two sets make no difference regarding what obligations result, etc. Thus, though imperatives or norms have no truth values and no 'logic', they still can be meaningfully used to determine what obligations arise in a certain situation. This view lies at the bottom of my approach, and also that of Makinson's and van der Torre's 'input/output logic' (cf. [68], [69]).[52] In the mentioned papers I have outlined operators for which the above statements are true. Also, I have shown that there are quite natural ways to define notions of ought that make the standard systems of monadic and dyadic deontic logic still

[52] Makinson & van der Torre explain that such a handling can also be used to devise a machine that 'hands out new norms' in a particular situation. But, as they also explain, this should not viewed as a kind of logical procedure. If it works (e.g. in the case of a computer automating the issuing of tax bills or parking tickets) it does so because the machine designers (authorities) *want* the machine to work in that way.

sound and complete with respect to such semantics. But there are also other definitions, like operators that handle conflicts in the ways described by (5) and (6) or (7). This multitude of possible definitions does not, in my view, mean that deontic logic has become a 'logic *à la carte*'. It rather responds to the different circumstances in which deliberation is required of an agent, and to different ways we talk about obligations, e.g. when we treat them as being *prima facie* or 'all things considered', or when we can or cannot stipulate that the norm-giver has been rational. Thinking about what sets of imperatives or other norms make which deontic propositions about the imperatives or norms true seems to me a promising way to go forward for deontic logic.

Deontic logic has been disparagingly called a "kind of ersatz truth", that merely mirrors logical relations that already exist between imperatives or norms, and so we should rather look for this logic than studying a deontic logic that only reflects it and so must result in a "dull isomorphism".[53] But it has been the 'logic of imperatives' that has kept escaping us, while sentences that use deontic expressions can easily be used to form valid arguments. So maybe it is not a logic of imperatives that is the 'proper' subject of study and makes deontic logic just an ersatz theory, but it is the other way round,[54] and the idea of a logic of imperatives has been a *fata morgana*, leading us to ever more futile attempts to explain inference relations between imperatives, to find analogues of truth values, or new logics to explain Dubislav's scheme, whereas any plausibility of this idea was just a reflection of the real, but quite distinct possibility of a logic about imperatives, namely of deontic logic. If my hypothesis holds, then it is the only logic regarding normative concepts such as obligation that we should be concerned with.

[53] This is Hare's view in [38] p. 325; also cf. Alchourrón [1] pp. 264–266; Kalinowski [50] p. 134; Weinberger [117] p. 58, [120]; Wagner & Haag [106] p. 102. The idea that deontic logic reflects the logical properties of norms is that of von Wright in [129] p. 134.

[54] Cf. Alchourrón & Bulygin [4] p. 463: "This logic of norms is, so to say, a reflection of the logic of normative propositions. It is because we regard as inconsistent a system in which it is true that $O_x p$ and $O_x \neg p$, that we say that the norms $!p$ and $!\neg p$ are incompatible. So it is the logic of norm propositions which yields the foundations for the logic of norms." .

References

1. Alchourrón, C. E., “Logic of Norms and Logic of Normative Propositions”, *Logique & Analyse*, **12**, 1969, 242–268.
2. Alchourrón, C. E., “The Intuitive Background of Normative Legal Discourse and its Formalization”, *Journal of Philosophicl Logic*, **1**, 1972, 447–463.
3. Alchourrón, C. E. and Bulygin, E., “The Expressive Conception of Norms”, in [42] 95–124.
4. Alchourrón, C. E. and Bulygin, E., “Pragmatic Foundations for a Logic of Norms”, *Rechtstheorie*, **15**, 1984, 453–464.
5. Alchourrón, C. E. and Bulygin, E., “On the Logic of Normative Systems”, in: Stachowiak, H. (ed.), *Handbuch pragmatischen Denkens*, Hamburg: Meiner, 1993, 273–293.
6. Alchourrón, C. E. and Martino, A. A., “Logic Without Truth”, *Ratio Juris*, **3**, 1990, 46–67.
7. Anscombe, G. E. M., *Intention*, Oxford: Blackwell, 1957.
8. Åqvist, L., “Interpretations of Deontic Logic”, *Mind*, **73**, 1964, 246–253.
9. Bailhache, P., “Several Possible Systems of Deontic Weak and Strong Norms”, *Logique & Analyse*, **21**, 1980, 89–100.
10. Bailhache, P., *Essai de logique déontique*, Paris: Librairie Philosophique J. Vrin, 1991.
11. Bentham, J., *Of Laws in General*, London: Athlone Press, 1970, edited by H. L. A. Hart. First appeared 1782.
12. Boole, G., *The Mathematical Analysis of Logic*, Cambridge: Macmillan, Barclay, and Machmillan, 1847.
13. Bulygin, E., “Existence of Norms”, in: Meggle, G. (ed.), *Actions, Norms, Values. Discussions with Georg Henrik von Wright. Papers presented at a colloquium held April 1996 at the Center for Interdisciplinary Research, University of Bielefeld*, Berlin: de Gruyter, 1999, 237–244.
14. Carmo, J. and Jones, A. J. I., “Deontic Logic and Contrary-to-duties”, in [23] 265-343.
15. Castañeda, H.-N., “Obligation and Modal Logic”, *Logique & Analyse*, **3**, 1960, 40–49.
16. Chellas, B. F., *The Logical Form of Imperatives*, Stanford: Perry Lane Press, 1969.
17. Dubislav, W., *Die Definition*, 3rd ed., Leipzig: Meiner, 1931.
18. Dubislav, W., “Zur Unbegründbarkeit der Forderungssätze”, *Theoria*, **3**, 1938, 330–342.
19. Engliš, K., “Die Norm ist kein Urteil”, *Archiv für Rechts- und Sozialphilosophie*, **50**, 1964, 305–316.
20. Føllesdal, D. and Hilpinen, R., “Deontic Logic: An Introduction”, in [41] 1–35.
21. Foot, P., “Moral Realism and Moral Dilemma”, *Journal of Philosophy*, **80**, 1983, 379–389.
22. Frey, G., “Idee einer Wissenschaftslogik. Grundzüge einer Logik imperativer Sätze”, *Philosophia Naturalis*, **4**, 1957, 434–491.
23. Gabbay, D. M. and Guenthner, F. (eds.), *Handbook of Philosophical Logic*, vol. 8, 2nd ed., Dordrecht: Kluwer, 2002.
24. Geach, P. T., “Imperative and Deontic Logic”, *Analysis*, **18**, 1958, 49–56.
25. Geach, P. T., “Imperative Inference”, *Analysis*, **23**, 1963, 37–42.
26. Gentzen, G., “Untersuchungen über das logische Schließen”, *Mathematische Zeitschrift*, **39**, 1934, 176–210; 405–431.

27. Goble, L., "A Logic of Good, Should, and Would", *Journal of Philosophical Logic*, **19**, 1990, 169–199.
28. Gombay, A., "What is Imperative Inference?", *Analysis*, **27**, 1967, 145–152.
29. Hamblin, C. L., *Imperatives*, Oxford: Blackwell, 1987.
30. Hansen, J., "Sets, Sentences, and Some Logics about Imperatives", *Fundamenta Informaticae*, **48**, 2001, 205–226.
31. Hansen, J., "Problems and Results for Logics about Imperatives", *Journal of Applied Logic*, **2**, 2004, 39–61.
32. Hansen, J., "Conflicting Imperatives and Dyadic Deontic Logic", *Journal of Applied Logic*, **3**, 2005, 484–511.
33. Hansen, J., "Deontic Logics for Prioritized Imperatives", *Artificial Intelligence and Law*, **14**, 2006, 1–34.
34. Hansen, J., "Prioritized Conditional Imperatives: Problems and a New Proposal", *Autonomous Agents and Multi-Agent Systems*, 2007, forthcoming. Online available from <http://www.springerlink.com/content/p253015562l43212>.
35. Hanson, W. H., "A Logic of Commands", *Logique & Analyse*, **9**, 1966, 329–343.
36. Hare, R. M., "Imperative Sentences", *Mind*, **58**, 1949, 21–39.
37. Hare, R. M., *The Language of Morals*, Oxford: Oxford University Press, 1952.
38. Hare, R. M., "Some Alleged Differences Between Imperatives and Indicatives", *Mind*, **76**, 1967, 309–326.
39. Harris, Z. S., "Transformational Theory", *Language*, **41**, 1965, 363–401.
40. Hempel, C. G., "Review of Alf Ross: 'Imperatives and logic"', *Journal of Symbolic Logic*, **6**, 1941, 105–106.
41. Hilpinen, R. (ed.), *Deontic Logic: Introductory and Systematic Readings*, Dordrecht: Reidel, 1971.
42. Hilpinen, R. (ed.), *New Studies in Deontic Logic*, Dordrecht: Reidel, 1981.
43. Hintikka, J., "The Ross Paradox as Evidence for the Reality of Semantical Games", in: Saarinen, E. (ed.), *Game-Theoretical Semantics*, Dordrecht: Reidel, 1977, 329–345.
44. Hodges, W., *Logic*, Harmondsworth: Penguin, 1977.
45. Hofstadter, A. and McKinsey, J. C. C., "On the Logic of Imperatives", *Philosophy of Science*, **6**, 1938, 446–457.
46. Holländer, P., *Rechtsnorm, Logik und Wahrheitswerte. Versuch einer kritischen Lösung des Jörgensenschen Dilemmas*, Baden-Baden: Nomos, 1993.
47. Hume, D., *A Treatise of Human Nature*, Oxford: Oxford University Press, 1888, edited by L. A. Selby-Bigge. First appeared 1739.
48. Jørgensen, J., "Imperatives and Logic", *Erkenntnis*, **7**, 1938, 288–296.
49. Kalinowski, G., *Le Problème de la Vérite en Morale et en Droit*, Lyon: Emmanuel Vitte, 1967.
50. Kalinowski, G., *Einführung in die Normenlogik*, Frankfurt/M.: Athenäum, 1973.
51. Kalinowski, G., "Über die deontischen Funktoren", in: Lenk, H. (ed.), *Normenlogik. Grundprobleme der deontischen Logik*, Pullach: Dokumentation, 1974, 39–63.
52. Kalinowski, G., "Über die Bedeutung der Deontik für Ethik und Rechtsphilosophie", in: Conte, A. G., Hilpinen, R. and von Wright, G. H. (eds.), *Deontische Logik und Semantik*, Wiesbaden: Athenaion, 1977, 101–129.
53. Kamp, H., "Free Choice Permission", *Proceedings of the Aristotelian Society*, **74**, 1973/74, 57–74.
54. Kamp, H., "Semantics versus Pragmatics", in: Guenthner, F. and Schmidt, S. J. (eds.), *Formal Semantics and Pragmatics for Natural Languages*, Dordrecht: Reidel, 1978, 255–287.

55. Kanger, S., “New Foundations for Ethical Theory: Part 1”, duplic., 42 p., 1957, reprinted in [41] 36–58.
56. Keene, G. B., “Can Commands Have Logical Consequences?”, *American Philosophical Quarterly*, **3**, 1966, 57–63.
57. Kelsen, H., *Allgemeine Theorie der Normen*, Wien: Manz, 1979.
58. Kenny, A. J., “Practical Inference”, *Analysis*, **26**, 1966, 65–75.
59. Keuth, H., “Deontische Logik und Logik der Normen”, in: Lenk, H. (ed.), *Normenlogik. Grundprobleme der deontischen Logik*, Pullach: Dokumentation, 1974, 64–86.
60. von Kutschera, F., *Einführung in die Logik der Normen, Werte und Entscheidungen*, Freiburg/München: Karl Alber, 1973.
61. Ledig, G., “Zur Klärung einiger Grundbegriffe. Imperativ, Rat, Bitte, Beschluß, Versprechen”, *Revue Internationale de la Théorie du Droit*, **3**, 1928/29, 260–270.
62. Ledig, G., “Zur Logik des Sollens”, *Der Gerichtssaal*, **100**, 1931, 368–385.
63. Lemmon, E. J., “Deontic Logic and the Logic of Imperatives”, *Logique & Analyse*, **8**, 1965, 39–71.
64. Lemmon, E. J., *Beginning Logic*, 2nd ed., London: Chapman and Hall, 1987.
65. Leonard, H. S., “Interrogatives, Imperatives, Truth, Falsity and Lies”, *Philosophy of Science*, **26**, 1959, 172–186.
66. MacKay, A. F., “Inferential Validity and Imperative Inference Rules”, *Analysis*, **29**, 1969, 145–156.
67. Makinson, D., “On a Fundamental Problem of Deontic Logic”, in: McNamara, P. and Prakken, H. (eds.), *Norms, Logics and Information Systems*, Amsterdam: IOS, 1999, 29–53.
68. Makinson, D. and van der Torre, L., “Input/Output Logics”, *Journal of Philosophical Logic*, **29**, 2000, 383–408.
69. Makinson, D. and van der Torre, L., “Constraints for Input/Output Logics”, *Journal of Philosophical Logic*, **30**, 2001, 155–185.
70. Mally, E., *Grundgesetze des Sollens. Elemente der Logik des Willens*, Graz: Leuschner & Lubensky, 1926.
71. Mates, B., *Elementary Logic*, 2nd ed., Oxford: Oxford University Press, 1972.
72. Menger, K., “A Logic of the Doubtful. On Optative and Imperative Logic”, *Reports of a Mathematical Colloquium*, **2**, 1939, 53–64, reprinted in [73] 91–102.
73. Menger, K., *Selected Papers in Logic and Foundations, Didactics, Economics*, Dordrecht: Reidel, 1979.
74. Moritz, M., “Der praktische Syllogismus und das juridische Denken”, *Theoria*, **20**, 1954, 78–127.
75. Moutafakis, N. J., *Imperatives and their Logic*, New Delhi: Sterling, 1975.
76. Niiniluoto, I., “Truth and Legal Norms”, *Archiv für Rechts- und Sozialphilosophie Beiheft*, **25**, 1985, 168–190.
77. Opałek, K., “On the Logical-Semantic Structure of Directives”, *Logique & Analyse*, **13**, 1970, 169–196.
78. Opałek, K., *Theorie der Direktiven und Normen*, Wien: Springer, 1986.
79. Opałek, K. and Woleñski, J., “Is, Ought, and Logic”, *Archiv für Rechts- und Sozialphilosophie*, **73**, 1987, 373–385.
80. Opałek, K. and Woleñski, J., “Normative Systems, Permission and Deontic Logic”, *Ratio Juris*, **4**, 1991, 334–348.
81. Peczenik, A., “Doctrinal Study of Law and Science”, *Österreichische Zeitschrift für öffentliches Recht*, **17**, 1967, 128–141.
82. Peczenik, A., “Norms and Reality”, *Theoria*, **34**, 1968, 117–133.

83. Philipp, P., "Logik deskriptiver normativer Begriffe", 1989/90, In [85] pp. 241–290. First published as reports of the Karl-Marx-Universität Leipzig, *Untersuchungen zur Logik und zur Methodologie* **6**:65–88, 1989, and **7**:51–82, 1990.
84. Philipp, P., "Normative Logic Without Norms", 1991, In [85] pp. 291–301. Edited version of a presentation to the Meeting of the Central Division of the American Philosophical Association, Chicago, 1991.
85. Philipp, P., *Logisch-philosophische Untersuchungen. Edited by Ingolf Max and Richard Raatzsch.*, Berlin: de Gruyter, 1998.
86. Philipps, L., "Braucht die Rechtswissenschaft eine deontische Logik?", in: G., J. and Maihofer, W. (eds.), *Rechtstheorie. Beiträge zur Grundlagendiskkussion*, Frankfurt: Klostermann, 1971, 352–368.
87. Poincaré, H., *Dernières Pensées*, Paris: Ernest Flammarion, 1913.
88. Rescher, N., *The Logic of Commands*, London: Routledge & Kegan Paul, 1966.
89. Rescher, N., "Assertion Logic", in: Rescher, N. (ed.), *Topics in Philosophical Logic*, chap. XIV, Dordrecht: Reidel, 1968.
90. Rödig, J., "Kritik des Normlogischen Schließens", *Theory and Decision*, **2**, 1971, 79–93.
91. Rödig, J., "Über die Notwendigkeit einer besonderen Logik der Normen", *Jahrbuch für Rechtssoziologie und Rechtstheorie*, **2**, 1972, 163–185.
92. Ross, A., "Imperatives and Logic", *Theoria*, **7**, 1941, 53–71, Reprinted with only editorial changes in *Philosophy of Science* **11**:30–46, 1944.
93. Sellars, W., "Imperatives, Intentions and the Logic of "Ought"", *Methodos*, **8**, 1956, 227–268.
94. Sigwart, C., *Logic*, vol. I. The Judgment, Concept, and Inference, 2nd ed., London: Swan Sonnenschein & Co., 1895.
95. Simon, H. A., "The Logic of Rational Decision", *The British Journal for the Philosophy of Science*, **16**, 1965, 169–186.
96. Sosa, E., "The Logic of Imperatives", *Theoria*, **32**, 1966, 224–235.
97. Sosa, E., "On Practical Inference and the Logic of Imperatives", *Theoria*, **32**, 1966, 211–223.
98. Sosa, E., "The Semantics of Imperatives", *American Philosophical Quarterly*, **4**, 1967, 57–64.
99. Stenius, E., "The Principles of a Logic of Normative Systems", *Acta Philosophica Fennica*, **16**, 1963, 247–260.
100. Stenius, E., "Ross' Paradox and Well-Formed Codices", *Theoria*, **48**, 1982, 49–77.
101. Stranzinger, R., "Ein paradoxienfreies deontisches System", in: Tammelo, I. and Schreiner, H. (eds.), *Strukturierungen und Entscheidungen im Rechtsdenken*, Wien: Springer, 1978, 183–193.
102. Świrydowicz, K., "Normative Consequence Relation and Consequence Operations on the Language of Dyadic Deontic Logic", *Theoria*, **60**, 1994, 27–47.
103. Tarski, A., "Über einige fundamentale Begriffe der Metamathematik", *Comptes Rendus des Séances de la Sociéte des Sciences et des Lettres de Varsovie*, **23**, 1930, 22–29, published in English under the title "On Some Fundamental Concepts of Metamathematics" in [105] pp. 30–37.
104. Tarski, A., "Der Wahrheitsbegriff in den formalisierten Sprachen", *Studia Philosophica*, **I**, 1935, 261–405.
105. Tarski, A., *Logic, Semantics, Metamathematics*, Oxford: Oxford University Press, 1956.
106. Wagner, H. and Haag, K., *Die moderne Logik in der Rechtswissenschaft*, Bad Homburg: Gehlen, 1970.

107. Walter, R., "Jörgensen's Dilemma and How to Face It", *Ratio Juris*, **9**, 1996, 168–171.
108. Walter, R., "Some Thoughts on Peczenik's Replies to 'Jörgensen's Dilemma and How to Face It' (with Two Letters by A. Peczenik)", *Ratio Juris*, **10**, 1997, 392–396.
109. Weinberger, O., "Über die Negation von Sollsätzen", *Theoria*, **23**, 1957, 102–132.
110. Weinberger, O., *Die Sollsatzproblematik in der modernen Logik: Können Sollsätze (Imperative) als wahr bezeichnet werden?*, Prague: Nakladatelství Ceskoslovenské Akademie Ved, 1958, page numbers refer to the reprint in [114] pp. 59–186.
111. Weinberger, O., "Bemerkungen zur Grundlegung der Theorie des juristischen Denkens", *Jahrbuch für Rechtssoziologie und Rechtstheorie*, **2**, 1972, 134–161.
112. Weinberger, O., "Der Begriff der Nicht-Erfüllung und die Normenlogik", *Ratio*, **14**, 1972, 15–32.
113. Weinberger, O., "Ideen zur logischen Normensemantik", in: Haller, R. (ed.), *Jenseits von Sein und Nichtsein. Beiträge zur Meinong-Forschung*, Graz: Akademische Druck- und Verlagsanstalt, 1974, 295–311, Reprinted in [114] pp. 259–277.
114. Weinberger, O., *Studien zur Normlogik und Rechtsinformatik*, Berlin: Schweitzer, 1974.
115. Weinberger, O., "Ex falso quodlibet in der deskriptiven und präskriptiven Sprache", *Rechtstheorie*, **6**, 1975, 17–32.
116. Weinberger, O., *Normentheorie als Grundlage der Jurisprudenz und Ethik*, Berlin: Duncker&Humblot, 1981.
117. Weinberger, O., "Der normenlogische Skeptizismus", *Rechtstheorie*, **17**, 1986, 13–81, reprinted in [121] pp. 431–499.
118. Weinberger, O., *Rechtslogik*, 2nd ed., Berlin: Duncker&Humblot, 1989.
119. Weinberger, O., "The Logic of Norms Founded on Descriptive Language", *Ratio Juris*, **4**, 1991, 284–307.
120. Weinberger, O., "The Logic of Norms Founded on Descriptive Language", *Ratio Juris*, **4**, 1991, 284–307.
121. Weinberger, O., *Moral und Vernunft*, Wien: Böhlau, 1992.
122. Weinberger, O., *Alternative Handlungstheorie*, Wien: Böhlau, 1996.
123. Weinberger, O., "Logical Analysis in the Realm of Law", in: Meggle, G. (ed.), *Actions, Norms, Values. Discussions with Georg Henrik von Wright. Papers presented at a colloquium held April 1996 at the Center for Interdisciplinary Research, University of Bielefeld*, Berlin: de Gruyter, 1999, 291–304.
124. Weingartner, P., "Reasons from Science for Limiting Classical Logic", in: Weingartner, P. (ed.), *Springer*, Berlin, 2003, 233–248.
125. Weingartner, P. and Schurz, G., "Paradoxes Solved by Simple Relevance Criteria", *Logique & Analyse*, **29**, 1986, 3–40.
126. Williams, B. A. O., "Imperative Inference", *Analysis*, **23**, 1963, 30–36.
127. Wittgenstein, L., *Philosophical Investigations*, Oxford: Blackwell, 1953.
128. von Wright, G. H., "Deontic Logic", *Mind*, **60**, 1951, 1–15.
129. von Wright, G. H., *Norm and Action*, London: Routledge & Kegan Paul, 1963.
130. von Wright, G. H., *An Essay in Deontic Logic and the General Theory of Action*, Amsterdam: North Holland, 1968.
131. von Wright, G. H., "Is there a Logic of Norms?", *Ratio Juris*, **4**, 1991, 265–283.
132. von Wright, G. H., "A Pilgrim's Progress", in: von Wright, G. H. (ed.), *The Tree of Knowledge and Other Essays*, Leiden: Brill, 1993, 103–113.
133. Yoshino, H., "Über die Notwendigkeit einer besonderen Normenlogik als Methode der juristischen Logik", in: Klug, U. (ed.), *Gesetzgebungstheorie, Juristische Logik, Zivil- und Prozessrecht. Gedächtnisschrift für Jürgen Rödig*, Berlin: Springer, 1978, 141–161.

134. Ziemba, Z., “Deontic Syllogistics”, *Studia Logica*, **28**, 1971, 139–159.
135. Ziemba, Z., “Deontic Logic”, 1976, Appendix in [136], pp. 360–430.
136. Ziembiński, Z., *Practical Logic*, Dordrecht: Reidel, 1976.

What is Input/Output Logic? Input/Output Logic, Constraints, Permissions[⋆]

David Makinson[1] and Leendert van der Torre[2]

[1] david.makinson@googlemail.com
[2] University of Luxembourg, Computer Science and Communications (CSC)
1359, Luxembourg, 6 rue Richard Coudenhove Kalergi, Luxembourg
leendert@vandertorre.com

Abstract. We explain the *raison d'être* and basic ideas of input/output logic, sketching the central elements with pointers to other publications for detailed developments. The motivation comes from the logic of norms. Unconstrained input/output operations are straightforward to define, with relatively simple behaviour, but ignore the subtleties of contrary-to-duty norms. To deal with these more sensitively, we constrain input/output operations by means of consistency conditions, expressed via the concept of an outfamily. They also provide a convenient platform for distinguishing and analysing several different kinds of permission.

Keywords. Deontic logic, input/output logic, constraints, permissions

1 Motivation

Input/output logic takes its origin in the study of conditional norms. These may express desired features of a situation, obligations under some legal, moral or practical code, goals, contingency plans, advice, etc. Typically they may be expressed in terms like: *In such-and-such a situation, so-and-so should be the case*, or *. . . should be brought about*, or *. . . should be worked towards*, or *. . . should be followed* – these locutions corresponding roughly to the kinds of norm mentioned.

To be more accurate, input/output logic has its source in a tension between the philosophy of norms and formal work of deontic logicians.

Philosophically, it is widely accepted that a distinction may be drawn between norms on the one hand, and declarative statements on the other. Declarative statements may bear truth-values, in other words are capable of being true or false; but norms are items of another kind. They may be respected (or not), and may also be assessed from the standpoint of other norms, for example when a legal norm is judged from a moral point of view (or vice versa). But it makes no sense to describe norms as true or as false.

However the formal work of deontic logicians often goes on as if such a distinction had never been heard of. The usual presentations of deontic logic, whether

[⋆] This paper extends [11] with Section 6 on permissions.

Dagstuhl Seminar Proceedings 07122
Normative Multi-agent Systems
http://drops.dagstuhl.de/opus/volltexte/2007/928

axiomatic or semantic, treat norms as if they could bear truth-values. In particular, the truth-functional connectives *and*, *or* and most spectacularly *not* are routinely applied to norms, forming compound norms out of elementary ones. Semantic constructions using possible worlds go further by offering rules to determine, in a model, the truth-value of a norm.

This anomaly was noticed more than half a century ago, by Dubislav [4] and Jørgensen [5], but little was done about it. Indeed, from the 1960s onwards, the semantic approach in terms of possible worlds deepened the gap. The first serious attempt by a logician to face the problem appears to be due to Stenius [15], followed by Alchourrón and Bulygin [2] for unconditional norms, then Alchourrón [1] and Makinson [7] for conditional ones. Input/output logic may be seen as an attempt to extract the essential mathematical structure behind these reconstructions of deontic logic.

Like every other approach to deontic logic, input/output logic must face the problem of accounting adequately for the behaviour of what are called 'contrary-to-duty' norms. The problem may be stated thus: given a set of norms to be applied, how should we determine which obligations are operative in a situation that already violates some among them? It appears that input/output logic provides a convenient platform for dealing with this problem by imposing consistency constraints on the generation of output.

We begin by outlining the central ideas and constructions of unconstrained input/output logic. These are quite straightforward, and provide the basic framework of the theory. We then sketch a strategy for constraining those operations so as to deal more sensitively with contrary-to-duty situations. Finally, we explain how the same operations may be deployed in the analysis of permission.

For further details, the reader is invited to refer to Makinson and van der Torre [8,9].

2 Unconstrained Input/Output Operations

We avoid assuming that conditional norms bear truth-values. They are not embedded in compound formulae using truth-functional connectives. To avoid all confusion, they are not even treated as formulae, but simply as ordered pairs (a, x) of purely boolean (or eventually first-order) formulae.

Technically, a normative code is seen as a set G of conditional norms, *i.e.*, a set of such ordered pairs (a, x). For each such pair, the body a is thought of as an *input*, representing some condition or situation, and the head x is thought of as an *output*, representing what the norm tells us to be desirable, obligatory or whatever in that situation. The task of logic is seen as a modest one. It is not to create or determine a distinguished set of norms, but rather to prepare information before it goes in as input to such a set G, to unpack output as it emerges and, if needed, coordinate the two in certain ways. A set G of conditional norms is thus seen as a transformation device, and the task of logic is to act as its 'secretarial assistant'.

The simplest kind of unconstrained input/output operation is depicted in Figure 1. A set A of propositions serves as explicit input, which is prepared by being expanded to its classical closure $Cn(A)$. This is then passed into the 'black box' or 'transformer' G, which delivers the corresponding immediate output

$$G(Cn(A)) = \{x \mid \text{ for some } a \in Cn(A), (a, x) \in G\}.$$

Finally, this is expanded by classical closure again into the full output $out_1(G, A) = Cn(G(Cn(A)))$. We call this *simple-minded output.*

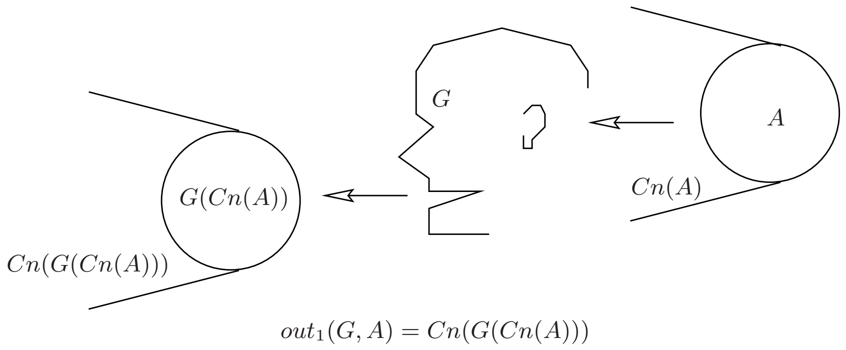


Fig. 1. Simple-minded Output

This is already an interesting operation. As desired, it does not satisfy the principle of identity, which in this context we call *throughput*, *i.e.*, in general we do not have $a \in out_1(G, \{a\})$ – which we write briefly, dropping the parentheses, as $out_1(G, a)$. It is characterized by three rules. Writing $x \in out_1(G, a)$ as $(a, x) \in out_1(G)$ and dropping the right hand side as G is held constant, these rules are:

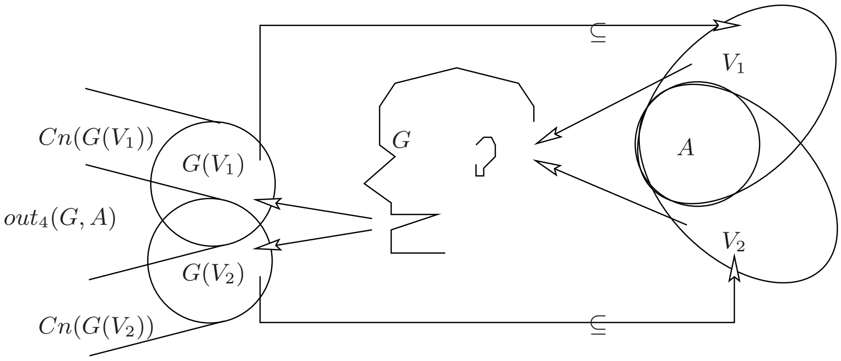
Strengthening Input (SI):	From (a, x) to (b, x) whenever $a \in Cn(b)$
Conjoining Output (AND):	From (a, x), (a, y) to $(a, x \wedge y)$
Weakening Output (WO):	From (a, x) to (a, y) whenever $y \in Cn(x)$

But simple-minded output lacks certain features that may be desirable in some contexts. In the first place, the preparation of inputs is not very sophisticated. Consider two inputs a and b. By classical logic, if $x \in Cn(a)$ and $x \in Cn(b)$ then $x \in Cn(a \vee b)$. But there is nothing to tell us that if $x \in out_1(G, a) = Cn(G(Cn(a)))$ and $x \in out_1(G, b) = Cn(G(Cn(b)))$ then $x \in out_1(G, a \vee b) = Cn(G(Cn(a \vee b)))$.

In the second place, even when we do not want inputs to be automatically carried through as outputs, we may still want outputs to be reusable as inputs – which is quite a different matter.

Operations satisfying each of these two features can be provided with explicit definitions, pictured by diagrams in the same spirit as that for simple-minded output, and characterized by straightforward rules. We thus have four very natural systems of input/output, which are labelled as follows: *simple-minded* alias out_1 (as above), *basic* (simple-minded plus input disjunction: out_2), *reusable* (simple-minded plus reusability: out_3), and *reusable basic* (all together: out_4).

For example, reusable basic output may be given a diagram and definition as in Figure 2. In the definition, a complete set is one that is either maximally consistent or equal to the set of all formulae.



$$out_4(G, A) = \cap\{Cn(G(V)) \mid A \subseteq V \supseteq G(V), V \text{ complete}\}$$

Fig. 2. Basic Reusable Output

The three stronger systems may also be characterized by adding one or both of the following rules to those for simple-minded output:

Disjoining input (OR): From (a, x), (b, x) to $(a \vee b, x)$
Cumulative transitivity (CT): From (a, x), $(a \wedge x, y)$ to (a, y)

These four operations have four counterparts that also allow *throughput*. Intuitively, this amounts to requiring $A \subseteq G(A)$. In terms of the definitions, it is to require that G is expanded to contain the diagonal, *i.e.*, all pairs (a, a). Diagrammatically it is to add arrows from G's ear to mouth. Derivationally, it is to allow arbitrary pairs of the form (a, a) to appear as leaves of a derivation; this is called the zero-premise identity rule ID.

All eight systems are distinct, with one exception: basic throughput, which we write as out_2^+, authorizes reusability, so that $out_2^+ = out_4^+$. This may be shown directly in terms of the definitions, or using the following simple derivation of CT from the other rules.

$$\dfrac{\dfrac{\dfrac{\dfrac{(a,x)}{(a\wedge\neg x,x)}\,\text{SI}\quad\dfrac{-}{(a\wedge\neg x,a\wedge\neg x)}\,\text{ID}}{(a\wedge\neg x,x\wedge(a\wedge\neg x))}\,\text{AND}}{(a\wedge\neg x,y)}\,\text{WO}\qquad (a\wedge x,y)}{(a,y)}\,\text{OR}$$

The application of WO here is justified by the fact that we have $y \in Cn(x \wedge (a \wedge \neg x))$ since the right hand formula is a contradiction. Note that all rules available in basic throughput (including, in particular, identity) are needed in the derivation, reflecting the fact that CT is not derivable in the weaker systems.

This strong system indeed collapses into classical consequence, in the sense that $out_4^+(G, A) = Cn(m(G) \cup A)$ where $m(G)$ is the materialization of G, *i.e.*, the set of all formulae $a \to x$ where $(a, x) \in G$.

The authors' papers [8] and [9, section 1] investigate these systems in detail – semantically, in terms of their explicit definitions, derivationally, in terms of the rules determining them, both separately and in relation to each other. We do not attempt to summarize the results here, but hope that the reader is tempted to follow further.

3 Why constrain?

As mentioned in section 1, all approaches to deontic logic must face the problem of dealing with contrary-to-duty norms. In general terms, we recall, the problem is: given a set of norms, how should we determine which obligations are operative in a situation that already violates some among them.

The following simple example is adapted from Prakken and Sergot [13].[1] Suppose we have the following two norms: *The cottage should not have a fence or a dog; if it has a dog it must have both a fence and a warning sign.*

In the usual deontic notation: $O(\neg(f \vee d)/t)$, $O(f \wedge w/d)$, where t stands for a tautology; in the notation of input/output logic: $(t, \neg(f \vee d))$, $(d, f \wedge w)$. Suppose further that we are in the situation that the cottage has a dog, thus violating the first norm. What are our current obligations?

Unrestricted input/output logic gives f: *the cottage has a fence* and w: *the cottage has a warning sign.* Less convincingly, because unhelpful if the presence of a dog is regarded as unalterable, it also gives $\neg d$: *the cottage does not have a*

[1] There are many examples in the literature. Most of them involve ingredients that, while perfectly natural in ordinary discourse, are extraneous to the essential problem and thus invite false analyses. These ingredients include defeasibility, causality, the passage of time, and the use of questionable rules such as CT and OR in deriving output. We have chosen a very simple example that avoids all those elements. There is one respect in which it could perhaps be further purified: under input d, the output is not only inconsistent with the input, but also itself inconsistent. This matter is discussed at the end of section 5.

dog. Even less convincingly, it gives $\neg f$: *the cottage does not have a fence*, which is the opposite of what we want.

These results hold even for simple-minded output, without reusability or disjunction of inputs. The only rules needed are SI and WO, as shown by the following derivation of $\neg f$.

$$\cfrac{\cfrac{(t,\neg(f\vee d))}{(t,\neg f)}\ \text{WO}}{(d,\neg f)}\ \text{SI}$$

A common reaction to examples such as these is to ask: why not just drop the rule SI of strengthening the input? In semantic terms, why not cut back the definition of simple-minded output from $Cn(G(Cn(A)))$ to $Cn(G(A))$, and in similar (but more complex) fashion with the others? Indeed, this is a possible option, and the strategy that we will describe below does have the effect of disallowing certain applications of SI. But simply to drop SI is, in the view of the authors, too heavy-handed. We need to know *why* SI is not always appropriate and, especially, *when* it remains justified.

4 A Strategy for Constraint: Maxfamilies and their Outfamilies

Our strategy is to adapt a technique that is well known in the logic of belief change – cut back the set of norms to just below the threshold of making the current situation contrary-to-duty. In effect, we carry out a contraction on the set G of given norms.

Specifically, we look at the maximal subsets $G' \subseteq G$ such that$out(G',A)$ is consistent with input A. In [8], the family of such G' is called the *maxfamily* of (G,A), and the family of outputs $out(G',A)$ for G' in the maxfamily, is called the *outfamily* of (G,A).[2]

To illustrate this, consider $G = \{(t,\neg(f\vee d)),(d,f\wedge w)\}$, with the contrary-to-duty input d. Using simple-minded output, $maxfamily(G,d)$ has just one element $\{(d,f\wedge w)\}$, and so $outfamily(G,d)$ has one element, namely $Cn(f\wedge w)$.

[2] So defined, the outfamily is not in general the same as the family of all maximal values of $out(G',A)$ consistent with A, for G' ranging over subsets of G. Every maximal value of $out(G',A)$ is in the outfamily, but not always conversely. For certain of our output operations, the two families do coincide, but not for others.

This can be shown by simple examples, such as the Möbius strip of Makinson [6,7]. Put $G = \{(a,x),(x,y),(y,\neg a)\}$. Then, for $out = out_3$ or $out = out_4$, $maxfamily(G,a)$ has three elements, namely the three two-element subsets of G. As a result, $outfamily(G,a)$ also has three elements – $Cn(\emptyset)$, $Cn(x)$, and $Cn(\{x,y\})$. Of these, only the last is a maximal value of $out(G',A)$ consistent with A for G' ranging over subsets of G.

We add that in this example, not even $Cn(\{x,y\})$ is a maximal subset of $out(G,a)$ that is consistent with a, for clearly $Cn(\{x,y\}) \subset Cn(\{x,y,\neg a\vee z\}) \subset out(G,a)$. Care is thus needed to avoid confusing maxfamilies with related maximal sets.

Although the outfamily strategy is designed to deal with contrary-to-duty norms, its application turns out to be closely related to belief revision and nonmonotonic reasoning when the underlying input/output operation authorizes throughput.

When all elements of G are of the form (t, x), then for the degenerate input/output operation $out_2^+(G,a) = out_4^+(G,a) = Cn(m(G) \cup \{a\})$, the elements of $outfamily(G,a)$ are just the maxichoice revisions of $m(G)$ by a, in the sense of Alchourrón, Gärdenfors and Makinson [3]. These coincide, in turn, with the extensions of the default system $(m(G), a, \emptyset)$ of Poole [12].

More surprisingly, there are close connections with the default logic of Reiter, falling a little short of identity. Read elements (a, x) of G as normal default rules $a; x/x$ in the sense of Reiter [14], and write $extfamily(G,A)$ for the set of extensions of (G, A). Then, for reusable simple-minded throughput out_3^+, it can be shown that $extfamily(G,A) \subseteq outfamily(G,A)$ and indeed that $extfamily(G,A)$ consists of precisely the maximal elements (under set inclusion) of $outfamily(G,A)$.

These results and related ones are proven in Makinson and van der Torre [9]. But in accord with the motivation from the logic of norms, the main focus in that paper is on input/output logics *without* throughput. Two kinds of question are investigated in detail there.

4.1 The search for truth-functional reductions of the consistency constraint

From the point of view of computation, it is convenient to make consistency checks as simple as possible, and executable using no more than already existing programs. For this reason, it is of interest to ask: under what conditions is the consistency of A with $out(G,A)$ reducible to the consistency of A with the materialization $m(G)$ of G, *i.e.*, with the set of all formulae $a \to x$ where $(a, x) \in G$?

It is easy to check that the latter consistency implies the former for all seven of our input/output operations. It turns out that we have equivalence for just two of them (reusable basic with and without identity).

On the level of derivations, the question can take a rather different form, with different answers. Given a derivation of (a, x) with leaves L, under what conditions is the consistency of a with $out(L,a)$ equivalent to its consistency with $m(L)$? Curiously, this holds for a wider selection of our input/output operations – in fact, for all of them except basic output. Even more surprisingly, for some of the operations (those without OR), the same reduction also holds with respect to the set $h(L)$ of heads x, and the set $f(L)$ of fulfilments $a \wedge x$, of elements (a, x) of L.

From this result on derivations, we can go back and sharpen the semantic one. When G is a *minimal* set with $x \in out(G,a)$ then, for each of our input/output operations other than basic output, a is consistent with $out(G,a)$ iff it is consistent with $m(G)$ – and for the operations without OR, with $h(G)$, $f(G)$.

4.2 More severe applications of the consistency check

From a practical point of view, whenever we constrain an operation to avoid excess production, the question arises: how cautious (timid) or brave (foolhardy) do we want to be? For input/output operations, this issue arises in different ways on the semantic and derivational levels. On the semantic level, once we have formed an outfamily we may ask: should we intersect, join, or choose from its elements to obtain a unique restrained output? On the level of derivations, it is natural to ask: do we want to apply the consistency check only at the root of a derivation, or at every step within it?

The policy of checking only at the root corresponds to the option, on the semantic level, of forming the join of the outfamily; while the stricter policy of checking at every step is an essentially derivational requirement. But whichever of the two we choose, it is of interest to know under what conditions they coincide. In other words, given a derivation of (a, x) with leaves L such that a is consistent with $out(L, a)$, under what conditions does it follow that for every node (b, y) in the derivation, b is consistent with $out(L, b)$? It turns out that for certain of the seven input/output operations (again, those without the OR rule) this result holds. For operations with OR but without the rule CT, a rather subtler result may be obtained.

One lesson of these rather intricate investigations is that the behaviour of the consistency constraint depends very much on the choice of input/output operation; in particular, the presence of the rule OR destroys some properties. Another lesson is that questions can take different forms, with different answers, on the semantic and derivational levels. Thirdly, a detour through derivations can sometimes sharpen semantic results.

5 Doubts and Queries

The investigation of constrained output is a much more complex matter than that of unconstrained output. It is also more open to doubts and queries. We put the main ones on the table.

5.1 Dependence on the formulation of G

The outfamily construction, at least in its present form, depends heavily on the formulation of the generating set G. To illustrate this, we go back to the cottage example of Prakken and Sergot [13] considered in sections 3 and 4. Here $G = \{(t, \neg(f \vee d)), (d, f \wedge w)\}$, and we consider the contrary-to-duty input d. As we have seen, using simple-minded output, $maxfamily(G, d)$ has unique element $\{(d, f \wedge w)\}$ and $outfamily(G, d)$ has unique element $Cn(f \wedge w)$. But if we split the first element of G into $(t, \neg f), (t, \neg d)$ then we get a different result. The maxfamily has two elements $\{(t, \neg f)\}$, $\{(d, f \wedge w)\}$ and the outfamily has two elements $Cn(\neg f)$ and $Cn(f \wedge w)$. Is this dependence on formulation of G a virtue, or a vice?

5.2 Are we cutting too deeply?

This problem is related to the first one. In some cases, the outfamily construction cuts deeply, perhaps too much. Consider again the cottage example, but this time with just one rule $(t, \neg(f \vee d))$ in G. Consider the same contrary-to-duty input d. Then the maxfamily has the empty set as its unique element, and so the outfamily has $Cn(\emptyset)$ as its unique element. Is this cutting too deeply? Shouldn't $Cn(\neg f)$ be retained?

5.3 Should we pre-process G?

If we wish to cut less deeply, then a possible procedure might be to 'pre-process' G. In the last example, when we decompose the sole element $(t, \neg(f \vee d))$ of G into $(t, \neg f)$, $(t, \neg d)$ then $Cn(\neg f)$ becomes the unique element of outfamily in the contrary-to-duty situation d. In general, for each element (a, x) of G, we could rewrite the head x in conjunctive normal form $x_1 \wedge \ldots \wedge x_n$, and then split (a, x) into $(a, x_1), \ldots, (a, x_n)$. This manoeuvre certainly meets the particular example. But is it appropriate for other examples of the same form with different content? And does it suffice for more complex examples? It looks suspiciously like hacking.

5.4 Avoid inconsistency with what?

On our definition, $maxfamily(G, A)$ is the family of maximal subsets $G' \subseteq G$ such that $out(G', A)$ is consistent with input A. It may be suggested that this is too radical – so long as $out(G, A)$ is consistent we should apply it without constraint.

To illustrate this, take another variation on the cottage example. Put $G = \{(t, \neg(f \vee d)), (d, w)\}$. The second norm no longer requires a fence when there is a dog, only a warning sign. Consider again the contrary-to-duty input d. Now $out(G, d) = Cn(\{(\neg f, \neg d, w\})$ which is inconsistent with the input d, but itself perfectly consistent. Should we cut it at all? Perhaps 'yes' if the input d is considered as unalterably true, but 'no' if it is presented as true but changeable.

6 Conditional Permission from an Input/output Perspective

In philosophical discussion of norms it is common to distinguish between two kinds of permission, negative and positive. Negative permission is easy to describe: something is permitted by a code iff it is not prohibited by that code, i.e. iff *nihil obstat*. In other words, taking prohibition in the usual way, something is negatively permitted by a code iff there is no obligation to the contrary.

Positive permission is more elusive. As a first approximation, one may say that something is positively permitted by a code iff the code explicitly presents it as such. But this leaves the central logical question unanswered. As well as the items that a code explicitly pronounces to be permitted, there are presumably others that in some sense follow from the explicit ones. The problem is to make it clear what kind of 'following' this is.

From the point of view of input/output logic, negative permission is straightforward to define: we simply put $(a, x) \in \mathit{negperm}(G)$ iff $(a, \neg x) \notin \mathit{out}(G)$, where *out* is any one of the four input/output operations that we have already discussed.

Because of its negative character, *negperm* fails the rule SI (strengthening the input). In other words, we don't have: $(a, x) \in \mathit{negperm}(G) \& a \in Cn(b) \Rightarrow (b, x) \in \mathit{negperm}(G)$. Indeed, it satisfies the opposite rule WI (weakening the input): $(a, x) \in \mathit{negperm}(G) \& b \in Cn(a) \Rightarrow (b, x) \in \mathit{negperm}(G)$. For if $(a, \neg x) \notin \mathit{out}(G)$ and $b \in Cn(a)$ then by SI for the underlying output operation, $(b, \neg x) \notin \mathit{out}(G)$ so $(b, x) \in \mathit{negperm}(G)$. This is a particular instance of a quite general pattern: whenever out satisfies a Horn rule (HR) then the corresponding *negperm* operation satisfies an 'inverse' Horn rule $(\text{HR})^{-1}$.

How should we define positive permission for conditional norms? Let G, P be sets of ordered pairs of propositions, where G represents the explicitly given conditional obligations of a code and P its explicitly given conditional permissions. The operation of *forward positive permission* is defined by putting:

$(a, x) \in \mathit{forperm}(P, G)$ iff $(a, x) \in \mathit{out}(G \cup Q)$ for some singleton or empty $Q \subseteq P$

i.e. in the principal case that P is not itself empty,

$(a, x) \in \mathit{forperm}(P, G)$ iff $(a, x) \in \mathit{out}(G(c, z))$

for some pair $(c, z) \in P$. This tells us that (a, x) is permitted whenever there is some explicitly given permission (c, z) such that when we treat it as if it were an obligation, joining it with G and applying the output operation to the union, then we get (a, x). Permissions are thus treated like weak obligations, the only difference being that while the latter may be used jointly, the former may only be applied one by one.

On the other hand, the operation of *backward positive permission* is defined by setting:

$(a, x) \in \mathit{backperm}(P, G)$ iff $(c, \neg z) \notin \mathit{out}(G \cup \{(a, x)\})$ for some pair $(c, z) \in P$ with c consistent.

This tells us that (a, x) is permitted whenever, given the obligations already present in G, we can't forbid x under the condition a without thereby committing ourselves to forbid something that has been explicitly permitted. With this in mind, one could also speak of the operation as one of *prohibition immunity.*

What do these two notions mean in ordinary life? Forward permission answers to the needs of the citizen, who needs to know whether an action that he is entertaining is permitted in the current situation. It also corresponds to the needs of authorities assessing the action once it is performed. If there is some explicit permission that 'covers' the action in question, then it is itself implicitly permitted.

On the other hand, backward permission fits the needs of the legislator, who needs to anticipate the effect of adding a prohibition to an existing corpus of norms. If prohibiting x in condition a would commit us to forbid something that has been explicitly permitted, then adding the prohibition is inadmissible under pain of incoherence, and the pair (a, x) is to that extent protected from prohibition.

Forperm and *backperm* are very different operations. Whereas *forperm* satisfies SI, *backperm* satisfies WI. Like negative permission, *backperm* satisfies the 'inverse' rule $(\mathrm{HR})^{-1}$ of any Horn rule (HR) satisfied by out; but *forperm* satisfies instead a 'subverse' rule $(\mathrm{HR})^{\downarrow}$.

Backperm may be characterized in a rather different way, using an idea of Makinson, [7]. Let us say that G is cross-coherent with P iff there is no $(c, z) \in P$ with c consistent, such that $(c, \neg z) \in \mathit{out}(G)$. Then it is easy to check that $(a, x) \in \mathit{backperm}(P, G)$ iff $(a, x) \in \mathit{negperm}(H)$ for every $H \supset G$ that is cross-coherent with P. From this it follows, in particular, that when G is cross-coherent with P then $\mathit{backperm}(P, G) \subseteq \mathit{negperm}(G)$. In this sense, we can say that under 'normal conditions' backward permission is a strengthened negative permission.

Further details of the behaviour of these operations may be found in Makinson and van der Torre [10].

7 Conclusions

Drawing together the threads of this paper, we emphasize the main points.

- Input/output logic seeks to extract the essential mathematical structure behind recent attempts to reconstruct deontic logic that avoid treating norms as if they had truth-values.
- Unconstrained input/output provides us with a simple and elegant construction with straightforward behaviour, but whose application to norms totally ignores the subtleties of contrary-to-duty obligations.
- On the other hand, output constrained using the outfamily strategy provides a way of dealing with contrary-to-duty obligations. Its behaviour is quite subtle, and depends considerably on the choice of background input/output operation, in particular on whether or not it authorizes the rule of disjunction of inputs.
- However, our definition of an outfamily has features that might be regarded as shortcomings. Its effect depends on the formulation of the generating set of norms; in some examples it gives what may be regarded as a wrong result unless some pre-processing as carried out on the generating set; and in some contexts the requirement of consistency of output with input may be too strong. These are delicate issues, and it remains possible that they have no unique solution definable in purely formal terms.

– Input/output operations also enable us to give a clear formal articulation of the well-known distinction between negative and positive permission. They also enable us, for the first time, to distinguish two very different kinds of positive permission, with quite different uses in practical life.

A topic of further research is the analysis of structured assemblies of input/output operations. Such structures, called logical input/output nets, or lions for short, are graphs, with the nodes labelled by pairs (G, out) where G is a normative code and out is an input/output operations (or recursively, by other lions). The relation of the graph indicates which nodes have access to others, providing passage for the transmission of local outputs as local inputs. The graph is further equipped with an entry point and an exit point, for global input and output.

References

1. Alchourrón, C., "Philosophical foundations of deontic logic and the logic of defeasible conditionals", in: Meyer, J. and Wieringa, R. (eds.), *Deontic Logic in Computer Science*, New York: Wiley, 1993, 43–84.
2. Alchourrón, C. and Bulygin, E., "The expressive conception of norms", in: Hilpinen, R. (ed.), *New Essays in Deontic Logic*, Dordrecht: Reidel, 1981, 95–124.
3. Alchourrón, C., Grdenfors, P. and Makinson, D., "On the logic of theory change: partial meet contraction and revision functions", *The Journal of Symbolic Logic*, **50**, 1985, 510–530.
4. Dubislav, W., "Zur Unbegrndbarkeit der Forderungstze", *Theoria*, **3**, 1937, 330–342.
5. Jørgensen, J., "Imperatives and logic", *Erkenntnis*, **7**, 1937-8, 288–296.
6. Makinson, D., "General Patterns in Nonmonotonic Reasoning", in: Gabbay, H. and Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 3, Oxford University Press, 1994, 35–110.
7. Makinson, D., "On a fundamental problem of deontic logic", in: McNamara, P. and Prakken, H. (eds.), *Norms, Logics and Information Systems. New Studies in Deontic Logic and Computer Science*, vol. 49 of *Frontiers in Artificial Intelligence and Applications*, Amsterdam: IOS Press, 1999, 29–53.
8. Makinson, D. and van der Torre, L., "Input/output logics", *J. Philosophical Logic*, **29**, 2000, 383–408.
9. Makinson, D. and van der Torre, L., "Constraints for input/output logics", *J. Philosophical Logic*, **30(2)**, 2001, 155–185.
10. Makinson, D. and van der Torre, L., "Permission from an input/output perspective", *J. Philosophical Logic*, **32(4)**, 2003, 391–416.
11. Makinson, D. and van der Torre, L., "What is Input/Output Logic?", in: *Foundations of the Formal Sciences II: Applications of Mathematical Logic in Philosophy and Linguistics*, vol. 17 of *Trends in Logic*, Kluwer, 2003.
12. Poole, D., "A logical framework for default reasoning", *Artificial Intelligence*, **36**, 1988, 27–47.
13. Prakken, H. and Sergot, M., "Contrary-to-duty obligations", *Studia Logica*, **57**, 1996, 91–115.
14. Reiter, R., "A logic for default reasoning", *Artificial Intelligence*, **13**, 1980, 81–132.
15. Stenius, E., "Principles of a logic of normative systems", *Acta Philosophica Fennica*, **16**, 1963, 247–260.

Ten Philosophical Problems in Deontic Logic

Jörg Hansen[1], Gabriella Pigozzi[2] and Leendert van der Torre[2]

[1] University of Leipzig, Institut für Philosophie
04107, Leipzig, Beethovenstraße 15, Germany
jhansen@uni-leipzig.de
[2] University of Luxembourg, Computer Science and Communications (CSC)
1359, Luxembourg, 6 rue Richard Coudenhove Kalergi, Luxembourg
{gabriella.pigozzi,leon.vandertorre}@uni.lu

Abstract. The paper discusses ten philosophical problems in deontic logic: how to formally represent norms, when a set of norms may be termed 'coherent', how to deal with normative conflicts, how contrary-to-duty obligations can be appropriately modeled, how dyadic deontic operators may be redefined to relate to sets of norms instead of preference relations between possible worlds, how various concepts of permission can be accommodated, how meaning postulates and counts-as conditionals can be taken into account, and how sets of norms may be revised and merged. The problems are discussed from the viewpoint of input/output logic as developed by van der Torre & Makinson. We argue that norms, not ideality, should take the central position in deontic semantics, and that a semantics that represents norms, as input/output logic does, provides helpful tools for analyzing, clarifying and solving the problems of deontic logic.

Keywords. Deontic logic, normative systems, input/output logic

Introduction

Deontic logic is the field of logic that is concerned with normative concepts such as obligation, permission, and prohibition. Alternatively, a deontic logic is a formal system that attempts to capture the essential logical features of these concepts. Typically, a deontic logic uses Ox to mean that it is obligatory that x, (or it ought to be the case that x), and Px to mean that it is permitted, or permissible, that x. The term 'deontic' is derived from the ancient Greek *déon*, meaning that which is binding or proper.

So-called Standard Deontic Logic (SDL) is a normal propositional modal logic of type KD, which means that it extends the propositional tautologies with the axioms $K : O(x \to y) \to (Ox \to Oy)$ and $D : \neg(Ox \land O\neg x)$, and it is closed under the inference rules *modus ponens* $x, x \to y/y$ and Necessitation x/Ox. Prohibition and permission are defined by $Fx = O\neg x$ and $Px = \neg O\neg x$. SDL is an unusually simple and elegant theory. An advantage of its modal-logical setting is that it can easily be extended with other modalities like epistemic or temporal operators and modal accounts of actions.

Dagstuhl Seminar Proceedings 07122
Normative Multi-agent Systems
http://drops.dagstuhl.de/opus/volltexte/2007/941

Not surprisingly for such a highly simplified theory, there are many features of actual normative reasoning that SDL does not capture. Notorious are the so-called 'paradoxes of deontic logic', which are usually dismissed as consequences of the simplifications of SDL. E.g. Ross's paradox [48], the counterintuitive derivation of "you ought to mail or burn the letter" from "you ought to mail the letter", is typically viewed as a side effect of the interpretation of 'or' in natural language. Many researchers seem to believe that the subject of deontic logic may be more or less finished, and we can focus on the use of deontic logic in computer science and agent theory, since there is nothing important left to add to it. In our view, this is far from the truth. On the contrary, there is a large number of important open problems in this field of research.

In this paper we discuss ten philosophical problems in deontic logic. All of these problems have been discussed in previous literature, and solutions have been offered, but we believe that all of them should be considered open and thus meriting further research. These problems are how deontic logic relates or applies to given sets of norms (imperatives, rules, aims) (sec. 1), what it means that a set of norms should be coherent (sec. 2), how conflicts of norms can be taken into account (sec. 3), how deontic logic should react to contrary-to-duty situations in which some norms are invariably violated (sec. 4), how to interpret dyadic deontic operators that formalize 'it ought to be that x on conditions α' as $O(x/\alpha)$ (sec. 5), how explicit permissions relate to, and change, an agent's obligations (sec. 6), how meaning postulates – norms that define legal terms – and constitutive norms, that create normative states of affairs, can be modeled (sec. 7 and 8), and how normative systems may be revised (sec. 9) and merged (sec. 10). Our choice is motivated by our aim at providing ourselves with models of normative reasoning of actual agents which may be human beings or computers, but the list of open problems is by no means final. Other problems may be considered equally important, such as how a hierarchy of norms (or of the norm-giving authorities) is to be respected, or how general norms relate to individual obligations, but we hope that our discussion provides the tools, and encourages the reader, to take a fresh look at these other problems, too.

To illustrate the problems, we use Makinson & van der Torre's input/output logic as developed in [42], [43], [44], and we therefore assume familiarity with this approach (cf. [45] for a good introduction). Input/output logic takes a very general view at the process used to obtain conclusions (more generally: outputs) from given sets of premises (more generally: inputs). While the transformation may work in the usual way, as an 'inference motor' to provide logical conclusions from a given set of premises, it might also be put to other, perhaps non-logical uses. Logic then acts as a kind of secretarial assistant, helping to prepare the inputs before they go into the machine, unpacking outputs as they emerge, and, less obviously, coordinating the two. The process as a whole is one of logically assisted transformation, and is an inference only when the central transformation is so. This is the general perspective underlying input/output logic. It is one of logic at work rather than logic in isolation; not some kind of non-classical logic, but a way of using the classical one.

1 Jørgensen's dilemma

While normative concepts are the subject of deontic logic, it is quite difficult how there can be a logic of such concepts at all. Norms like individual imperatives, promises, legal statutes, moral standards etc. are usually not viewed as being true or false. E.g. consider imperative or permissive expressions such as "John, leave the room!" and "Mary, you may enter now": they do not describe, but demand or allow a behavior on the part of John and Mary. Being non-descriptive, they cannot meaningfully be termed true or false. Lacking truth values, these expressions cannot – in the usual sense – be premise or conclusion in an inference, be termed consistent or contradictory, or be compounded by truth-functional operators. Hence, though there certainly exists a logical study of normative expressions and concepts, it seems there cannot be a logic of norms: this is Jørgensen's dilemma ([30], cf. [41]).

Though norms are neither true nor false, one may state that *according to the norms*, something ought to be (be done) or is permitted: the statements "John ought to leave the room", "Mary is permitted to enter", are then true or false descriptions of the normative situation. Such statements are sometimes called normative statements, as distinguished from norms. To express principles such as the principle of conjunction: $O(x \wedge y) \leftrightarrow (Ox \wedge Oy)$, with Boolean operators having truth-functional meaning at all places, deontic logic has resorted to interpreting its formulas Ox, Fx, Px not as representing norms, but as representing such normative statements. A possible logic of normative statements may then reflect logical properties of underlying norms – thus logic may have a "wider reach than truth", as von Wright [54] famously stated.

Since the truth of normative statements depends on a normative situation, like the truth of the statement "John ought to leave the room" depends on whether some authority ordered John to leave the room or not, it seems that norms must be represented in a logical semantics that models such truth or falsity. But semantics used to model the truth or falsity of normative statements mostly fail to include norms. Standard deontic semantics evaluates deontic formulas with respect to sets of worlds, in which some are ideal or better than others – Ox is then defined true if x is true in all ideal or the best reachable worlds. In our view, norms, not ideality, should take the central position from which normative statements are evaluated. Then the following question arises, pointedly asked by D. Makinson in [41]:

Problem 1. How can deontic logic be reconstructed in accord with the philosophical position that norms are neither true nor false?

In the older literature on deontic logic there has been a veritable 'imperativist tradition' of authors that have, deviating from the standard approach, in one way or other, tried to give truth definitions for deontic operators with respect to given sets of norms.[3] The reconstruction of deontic logic as logic about imperatives

[3] Cf. among others S. Kanger [32], E. Stenius [53], T. J. Smiley [51], Z. Ziemba [62], B. van Fraassen [15], Alchourrón & Bulygin [1] and I. Niiniluoto [47].

has been the project of one of the authors beginning with [19]. Makinson & van der Torre's input/output logic [42] is another reconstruction of a logic of norms in accord with the philosophical position that norms direct rather than describe, and are neither true nor false. Suppose that we have a set G (meant to be a set of conditional norms), and a set A of formulas (meant to be a set of given facts). The problem is then: how may we reasonably define the set of propositions x making up the output of G given A, which we write $out(G, A)$? In particular, if we view the output as descriptions of states of affairs that ought to obtain given the norms G and the facts A, what is a reasonable output operation that enables us to define a deontic O-operator that describes the normative statements that are true given the norms and the facts, we say: the normative consequences given the situation? One such definition is the following:

$$G, A \models Ox \;\text{ iff }\; x \in out(G, A)$$

So Ox is true iff the output of G under A includes x. Note that this is rather a description of how we think such an output should or might be interpreted, whereas 'pure' input/output logic does not discuss such definitions. For a simple case, let G include a conditional norm that states that if a is the case, x should obtain (we write $(a, x) \in G$).[4] If a can be inferred from A, i.e. if $a \in Cn(A)$, and z is logically implied by x, then z should be among the normative consequences of G given A. An operation that does this is simple-minded output out_1:

$$out_1(G, A) \;=\; Cn(G(Cn(A)))$$

where $G(B) = \{y \mid (b, y) \in G \text{ and } b \in B\}$. So in the given example, Oz is true given $(a, x) \in G$, $a \in Cn(A)$ and $z \in Cn(x)$.

Simple-minded output may, however, not be strong enough. Sometimes, legal argumentation supports reasoning by cases: if there is a conditional norm (a, x) that states that an agent must bring about x if a is the case, and a norm (b, x) that states that the same agent must also bring about x if b is the case, and $a \vee b$ is implied by the facts, then we should be able to conclude that the agent must bring about x. An operation that supports such reasoning is basic output out_2:

$$out_2(G, A) \;=\; \cap\{Cn(G(V)) \mid v(A) = 1\}$$

where v ranges over Boolean valuations plus the function that puts $v(b) = 1$ for all formulae b, and $V = \{b \mid v(b) = 1\}$. It can easily be seen that now Ox is true given $\{(a, x), (b, x)\} \subseteq G$ and $a \vee b \in Cn(A)$.

It is quite controversial whether reasoning with conditional norms should support 'normative' or 'deontic detachment', i.e. whether it should be accepted that if one norm (a, x) commands an agent to make x true in conditions a, and another norm (x, y) directs the agent to make y true given x is true, then the agent has an obligation to make y true if a is factually true. Some would argue that as long as the agent has not in fact realized x, the norm to bring about y is not 'triggered'; others would maintain that obviously the agent has an obligation to make $x \wedge y$ true given that a is true. If such detachment is viewed

[4] As has become usual, an unconditional norm that commits the agent to realizing x is represented by a conditional norm $(\top, x)$, where $\top$ means an arbitrary tautology.

as permissible for normative reasoning, then one might use reusable output out_3 that supports such reasoning:

$$out_3(G, A) \;=\; \cap\{Cn(G(B)) \mid A \subseteq B = Cn(B) \supseteq G(B)\}$$

An operation that combines reasoning by cases with deontic detachment is then reusable basic output out_4:

$$out_4(G, A) \;=\; \cap\{Cn(G(V)) : v(A) = 1 \text{ and } G(V) \subseteq V\}$$

Finally, it is often required to reconsider the facts when drawing conclusions about what an agent must do: suppose there is an unconditional norm $(\top, x \vee y)$ to bring about $x \vee y$, but that the agent cannot realize x as the facts include $\neg x$. We would like to say that then the agent must bring about y, as this is the only possible way left to satisfy the norm. To do this, one may use the throughput versions out_n^+ of any of the output operations $out_1, out_2, out_3, out_4$,

$$out_n^+(G, A) \;=\; out_n(G^+, A),$$

where $G^+ = G \cup I$ and I is the set of all pairs (a, a) for formulae a. The choice of the throughput versions might appear questionable, since each makes Ox true in case $x \in Cn(A)$, i.e. it makes the unalterable facts obligatory.

It may turn out that further modifications of the output operation are required in order to produce reasonable results for normative reasoning. Also, the proposal to employ input/output logic to reconstruct deontic logic may lead to competing solutions, depending on what philosophical views as to what transformations should be acceptable one subscribes to. All this is what input/output logic is about. However, it should be noted that input/output logic succeeds in representing norms as entities that are neither true nor false, while still permitting normative reasoning about such entities.

2 Coherence

Consider norms which on one hand require you to leave the room, while on the other requiring you not to leave the room at the same time. In such cases, we are inclined to say that there is something wrong with the normative system. This intuition is captured by the SDL axiom $D : \neg(Ox \wedge O\neg x)$ that states that there cannot be co-existing obligations to bring about x and to bring about $\neg x$, or, using the standard cross-definitions of the deontic modalities: x cannot be both, obligatory and forbidden, or: if x is obligatory then it is also permitted. But what does this tell us about the normative system?

Since norms do not bear truth values, we cannot, in any usual sense, say that such a set of norms is inconsistent. All we can consider is the consistency of the output of a set of norms. We like to use the term *coherence* with respect to a set of norms with consistent output, and define:

(1) A set of norms G is coherent iff $\perp \notin out(G, A)$.

However, this definition seems not quite sufficient: one might argue that one should be able to determine whether a set of norms G is coherent or not regardless of what arbitrary facts A might be assumed. A better definition would be $(1a)$:

($1a$) A set of norms G is coherent iff there exists a set of formulas A such that $\bot \notin out(G, A)$.

For ($1a$) it suffices that there exists a situation in which the norms can be, or could have been, fulfilled. However, consider the set of norms $G = \{(a, x), (a, \neg x)\}$ that requires both x to be realized and $\neg x$ to be realized in conditions a: it is immediate that e.g. for all output operations $out_n^{(+)}$, we have $\bot \notin out_n^{(+)}(G, \neg a)$: no conflicting demands arise when $\neg a$ is factually assumed. Yet something seems wrong with a normative system that explicitly considers a fact a only to tie to it conflicting normative consequences. The dual of ($1a$) would be

($1b$) A set of norms G is coherent iff for all sets of formulas A, $\bot \notin out(G, A)$.

Now a set G with $G = \{(a, x), (a, \neg x)\}$ would no longer be termed coherent. ($1b$) makes the claim that for no situation A, two norms $(a, x), (b, y)$ would ever come into conflict, which might seem too strong. We may wish to restrict A to sets of facts that are consistent, or that are not in violation of the norms. The question is, basically, how to distinguish situations that the norm-givers should have taken care of, from those that describe misfortune of otherwise unhappy circumstances. A weaker claim than ($1b$) would be ($1c$):

($1c$) A set of norms G is coherent iff for all a with $(a, x) \in G$, $\bot \notin out(G, a)$.

By this change, consistency of output is required just for those factual situations that the norm-givers have foreseen, in the sense that they have explicitly tied normative consequences to such facts. Still, ($1c$) might require further modification, since if a is a foreseen situation, and so is b, then also $a \vee b$ or $a \wedge b$ might be counted as foreseen situations for which the norms should be coherent.

However, there is a further difficulty: let G contain a norm $(a, \neg a)$ that, for conditions in which a is unalterably true, demands that $\neg a$ be realized. We then have $\neg a \in out_n(G, a)$ for the principal output operations out_n, but not $\bot \in out_n(G, a)$. Certainly the term 'incoherent' should apply to a normative system that requires the agent to accomplish what is – given the facts in which the duty arises – impossible. But since not every output operation supports 'throughput', i.e. the input is not necessarily included in the output, neither (1) nor its variants implies that the agent can actually realize all propositions in the output, though they might be logically consistent. We might therefore demand that the output is not consistent *simpliciter*, but consistent with the input:

(2) A set of norms G is coherent iff $\bot \notin out(G, A) \cup A$.

But with definition (2) we obtain the questionable result that for any case of norm-violation, i.e. for any case in which $(a, x) \in G$ and $(a \wedge \neg x) \in Cn(A)$, G must be termed incoherent – Adam's fall would only indicate that there was something wrong with God's commands. One remedy would be to leave aside all those norms that are invariably violated, i.e. instead of $out(G, A)$ consider $out(\{(a, x) \in G \mid (a \wedge \neg x) \notin Cn(A)\}, A)$ – but then a set G such that $(a, \neg a) \in G$ would not be incoherent. It seems it is time to formally state our problem:

Problem 2. When is a set of norms to be termed 'coherent'?

As can be seen from the discussion above, input/output logic provides the tools to formally discuss this question, by rephrasing the question of coherence of the norms as one of consistency of output, and of output with input. Both notions have been explored in the input/output framework as 'output under constraints':

Definition (Output under constraints) *Let G be a set of conditional norms and A and C two sets of propositional formulas. Then G is coherent in A under constraints C when $out(G, A) \cup C$ is consistent.*

Future study must define an output operation, determine the relevant states A, and find the constraints C, such that any set of norms G would be appropriately termed coherent or incoherent by this definition.

3 Normative conflicts and dilemmas

There are essentially two views on the question of normative conflicts: in the one view, they do not exist. In the other view, conflicts and dilemmas are ubiquitous.

According to the view that normative conflicts are ubiquitous, it is obvious that we may become the addressees of conflicting normative demands at any time. My mother may want me to stay inside while my brother wants me to go outside with him and play games. I may have promised to finish a paper until the end of a certain day, while for the same day I have promised a friend to come to dinner – now it is late afternoon and I realize I will not be able to finish the paper if I visit my friend. Social convention may require me to offer you a cigarette when I am lighting one for myself, while concerns for your health should make me not offer you one. Legal obligations might collide - think of the recent case where the SWIFT international money transfer program was required by US anti-terror laws to disclose certain information about its customers, while under European law that also applied to that company, it was required not to disclose this information. Formally, let there be two conditional norms (a, x) and (b, y): unless we have that either $(x \rightarrow y) \in Cn(a \wedge b)$ or $(y \rightarrow x) \in Cn(a \wedge b)$ there is a possible situation $a \wedge b \wedge \neg(x \wedge y)$ in which the agent can still satisfy each norm individually, but not both norms collectively. But to assume the former for any two norms (a, x) and (b, y) is clearly absurd.[5] So any logic about norms must take into account possible conflicts. But standard deontic logic SDL includes D: $\neg(Ox \wedge O\neg x)$ as one of its axioms, and it is not quite immediate how deontic reasoning could accommodate conflicting norms. The problem is thus:

Problem 3a. How can deontic logic accommodate possible conflicts of norms?

In an input/output setting one could say that there exists a conflict whenever $\bot \in Cn(out(G, A) \cup A)$, i.e. whenever the output is inconsistent with the input: then the norms cannot all be satisfied in the given situation. There appear to be two ways to proceed when such inconsistencies cannot be ruled out.[6]. For both, it is necessary to recur to the the notion of a $maxfamily(G, A, A)$, i.e. the

[5] Nevertheless, Lewis' [36], [37] and Hansson's [24] deontic semantics imply that there exists a 'system of spheres', in our setting: a sequence of boxed contrary-to-duty norms $(\top, x_1), (\neg x_1, x_2), (\neg x_1 \wedge \neg x_2, x_3), ...$ that satisfies this condition.

[6] For the concepts underlying the 'some-things-considered' and 'all-things-considered' O-operators defined below cf. Horty [28] and Hansen [20], [21]

family of all maximal $H \subseteq G$ such that $out(H, A) \cup A$ is consistent. On this basis, input/output logic defines the following two output operations $out^{\cup}$ and $out^{\cap}$:

$$\begin{aligned} out^{\cup}(G, A) &= \bigcup\{out(H, A) \mid H \in maxfamily(G, A, A)\} \\ out^{\cap}(G, A) &= \bigcap\{out(H, A) \mid H \in maxfamily(G, A, A)\} \end{aligned}$$

Note that $out^{\cup}$ is a non-standard output operation that is not closed under consequences, i.e. we do not generally have $Cn(out^{\cup}(G, A)) = out^{\cup}(G, A)$. Finally we may use the intended definition of an O-operator

$$G, A \models Ox \text{ iff } x \in out(G, A)$$

to refer to the operations $out^{\cup}$ and $out^{\cap}$, rather than the underlying operation $out(G, A)$ itself, and write $O^{\cup}x$ and $O^{\cap}x$ to mean that $x \in out^{\cup}(G, A)$ and $x \in out^{\cap}(G, A)$, respectively. Then the 'some-things-considered', or 'bold' O-operator $O^{\cup}$ describes x as obligatory given the set of norms G and the facts A if x is in the output of some $H \in maxfamily(G, A, A)$, i.e. if some subset of non-conflicting norms, or: some coherent normative standard embedded in the norms, requires x to be true. It is immediate that neither the SDL axiom $D : \neg(Ox \wedge O\neg x)$ nor the agglomeration principle $C : Ox \wedge Oy \rightarrow O(x \wedge y)$ holds for $O^{\cup}$, as there may be two competing standards demanding x and $\neg x$ to be realized, while there may be none that demands the impossible $x \wedge \neg x$. On the other hand, the 'all-things-considered', or 'sceptic', O-operator $O^{\cap}$ describes x as obligatory given the norms G and the facts A if x is in the outputs of all $H \in maxfamily(G, A, A)$, i.e. it requires that x must be realized according to all coherent normative standards. Note that by this definition, both SDL theorems D and C are validated.

The opposite view, that normative conflicts do not exist, appeals to the very notion of obligation: it is essential for the function of norms to direct human behavior that the subject of the norms is capable of following them. To state a norm that cannot be fulfilled is a meaningless use of language. To state two norms which cannot both be fulfilled is confusing the subject, not giving him or her directions. To say that a subject has two conflicting obligations is therefore a misuse of the term 'obligation'. So there cannot be conflicting obligations, and if things appear differently, a careful inspection of the normative situation is required that resolves the dilemma in favor of the one or other of what only appeared both to be obligations. In particular, this inspection may reveal a priority ordering of the apparent obligations that helps resolve the conflict (this summarizes viewpoints prominent e.g. in Ross [49], von Wright [59], [60], and Hare [25]). The problem that arises for such a view is then how to determine the 'actual obligations' in face of apparent conflicts, or, put differently, in the face of conflicting 'prima facie' obligations.

Problem 3b. How can the resolution of apparent conflicts be semantically modeled?

Again, both the $O^{\cup}$ and the $O^{\cap}$-operator may help to formulate and solve the problem: $O^{\cup}$ names the conflicting *prima facie* obligations that arise from a set of norms G in a given situation A, whereas $O^{\cap}$ resolves the conflict by telling the agent to do only what is required by all maximal coherent subsets of the norms: so there might be conflicting 'prima facie' $O^{\cup}$-obligations, but no conflicting 'all things considered' $O^{\cap}$-obligations. The view that a priority ordering helps to resolve conflicts seems more difficult to model. A good approach appears to be to let the priorities help us to select a set $\mathscr{P}(G,A,A)$ of preferred maximal subsets $H \in \mathit{maxfamily}(G,A,A)$. We may then define the $O^{\cap}$-operator not with respect to the whole of $\mathit{maxfamily}(G,A,A)$, but only with respect to its selected preferred subsets $\mathscr{P}(G,A,A)$. Ideally, in order to resolve all conflicts, the priority ordering should narrow down the selected sets to $\mathit{card}(\mathscr{P}(G,A,A)) = 1$, but this generally requires a strict ordering of the norms in G. The demand that all norms can be strictly ordered is itself subject of philosophical dispute: some moral requirements may be incomparable (this is Sartre's paradox, where the requirement that Sartre's student stays with his ailing mother conflicts with the requirement that the student joins the resistance against the German occupation), while others may be of equal weight (e.g. two simultaneously obtained obligations towards identical twins, of which only one can be fulfilled). The difficult part is then to define a mechanism that determines the preferred maximal subsets by use of the given priorities between the norms. There have been several proposals to this effect, not all of them successful, and the reader is referred to the discussions in Boella & van der Torre [8] and Hansen [22], [23].

4 Contrary-to-duty reasoning

Suppose we are given a code G of conditional norms, that we are presented with a condition (input) that is unalterably true, and asked what obligations (output) it gives rise to. It may happen that the condition is something that should not have been true in the first place. But that is now water under the bridge: we have to "make the best out of the sad circumstances" as B. Hansson [24] put it. We therefore abstract from the deontic status of the condition, and focus on the obligations that are consistent with its presence. How to determine this in general terms, and if possible in formal ones, is the well-known problem of contrary-to-duty conditions as exemplified by the notorious contrary-to-duty paradoxes. Chisholm's paradox [13] consists of the following four sentences:

(1) It ought to be that a certain man go to the assistance of his neighbors.
(2) It ought to be that if he does go, he tell them he is coming.
(3) If he does not go then he ought not to tell them he is coming.
(4) He does not go.

Furthermore, intuitively, the sentences derive (5):

(5) He ought not to tell them he is coming.

Chisholm's paradox is a contrary-to-duty paradox, since it contains both a primary obligation to go, and a secondary obligation not to call if the agent does not go. Traditionally, the paradox was approached by trying to formalize each of the sentences in an appropriate language of deontic logic, and then consider the sets $\{Ox, O(x \rightarrow z), O(\neg x \rightarrow \neg z), \neg x\}$, or $\{Ox, x \rightarrow Oz, \neg x \rightarrow O\neg z, \neg x\}$, or $\{Ox, O(x \rightarrow z), \neg x \rightarrow O\neg z, \neg x\}$ or $\{Ox, x \rightarrow Oz, O(\neg x \rightarrow \neg z), \neg x\}$. But whatever approach is taken, it turned out that either the set of formulas is traditionally inconsistent or inconsistent in SDL, or one formula is a logical consequence – by traditional logic or in SDL – of another formula. Yet intuitively the natural-language expressions that make up the paradox are consistent and independent from each other: this is why it is called a paradox. Though the development of dyadic deontic operators as well as the introduction of temporally relative deontic logic operators can be seen as a direct result of Chisholm's paradox, the paradox seems so far unsolved. The problem is thus:

Problem 4. How do we reason with contrary-to-duty obligations which are in force only in case of norm violations?

In the input/output logic framework, the strategy for eliminating excess output is to cut back the set of generators to just below the threshold of yielding excess. To do that, input/output logic looks at the maximal non-excessive subsets, as described by the following definition:

Definition (Maxfamilies) *Let G be a set of conditional norms and A and C two sets of propositional formulas. Then maxfamily(G, A, C) is the set of maximal subsets $H \subseteq G$ such that $out(H, A) \cup C$ is consistent.*

For a possible solution to Chisholm's paradox, consider the following output operation $out^{\cap}$:

$$out^{\cap}(G, A) \;=\; \bigcap\{out(H, A) \mid H \in maxfamily(G, A, A)\}$$

So an output x is in $out^{\cap}(G, A)$ if it is in output $out(H, A)$ of all maximal norm subsets $H \subseteq G$ such that $out(H, A)$ is consistent with the input A. Let a deontic O-operator be defined in the usual way with regard to this output:

$$G, A \models O^{\cap}x \;\text{ iff }\; x \in out^{\cap}(G, A)$$

Furthermore, tentatively, and only for the task of shedding light on Chisholm's paradox, let us define an entailment relation between norms as follows:

Definition (Entailment relation) *Let G be a set of conditional norms, and (a, x) be a norm whose addition to G is under consideration. Then (a, x) is entailed by G iff for all sets of propositions A, $out^{\cap}(G \cup \{(a, x)\}, A) = out^{\cap}(G, A)$.*

So a (considered) norm is entailed by a (given) set of norms if its addition to this set would not make a difference for any set of facts A. Finally, let us use the following cautious definition of 'coherence from the start' (also called 'minimal coherence' or 'coherence per se'):

A set of norms G is 'coherent from the start' iff $\bot \notin out(G, \top)$.

Now consider a 'Chisholm norm set' $G = \{(\top, x), (x, z), (\neg x, \neg z), \}$, where $(\top, x)$ means the norm that the man must go to the assistance of his neighbors, (x, z) means the norm that it ought to be that if he goes he ought to tell them he is coming, and $(\neg x, \neg z)$ means the norm that if he does not go he ought not to tell them he is coming. It can be easily verified that the norm set G is 'coherent from the start' for all standard output operations $out_n^{(+)}$, since for these either $out(G, \top) = Cn(\{x\})$ or $out(G, \top) = Cn(\{x, z\})$, and both sets $\{x\}$ and $\{x, z\}$ are consistent. Furthermore, it should be noted that all norms in the norm set G

are independent from each other, in the sense that no norm $(a, x) \in G$ is entailed by $G \setminus \{(a, x)\}$ for any standard output operation $out_n^{(+)}$: for $(\top, x)$ we have $x \in out^{\cap}(G, \top)$ but $x \notin out^{\cap}(G \setminus \{(\top, x)\}, \top)$, for (x, z) we have $z \in out^{\cap}(G, x)$ but $z \notin out^{\cap}(G \setminus \{(x, z)\}, x)$, and for $(\neg x, \neg z)$ we have $\neg z \in out^{\cap}(G, \neg x)$ but $\neg z \notin out^{\cap}(G \setminus \{(\neg x, \neg z)\}, \top)$. Finally consider the 'Chisholm fact set' $A = \{\neg x\}$, that includes as an assumed unalterable fact the proposition $\neg x$, that the man will not go to the assistance of his neighbors: we have $maxfamily(G, A, A) = \{G \setminus \{(\top, x)\}\} = \{\{(x, z), (\neg x, \neg z), \}\}$ and either $out(G \setminus \{(\top, x)\}, A) = Cn(\{\neg z\})$ or $out(G \setminus \{(\top, x)\}, A) = Cn(\{\neg x, \neg z\})$ for all standard output operations $out_n^{(+)}$, and so $O^{\cap} \neg z$ is true given the norm and fact sets G and A, i.e. the man must not tell his neighbors he is coming.

5 Descriptive dyadic obligations

Dyadic deontic operators, that formalize e.g. 'x ought to be true under conditions a' as $O(x/a)$, were introduced over 50 years ago by G. H. von Wright [56]. Their introduction was due to Prior's paradox of derived obligation: often a primary obligation Ox is accompanied by a secondary, 'contrary-to-duty' obligation that pronounces y (a sanction, a remedy) as obligatory if the primary obligation is violated. At the time, the usual formalization of the secondary obligation would have been $O(\neg x \rightarrow y)$, but given Ox and the axioms of standard deontic logic SDL, $O(\neg x \rightarrow y)$ is derivable for any y. A bit later, Chisholm's paradox showed that formalizing the secondary obligation as $\neg x \rightarrow Oy$ produces similarly counterintuitive results. So to deal with such contrary-to-duty conditions, the dyadic deontic operator $O(x/a)$ was invented.

The perhaps best-known semantic characterization of dyadic deontic logic is Bengt Hansson's [24] system *DSLD*3, axiomatized by Spohn [52]. Hansson's idea was that the circumstances (the conditions a) are something which has actually happened (or will unalterably happen) and which cannot be changed afterwards. Ideal worlds in which $\neg a$ is true are therefore excluded. But some worlds may still be better than others, and there should then be an obligation to make "'the best out of the sad circumstances". Consequently, Hansson presents a possible worlds semantics in which all worlds are ordered by a preference (betterness) relation. $O(x/a)$ is then defined true if x is true in the best a-worlds. Here, we intend to employ semantics that do not make use of any prohairetic betterness relation, but that models deontic operators with regard to given sets of norms and facts, and the question is then

Problem 5. How to define dyadic deontic operators with regard to given sets of norms and facts?

Input/output logic assumes a set of (conditional) norms G, and a set of invariable facts A. The facts A may describe a situation that is inconsistent with the output $out(G, A)$: suppose there is a primary norm $(\top, a) \in G$ and a secondary norm $(\neg a, x) \in G$, i.e. $G = \{(\top, a), (\neg a, x)\}$, and $A = \{\neg a\}$. Though

$a \in out(G, A)$, it makes no sense to describe a as obligatory since a cannot be realized any more in the given situation – no crying over spilt milk. Rather, the output should include only the consequent of the secondary obligation x – it is the best we can make out of these circumstances. To do so, we return to the definitions of $maxfamily(G, A, A)$ as the set of all maximal subsets $H \subseteq G$ such that $out(H, A) \cup A$ is consistent, and the set $out^{\cap}(G, A)$ as the intersection of all outputs from $H \in maxfamily(G, A, A)$, i.e. $out^{\cap}(G, A) = \bigcap\{out(H, A) \mid H \in maxfamily(G, A, A)\}$. We may then define:

$$G \models O(x/a) \text{ iff } x \in out^{\cap}(G, \{a\})$$

Thus, relative to the set of norms G, $O(x/a)$ is defined true if x is in the output under a of all maximal sets H of norms such that their output under $\{a\}$ is consistent with a. In the example where $G = \{(\top, a), (\neg a, x)\}$ we therefore obtain $O(x/\neg a)$ but not $O(a/\neg a)$ as being true, i.e. only the consequent of the secondary obligation is described as obligatory in conditions $\neg a$.

In the above definition, the antecedent a of the dyadic formula $O(x/a)$ makes the inputs explicit: the truth definition does not make use of any facts other than a. This may be unwanted; one might consider an input set A of *given* facts, and employ the antecedent a only to denote an additional, *assumed* fact. Still, the output should contradict neither the given nor the assumed facts, and the output should include also the normative consequences x of a norm (a, x) given the assumed fact a. This may be realized by the following definition:

$$G, A \models O(x/a) \text{ iff } x \in out^{\cap}(G, A \cup \{a\})$$

So, relative to a set of norms G and a set of facts A, $O(x, a)$ is defined true if x is in the output under $A \cup \{a\}$ of all maximal sets H of norms such that their output under $A \cup \{a\}$ is consistent with $A \cup \{a\}$.

Hansson's description of dyadic deontic operators as describing defeasible obligations that are subject to change when more specific, namely contrary-to-duty situations emerge, may be the most prominent view, but it is by no means the only one. Earlier authors like von Wright [57] [58] and Anderson [4] have proposed more normal conditionals, which in particular support 'strengthening of the antecedent' SA $O(x/a) \rightarrow O(x/a \wedge b)$. From an input/output perspective, such operators can be accommodated by defining

$$G, A \models O(x/a) \text{ iff } x \in out(G, A \cup \{a\})$$

It is immediate that for all standard output operations $out_n^{(+)}$ this definition validates SA. The properties of dyadic deontic operators that are, like the above, semantically defined within the framework of input/output logic, have not been studied so far. The theorems they validate will inevitably depend on what output operation is chosen (cf. [23] for some related conjectures).

6 Permissive norms

In formal deontic logic, permission is studied less frequently than obligation. For a long time, it was naively assumed that it can simply be taken as a dual of obligation, just as possibility is the dual of necessity in modal logic. Permission is then defined as the absence of an obligation to the contrary, and the modal operator P defined by $Px =_{def} \neg O \neg x$. Today's focus on obligations is not only in stark contrast how deontic logic began, for when von Wright [55] started modern deontic logic in 1951, it was the P-operator that he took as primitive, and defined obligation as an absence of a permission to the contrary. Rather, more and more authors have come to realize how subtle and multi-faceted the concept of permission is. Much energy was devoted to solving the problem of 'free choice permission', where one may derive from the statement that one is permitted to have a cup of tea or a cup of coffee that it is permitted to have a cup of tea, and it is permitted to have a cup of coffee, or for short, that $P(x \vee y)$ implies Px and Py (cf. [31]. Von Wright, in his late work starting with [61], dropped the concept of inter-definability of obligations and permissions altogether by introducing P-norms and O-norms, where one may call something permitted only if it derives from the collective contents of some O-norms and at most one P-norm. This concept of 'strong permission' introduced deontic 'gaps': whereas in standard deontic logic SDL, $O\neg x \vee Px$ is a tautology, meaning that any state of affairs is either forbidden or permitted, von Wright's new theory means that in the absence of explicit P-norms only what is obligatory is permitted, and that nothing is permitted if also O-norms are missing. Perhaps most importantly, Bulygin [12] observed that an authoritative kind of permission must be used in the context of multiple authorities and updating normative systems: if a higher authority permits you to do something, a lower authority can no longer prohibit it. Summing up, the understanding of permission is still in a less satisfactory state than the understanding of obligation and prohibition. The problem can be phrased thus:

Problem 6. How to distinguish various kinds of permissions and relate them to obligations?

¿From the viewpoint of input/output logic, one may first try to define a concept

of negative permission in the line of the classic approach. Such a definition is the following:

$$G, A \models P^{neg}x \ \text{ iff } \ \neg x \notin out(G, A)$$

So something is permitted by a code iff its negation is not obligatory according to the code and in the given situation. As innocuous and standard as such a definition seems, questions arise as to what output operation out may be used. Simple-minded output out_1 and basic output out_2 produce counterintuitive results: consider a set of norms G of which one norm $(work, tax)$ demands that if I am employed then I have to pay tax. For the default situation $A = \{\top\}$ then $P^{neg}(a \wedge \neg x)$ is true, i.e. it is by default permitted that I am employed and do not pay tax. Stronger output operations out_3 and out_4 that warrant reusable output exclude this result, but their use in deontic reasoning is questionable for other reasons.

In contrast to a concept of negative permission, one may also define a concept of 'strong' or 'positive permission'. This requires a set P of explicit permissive norms, just as G is a set of explicit obligations. As a first approximation, one may say that something is positively permitted by a code iff the code explicitly presents it as such. But this leaves a central logical question unanswered as to how explicitly given permissive and obligating norms may generate permissions that – in some sense – follow from the explicitly given norms. In the line of von Wrights later approach, we may define:

$$G, P \models P^{stat}(x/a) \;\text{ iff }\; x \in out(G \cup \{(b,y)\}, a) \text{ for some } (b,y) \in P \cup \{(\top,\top)\}$$

So there is a permission to realize x in conditions a if x is generated under these conditions either by the norms in G alone, or the norms in G together with some explicit permission (b, y) in P. We call this a 'static' version of strong permission. For example, consider a set G consisting of the norm $(\mathit{work}, \mathit{tax})$, and a set P consisting of the sole license $(\mathit{18y}, \mathit{vote})$ that permits all adults to take part in political elections. Then all of the following are true: $P^{stat}(\mathit{tax}/\mathit{work})$, $P^{stat}(\mathit{vote}/\mathit{18y})$, $P^{stat}(\mathit{tax}/\mathit{work} \wedge \mathit{male})$ and also $P^{stat}(\mathit{vote}/\neg \mathit{work} \wedge \mathit{18y})$ (so even unemployed adults are permitted to vote).

Where negative permission is liberal, in the sense that anything is permitted that does not conflict with ones obligations, the concept of static permission is quite strict, as nothing is permitted that does not explicitly occur in the norms. In between, one may define a concept of 'dynamic permission' that defines something as permitted in some situation a if forbidding it for these conditions would prevent an agent from making use of some explicit (static) permission. The formal definition reads:

$$\begin{aligned} G, P \models P^{dyn}(x/a) \;\text{ iff }\; & \neg y \in out(G \cup \{(a, \neg x)\}, b) \text{ for some } y \text{ and conditions} \\ & b \text{ such that } G, P \models P^{stat}(y/b) \end{aligned}$$

Consider the above static permission $P^{stat}(\mathit{vote}/\neg \mathit{work} \wedge \mathit{18y})$ that even the unemployed adult populations is permitted to vote, generated by the sets $P = \{(\mathit{18y}, \mathit{vote})\}$ and $G = \{(\mathit{work}, \mathit{tax})\}$. We might also like to say, without reference to age, that the unemployed are protected from being forbidden to vote, and in this sense are permitted to vote, but $P^{stat}(\mathit{vote}/\neg \mathit{work})$ is not true. And we might like to say that adults are protected from being forbidden to vote unless they are employed, and in this sense are permitted to be both unemployed and take part in elections, but also $P^{stat}(\neg \mathit{work} \wedge \mathit{vote}/\mathit{18y})$ is not true. Dynamic permissions allow us to express such protections, and make both $P^{dyn}(\mathit{vote}/\neg \mathit{work})$ and $P^{dyn}(\neg \mathit{work} \wedge \mathit{vote}/\mathit{18y})$ true: if either $(\neg \mathit{work}, \neg \mathit{vote})$ or $(18y, (\neg \mathit{work} \rightarrow \neg \mathit{vote}))$ were added to G we would obtain $\neg \mathit{vote}$ as output in conditions $\neg \mathit{work} \wedge \mathit{18y})$ in spite of the fact that, as we have seen, $G, P \models P^{stat}(\mathit{vote}/\neg \mathit{work} \wedge \mathit{18y})$.

There are, ultimately, a number of questions for all these concepts of permissions that have been further explored in [44]. Other kinds of permissions have been discussed from an input/output perspective in the literature, too, for example permissions as exceptions of obligations [8]. But it seems input/output logic is able to help clarify the underlying concepts of permission better than traditional deontic semantics.

7 Meaning postulates and intermediate concepts

To define a deontic operator of individual obligation seems straightforward if the norm in question is an individual command or act of promising. For example, if you are the addressee α of the following imperative sentence

(1) You, hand me that screwdriver, please.

and you consider the command valid, then what you ought to do is to hand the screwdriver in question to the person β uttering the request. In terms of input/output logic, let x be the proposition that α hands the screwdriver to β: with the set of norms $G = \{(\top, x)\}$, the set of facts $A = \{\top\}$, and the truth definition Ox iff $x \in out(A, G)$: then we obtain that Ox is true, i.e. it is true that it ought to be that α hands the screwdriver to β.

Norms that belong to a legal system are more complex, and thus more difficult to reason about. Consider, for example

(2) An act of theft is punished by a prison sentence not exceeding 5 years or a fine.

Things are again easy if you are a judge and you know that the accused in front of you has committed an act of theft – then you ought to hand out a verdict that commits the accused to pay a fine or to serve a prison sentence not exceeding 5 years. But how does the judge arrive at the conclusion that an act of theft has been committed? ‘Theft’ is a legal term that is usually accompanied by a legal definition such as the following one:

(3) Someone commits an act of theft if that person has taken a movable object from the possession of another person into his own possession with the intention to own it, and if the act occurred without the consent of the other person or some other legal authorization.

It is noteworthy that (3) is not a norm in the strict sense – it does not prescribe or allow a behavior – but rather a stipulative definition, or, in more general terms, a *meaning postulate* that constitutes the legal meaning of theft. Such sentences are often part of the legal code. They share with norms the property of being neither true nor false. The significance of (3) is that it decomposes the complex legal term ‘theft’ into more basic legal concepts. These concepts are again the subject of further meaning postulates, among which may be the following:

(4) A person in the sense of the law is a human being that has been born.
(5) A movable object is any physical object that is not a person or a piece of land.
(6) A movable object is in the possession of a person if that person is able to control the uses and the location of the object.
(7) The owner of an object is – within the limits of the law – entitled to do with it whatever he wants, namely keep it, use it, transfer possession or ownership of the object to another person, and destroy or abandon it.

Not all of definitions (4)-(7) may be found in the legal statutes, though they may be viewed as belonging to the normative system by virtue of having been accepted in legal theory and judicial reasoning. They constitute ‘intermediate concepts': they link legal terms (person, movable object, possession etc.) to words describing natural facts (human being, born, piece of land, keep an object etc.).

Any proper representation of legal norms must include means of representing meaning postulates that define legal terms, decompose legal terms into more basic legal terms, or serve as intermediate concepts that link legal terms to terms that describe natural facts. But for deontic logic, with its standard possible worlds semantics, a comprehensive solution to the problem of representing meaning postulates is so far lacking (cf. Lindahl [39]). The problem is thus:

Problem 7. How can meaning postulates and intermediate terms be modeled in semantics for deontic logic reasoning?

The representation of intermediate concepts is of particular interest, since such concepts arguably reduce the number of implications required for the transition from natural facts to legal consequences and thus serve an economy of expression (cf. Lindahl & Odelstad [40]). Lindahl & Odelstad use the term 'ownership' as an example to argue as follows: let $F_1, ..., F_p$ be descriptions of some situations in which a person α acquires ownership of an object γ, e.g. by acquiring it from some other person β, finding it, building it from owned materials, etc., and let $C_1, ..., C_n$ be among the legal consequences of α's ownership of γ, e.g. freedom to use the object, rights to compensation when the object is damaged, obligations to maintain the object or pay taxes for it etc. To express that each fact F_i has the consequence C_j, $p \times n$ implications are required. The introduction of the term $\mathit{Ownership}(x, y)$ reduces the number of required implications to $p + n$: there are p implications that link the facts $F_1, ..., F_p$ to the legal term $\mathit{Ownership}(x, y)$, and n implications that link the legal term $\mathit{Ownership}(x, y)$ to each of the legal consequences $C_1, ..., C_n$. The argument obviously does not apply to all cases: one implication $(F_1 \vee ... \vee F_p) \rightarrow (C_1 \wedge ... \wedge C_n)$ may often be sufficient to represent the case that a variety of facts $F_1, ..., F_p$ has the same multitude of legal consequences $C_1, ..., C_n$. However, things may be different when norms that link a number of factual descriptions to the same legal consequences stem from different normative sources, may come into conflict with other norms, can be overridden by norms of higher priority, or be subject of individual exemption by norms that grant freedoms or licenses: in these cases, the norms must be represented individually. So it seems worthwhile to consider ways to incorporate intermediate concepts into a formal semantics for deontic logic.

In an input/output framework, a first step could be to employ a separate set T of theoretical terms, namely meaning postulates, alongside the set G of norms. Let T consists of intermediates of the form (a, x), where a is a factual sentence (e.g. that β is in possession of γ, and that α and β agreed that α should have γ, and that β hands γ to α), and x states that some legal term obtains (e.g. that α is now owner of γ). To derive outputs from the set of norms G, one may then use $A \cup \mathit{out}(T, A)$ as input, i.e. the factual descriptions together with the legal statements that obtain given the intermediates T and the facts A.

It may be of particular interest to see that such a set of intermediates may help resolve possible conflicts in the law. Let $(\top, \neg \mathit{dog})$ be a statute that forbids dogs on the premises, but let there also be a higher order principle that no blind person may be required to give up his or her guide dog. Of course the conflict may be solved by modifying the statute (e.g. add a condition that the dog in question is not a guide dog), but then modifying a statute is usually not something a judge, faced with such a norm, is allowed to do: the judge's duty is solely to consider the statute, interpret it according to the known or supposed will of the norm-giver, and apply it to the given facts. The judge may then come to the conclusion that a fair and considerate norm-giver would not have meant the statute to apply to guide dogs, i.e. the term "dog" in the statute is a theoretical term whose extension is smaller than the natural term. So the statute must be re-interpreted as reading $(\top, \neg tdog)$ with the additional intermediate $(dog \wedge \neg guidedog, tdog) \in T$, and thus no conflict arises for the case of blind persons that want to keep their guide dog. While this seems to be a rather natural view of how judicial conflict resolution works (the example is taken from an actual court case), the exact process of creating and modifying theoretical terms in order to resolve conflicts must be left to further study.

8 Constitutive norms

Constitutive norms like counts-as conditionals are rules that create the possibility of or define an activity. For example, according to Searle [50], the activity of playing chess is constituted by action in accordance with these rules. Chess has no existence apart from these rules. The institutions of marriage, money, and promising are like the institutions of baseball and chess in that they are systems of such constitutive rules or conventions. They have been identified as the key mechanism to normative reasoning in dynamic and uncertain environments, for example to realize agent communication, electronic contracting, dynamics of organizations, see, e.g., [9].

Problem 8. How to define counts-as conditionals and relate them to obligations and permissions?

For Jones and Sergot [29], the counts-as relation expresses the fact that a state of affairs or an action of an agent "is a sufficient condition to guarantee that the institution creates some (usually normative) state of affairs". They formalize this introducing a conditional connective $\Rightarrow_s$ to express the "counts-as" connection holding in the context of an institution s. They characterize the logic of $\Rightarrow_s$ as a conditional logic, with axioms for agglomeration $((x \Rightarrow_s y) \,\&\, (x \Rightarrow_s z)) \supset (x \Rightarrow_s (y \wedge z))$, left disjunction $((x \Rightarrow_s z) \,\&\, (y \Rightarrow_s z)) \supset ((x \vee y) \Rightarrow_s z)$ and transitivity $((x \Rightarrow_s y) \,\&\, (y \Rightarrow_s z)) \supset (x \Rightarrow_s z)$. The flat fragment can be phrased as an input/output logic as follows [7].

Definition 1. *Let L be a propositional action logic with $\vdash$ the related notion of derivability and Cn the related consequence operation $Cn(x) = \{y \mid x \vdash y\}$. Let CA be a set of pairs of L, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, read as '$x_1$ counts as y_1', etc.*

Moreover, consider the following proof rules conjunction for the output (AND), disjunction of the input (OR), and transitivity (T) defined as follows:

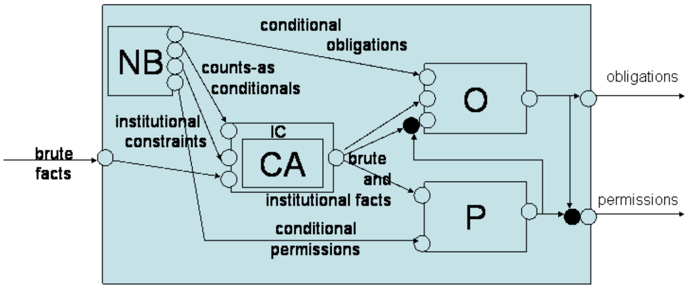
$$\frac{(x,y_1),(x,y_2)}{(x,y_1\wedge y_2)}AND \qquad \frac{(x_1,y),(x_2,y)}{(x_1\vee x_2,y)}OR \qquad \frac{(x,y_1),(y_1,y_2)}{(x,y_2)}T$$

For an institution s, the counts-as output operator out_{CA} is defined as closure operator on the set CA using the rules above, together with a silent rule that allows replacement of logical equivalents in input and output. We write $(x,y) \in out_{\mathrm{CA}}(CA,s)$. Moreover, for $X \subseteq L$, we write $y \in out_{\mathrm{CA}}(CA,s,X)$ if there is a finite $X' \subseteq X$ such that $(\wedge X', y) \in out_{\mathrm{CA}}(CA,s)$, indicating that the output y is derived by the output operator for the input X, given the counts-as conditionals CA of institution s. We also write $out_{\mathrm{CA}}(CA,s,x)$ for $out_{\mathrm{CA}}(CA,s,\{x\})$.

Example 1. If for some institution s we have $CA = \{(a,x),(x,y)\}$, then we have $out_{CA}(CA,s,a) = \{x,y\}$.

There is presently no consensus on the logic of counts-as conditionals, probably due to the fact that the concept is not studied in depth yet. For example, the adoption of the transitivity rule T for their logic is criticized by Artosi *et al.* [5]. Jones and Sergot say that “we have been unable to produce any counter-instances [of transitivity], and we are inclined to accept it”.[7]

The main issue in defining constitutive norms like counts-as conditionals is defining their relation with regulative norms like obligations and permissions. Boella and van der Torre [7] use the notion of a logical architecture combining several logics into a more complex logical system, also called logical input/output nets (or lions).



The notion of logical architecture naturally extends the input/output logic framework, since each input/output logic can be seen as the description of a ‘black box’. In the above figure there are boxes for counts-as conditionals (CA), institutional constraints (IC), obligating norms (O) and explicit permissions (P). The norm base (NB) component contains sets of norms or rules, which are used in the other components to generate the component’s output from its input. The

[7] Neither of these authors considers replacing transitivity by cumulative transitivity (CT): $((x \Rightarrow_s y) \,\&\, (x \wedge y \Rightarrow_s z)) \supset (x \Rightarrow_s z)$, that characterizes operations out_3, out_4 of input/output logic.

figure shows that the counts-as conditionals are combined with the obligations and permissions using iteration, that is, the counts-as conditionals produce institutional facts, which are input for the norms. Roughly, if we write $out(CA, G, A)$ for the output of counts-as conditionals together with obligations, $out(G, A)$ for obligations as before, then $out(CA, G, A) = out(G, out_{CA}(CA, A))$.

There are many open issues concerning constitutive norms, since their logical analysis has not attracted much attention yet. How to distinguish among various kinds of constitutive norms? How are constitutive norms (x counts as y) distinguished from classifications (x is a y)? What is the relation with intermediate concepts?

9 Revision of a set of norms

In general, a code G of regulations is not static, but changes over time. For example, a legislative body may want to introduce new norms or to eliminate some existing ones. A different (but related) type of change is the one induced by the fusion of two (or more) codes as it is addressed in the next section.

Little work exists on the logic of the revision of a set of norms. To the best of our knowledge, Alchourrón and Makinson were the first to study the changes of a legal code [2,3]. The addition of a new norm n causes an enlargement of the code, consisting of the new norm plus all the regulations that can be derived from n. Alchourrón and Makinson distinguish two other types of change. When the new norm is incoherent with the existing ones, we have an *amendment* of the code: in order to coherently add the new regulation, we need to reject those norms that conflict with n. Finally, *derogation* is the elimination of a norm n together with whatever part of G implies n.

In [2] a "hierarchy of regulations" is assumed. Few years earlier, Alchourrón and Bulygin [1] already considered the *Normenordnung* and the consequences of gaps in this ordering. For example, in jurisprudence the existence of precedents is an established method to determine the ordering among norms.

However, although Alchourrón and Makinson aim at defining change operators for a set of norms of some legal system, the only condition they impose on G is that it is a non-empty and finite set of propositions. In other words, a norm x is taken to be simply a formula in propositional logic. Thus, they suggest that "the same concepts and techniques may be taken up in other areas, wherever problems akin to inconsistency and derogation arise" ([2], p. 147).

This explains how their work (together with Gärdenfors' analysis of counterfactuals) could ground that research area that is now known as *belief revision*. Belief revision is the formal studies of how a set of propositions changes in view of a new information that may cause an inconsistency with the existing beliefs. Expansion, revision and contraction are the three belief change operations that Alchourrón, Gärdenfors and Makinson identified in their approach (called AGM) and that have a clear correspondence with the changes on a system of norms we mentioned above. Hence, the following question needs to be addressed:

Problem 9. How to revise a set of regulations or obligations? Does belief revision offer a satisfactory framework for norms revision?

Some of the AGM axioms seem to be rational requirements in a legal context, whereas they have been criticized when imposed on belief change operators. An example is the *success* postulate, requiring that a new input must always be accepted in the belief set. It is reasonable to impose such a requirement when we wish to enforce a new norm or obligation. However, it gives rise to irrational behaviors when imposed to a belief set, as observed for instance in [16].

On the other hand, when we turn to a proper representation of norms, like in the input/output logic framework, the AGM principles prove to be too general to deal with the revision of a normative system. For example, one difference between revising a set of propositions and revising a set of regulations is the following: when a new norm is added, coherence may be restored modifying some of the existing norms, not necessarily retracting some of them. The following example will clarify this point:

Example. If we have $\{(\top, a), (a, b)\}$ and we have that c is an exception to the obligation to do b, then we need to retract (c, b). Two possible solutions are $\{(\neg c, a), (a, b)\}$ or $\{(\top, a), (a \wedge \neg c, b)\}$.

Future research must investigate whether general patterns in the revision of norms exist and how to formalize them.

10 Merging sets of norms

In the previous section we have seen that the change over time of a system of norms raises questions that cannot be properly answered within the belief revision framework. We now want to turn to another type of change, that is the aggregation of regulations. This problem has been only recently addressed in the literature and therefore the findings are still very partial.

The first noticeable thing is the lack of general agreement about where the norms that are to be aggregated come from:

1. some works focus on the merging of conflicting norms that belong to the same normative system [14];
2. other works assume that the regulations to be fused belong to different systems [11]; and finally
3. some authors provide patterns of possible rules to be combined, and consider both cases (1) and (2) above [18].

The first situation seems to be more a matter of coherence of the whole system rather than a genuine problem of fusion of norms. However, such approaches have the merit to reveal the tight connections between fusion of norms, non-monotonic logics and defeasible deontic reasoning. The initial motivation for the study of belief revision was the ambition to model the revision of a set of regulations. On the contrary, the generalization of belief revision to *belief merging* is exclusively dictated by the goal to tackle the problem — arising in computer science — of combining information from different sources. The pieces of information are represented in a formal language and the aim is to merge them in an (ideally) unique knowledge base.[8]

Problem 10. Can the belief merging framework deal with the problem of merging sets of norms?

If (following Alchourrón and Makinson) we assume that norms are unconditional, then we could expect to use standard merging operators to fuse sets of norms. Yet, not only once we consider conditional norms, as in the input/output logic framework, problems arise again. But also, most of the fusion procedures proposed in the literature seem to be inadequate for the scope.

To see why this is the case, we need to explain the merging approach in few words. Let us assume that we have a finite number of belief bases $K_1, K_2, \ldots, K_n$ to merge. IC is the belief base whose elements are the integrity constraints (i.e., any condition that we want the final outcome to satisfy). Given a multi-set $E = \{K_1, K_2, \ldots, K_n\}$ and IC, a merging operator $\mathcal{F}$ is a function that assigns a belief base to E and IC. Let $\mathcal{F}_{IC}(E)$ be the resulting collective base from the IC fusion on E.

Fusion operators come in two types: model-based and syntax-based. The idea of a model-based fusion operator is that models of $\mathcal{F}_{IC}(E)$ are models of IC, which are preferred according to some criterion depending on E. Usually the preference information takes the form of a total pre-order on the interpretations induced by a notion of distance $d(w, E)$ between an interpretation w and E.

Syntax-based merging operators are usually based on the selection of some consistent subsets of E [6,34]. The bases K_i in E can be inconsistent and the result does not depend on the distribution of the wffs over the members of the group.[9]

Finally, the model-based aggregation operators for bases of equally reliable sources can be of two sorts. On the one hand, there are majoritarian operators that are based on a principle of distance-minimization [38]. On the other hand, there are egalitarian operators, which look at the distribution of the distances in E [33]. These two types of merging try to capture two intuitions that often guide the aggregation of individual preferences into a social one. One option is to let the majority decide the collective outcome, and the other possibility is to equally distribute the individual dissatisfaction.

Obviously, these intuitions may well serve in the aggregation of individual knowledge bases or individual preferences, but have nothing to say when we try to model the fusion of sets of norms. Hence, for this purpose, syntactic merging operators may be more appealing. Nevertheless, the selection of a coherent subset

[8] See [35] for a survey on logic-based approaches to information fusion.

[9] [34] refers the term 'combination' to the syntax-based fusion operators to distinguish them from the model-based approaches.

depends on additional information like an order of priority over the norms to be merged, or some other meta-principles.

As the application of belief merging to the aggregation of sets of norms turned out to be unfeasible, an alternative approach is to generalize existing belief change operators to merging rules. This is the approach followed in [11], where merging operators defined using a consolidation operation and possibilistic logic are applied to the aggregation of conditional norms in an input/output logic framework. However, at this preliminary stage, it is not clear whether such methodology is more fruitful for testing the flexibility of existing operators to tackle other problems than the ones they were created for, or if this approach can really shed some light to the new riddle at hand.

A different perspective is taken in [18]. Here, real examples from the Belgian-French bilateral agreement preventing double taxation are considered. These are fitted into a taxonomy of the most common legal rules with exceptions, and the combination of each pair of norms is analyzed. Moreover, both the situations in which the regulations come from the same system and those in which they come from different ones are contemplated, and some general principles are derived. Finally, a merging operator for rules with abnormality propositions is proposed. A limit of Grégoire's proposal is that only the aggregation of rules with the same consequence is taken into account and, in our opinion, this neglects other sorts of conflicts that may arise, as we see now.

The call for non-monotonic reasoning in the treatment of contradictions is also in Cholvy and Cuppens' [14]. A logic to reason when several contradictory norms are merged is presented. The proposal assumes an order of priority among the norms to be merged and this order is also the way to solve the incoherence. Even though this is quite a strong assumption, Cholvy and Cuppens' work take into consideration a broader type of incoherence than in [18]. In their example, an organization that works with secret documents has two rules. R_1 is "It is obligatory that any document containing some secret information is kept in a safe, when nobody is using this document". R_2 is "If nobody has used a given document for five years, then it is obligatory to destroy this document by burning it". As they observe, in order to deduce that the two rules are conflicting, we need to introduce the constraint that keeping a document and destroying it are contradictory actions. That is, the notion of coherence between norms can involve information that are not norms.

11 Conclusion: Deontic logic in context

In this paper we discussed problems of deontic logic that should be considered open and how input/output logic may be useful for analyzing these problems and finding fresh solutions. Jørgensen's dilemma might be overcome by distinguishing operations with norms, like the output $out(G, A)$ of a set of norms G under conditions A, from truth definitions that define what ought to obtain or be done given these norms and conditions. Coherence of a set of norms might be defined with respect to output under constraints, meaning that the set of norms should not generate output for certain conditions that is inconsistent with these constraints. Normative conflicts may be overcome by considering coherent subsets of norms and their output, or such subsets that are preferred given a priority ordering of the norms. Likewise, contrary-to-duty obligations, that obtain in conditions that represent violations, may be modeled by considering only output that is consistent with the input, i.e. the given conditions. Input/output logic provides two possible definitions of dyadic deontic operators, which reconstruct past discussions on whether such operators should be defeasible (in particular in contrary-to-duty conditions), or support strengthening of the antecedent that derives $O(x/a \wedge b)$ from $O(x/a)$. Input/output logic may take into account not just sets of obligating norms, but also explicit permissions, and thus helps shed light on the distinction between weak (negative) permission, where something is permitted if it does not conflict with the norms, and strong (positive) permission which requires an explicit license by the norm-givers. Meaning postulates and intermediate terms, common in legal reasoning but largely ignored by traditional deontic literature, can be taken into account by considering generators T that link natural facts to theoretical terms occurring in the norms, and for counts-as conditionals we may use a separate set of generators (normative institutions) that models how norms are created given an input of natural facts. Finally the questions of how to revise and merge given sets of norms may be approached by preparing the generators (norms) with the aid of standard revision and merging operators.

Lately, normative systems and deontic logic have received widespread attention in multiagent systems and artificial intelligence. A normative multiagent system is "a multiagent system together with normative systems in which agents can decide whether to follow the explicitly represented norms, and the normative systems specify how and in which extent the agents can modify the norms" [10]. Deontic logic, that attempts to formalize the normative consequences given a set of norms and a given situation, can be a helpful tool for devising such systems. In such a general setting, a setting of 'deontic logic in context', many new problems arise: how do deontic truths feature in agent planning and decision making? how do they interact with agent desires, goals, preferences and intentions? how do they feature in communication? how do we model the change of obligations over time, when agents violate or discharge their obligations, when the underlying norms are modified or retracted or when new norms come into existence? The clarification and solution of the problems outlined above, and others, may serve as a first step to make deontic logic fit to become a working component in such a larger setting.

Acknowledgments

The authors wish to thank David Makinson for helpful discussions in the preparation of this paper.

References

1. Alchourrón, C. E. and Bulygin, E., "The Expressive Conception of Norms", in [27] 95–124.
2. Alchourrón, C. E. and Makinson, D., "Hierarchies of Regulations and Their Logic", in [27] 125–148.
3. Alchourrón, C. E. and Makinson, D., "On the Logic of Theory Change: Contraction Functions and Their Associated Revision Functions", *Theoria*, **48**, 1982, 14–37.
4. Anderson, A. R., "On the Logic of Commitment", *Philosophical Studies*, **19**, 1959, 23–27.
5. Artosi, A., Rotolo, A. and Vida, S., "On the logical nature of count-as conditionals", in: *Procs. of LEA 2004 Workshop*, 2004.
6. Baral, C., Kraus, S., Minker, J. and Subrahmanian, V. S., "Combining knowledge bases consisting of first-order theories", *Computational Intelligence*, **8**, 1992, 45–71.
7. Boella, G. and van der Torre, L., "A Logical Architecture of a Normative System", in [17] 24–35.
8. Boella, G. and van der Torre, L., "Permissions and Obligations in Hierarchical Normative Systems", in: *Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003, June 24-28, Edinburgh, Scotland, UK*, ACM, 2003, revised version to appear in *Artificial Intelligence and Law*.
9. Boella, G. and van der Torre, L., "Constitutive Norms in the Design of Normative Multiagent Systems", in: *Computational Logic in Multi-Agent Systems, 6th International Workshop, CLIMA VI*, LNCS 3900, Springer, 2006, 303–319.
10. Boella, G., van der Torre, L. and Verhagen, H., "Introduction to normative multiagent systems", *Computational and Mathematical Organization Theory*, **12(2-3)**, 2006, 71–79.
11. Booth, R., Kaci, S. and van der Torre, L., "Merging Rules: Preliminary Version", in: *Proceedings of the NMR'06*, 2006.
12. Bulygin, E., "Permissive Norms and Normative Concepts", in: Martino, A. A. and Socci Natali, F. (eds.), *Automated Analysis of Legal Texts*, Amsterdam: North Holland, 1986, 211–218.
13. Chisholm, R., "Contrary-to-duty imperatives and deontic logic", *Analysis*, **24**, 1963, 3336.
14. Cholvy, L. and Cuppens, F., "Reasoning about Norms Provided by Conflicting Regulations", in [46] 247–264.
15. van Fraassen, B., "Values and the Heart's Command", *Journal of Philosophy*, **70**, 1973, 5–19.
16. Gabbay, D., Pigozzi, G. and Woods, J., "Controlled Revision — An algorithmic approach for belief revision", *Journal of Logic and Computation*, **13**, 2003, 3–22.
17. Goble, L. and Meyer, J.-J. C. (eds.), *Deontic Logic and Artificial Normative Systems. 8th International Workshop on Deontic Logic in Computer Scicence, DEON 2006, Utrecht, July 2006, Proceedings*, Berlin: Springer, 2006.
18. Grégoire, E., "Fusing legal knowledge", in: *Proceedings of the 2004 IEEE INt. Conf. on Information Reuse and Integration (IEEE-IRI'2004)*, 2004, 522–529.
19. Hansen, J., "Sets, Sentences, and Some Logics about Imperatives", *Fundamenta Informaticae*, **48**, 2001, 205–226.
20. Hansen, J., "Problems and Results for Logics about Imperatives", *Journal of Applied Logic*, **2**, 2004, 39–61.
21. Hansen, J., "Conflicting Imperatives and Dyadic Deontic Logic", *Journal of Applied Logic*, **3**, 2005, 484–511.

22. Hansen, J., "Deontic Logics for Prioritized Imperatives", *Artificial Intelligence and Law*, *forthcoming*, 2005.
23. Hansen, J., "Prioritized Conditional Imperatives: Problems and a New Proposal", *Autonomous Agents and Multi-Agent Systems*, 2007, submitted.
24. Hansson, B., "An Analysis of Some Deontic Logics", *Nôus*, **3**, 1969, 373–398, reprinted in [26] 121–147.
25. Hare, R. M., *Moral Thinking*, Oxford: Clarendon Press, 1981.
26. Hilpinen, R. (ed.), *Deontic Logic: Introductory and Systematic Readings*, Dordrecht: Reidel, 1971.
27. Hilpinen, R. (ed.), *New Studies in Deontic Logic*, Dordrecht: Reidel, 1981.
28. Horty, J. F., "Nonmonotonic Foundations for Deontic Logic", in: Nute, D. (ed.), *Defeasible Deontic Logic*, Dordrecht: Kluwer, 1997, 17–44.
29. Jones, A. and Sergot, M., "A formal characterisation of institutionalised power", *Journal of IGPL*, **3**, 1996, 427443.
30. Jørgensen, J., "Imperatives and Logic", *Erkenntnis*, **7**, 1938, 288–296.
31. Kamp, H., "Free Choice Permission", *Proceedings of the Aristotelian Society*, **74**, 1973/74, 57–74.
32. Kanger, S., "New Foundations for Ethical Theory: Part 1", duplic., 42 p., 1957, reprinted in [26] 36–58.
33. Konieczny, S., *Sur la Logique du Changement: Révision et Fusion de Bases de Connaissance*, Ph.D. thesis, University of Lille, France, 1999.
34. Konieczny, S., "On the difference between merging knowledge bases and combinig them", in: *Proceedings of KR'00*, vol. 8, Morgan Kaufmann, 2000, 135–144.
35. Konieczny, S. and Grégoire, E., "Logic-based approaches to information fusion", *Information Fusion*, **7**, 2006, 4–18.
36. Lewis, D., *Counterfactuals*, Oxford: Basil Blackwell, 1973.
37. Lewis, D., "Semantic Analyses for Dyadic Deontic Logic", in: Stenlund, S. (ed.), *Logical Theory and Semantic Analysis*, Dordrecht: Reidel, 1974, 1 – 14.
38. Lin, J. and Mendelzon, A., "Merging databases under constraints", *International Journal of Cooperative Information Systems*, **7**, 1996, 55–76.
39. Lindahl, L., "Norms, Meaning Postulates, and Legal Predicates", in: Garzón Valdés, E. (ed.), *Normative Systems in Legal and Moral Theory. Festschrift for Carlos E. Alchourrón and Eugenio Bulygin*, Berlin: Duncker & Humblot, 1997, 293–307.
40. Lindahl, L. and Odelstad, J., "Intermediate Concepts in Normative Systems", in [17] 187–200.
41. Makinson, D., "On a Fundamental Problem of Deontic Logic", in [46], 29–53.
42. Makinson, D. and van der Torre, L., "Input/Output Logics", *Journal of Philosophical Logic*, **29**, 2000, 383–408.
43. Makinson, D. and van der Torre, L., "Constraints for Input/Output Logics", *Journal of Philosophical Logic*, **30**, 2001, 155–185.
44. Makinson, D. and van der Torre, L., "Permissions from an Input/Output Perspective", *Journal of Philosophical Logic*, **32**, 2003, 391–416.
45. Makinson, D. and van der Torre, L., "What is Input/Output Logic", in: Löwe, B., Malzkom, W. and Räsch, T. (eds.), *Foundations of the Formal Sciences II : Applications of Mathematical Logic in Philosophy and Linguistics (Papers of a conference held in Bonn, November 10-13, 2000)*, Trends in Logic, vol. 17, Dordrecht: Kluwer, 2003, 163–174, reprinted in this volume.
46. McNamara, P. and Prakken, H. (eds.), *Norms, Logics and Information Systems*, Amsterdam: IOS, 1999.

47. Niiniluoto, I., "Hypothetical Imperatives and Conditional Obligation", *Synthese*, **66**, 1986, 111–133.
48. Ross, A., "Imperatives and Logic", *Theoria*, **7**, 1941, 53–71, Reprinted in *Philosophy of Science* **11**:30–46, 1944.
49. Ross, W. D., *The Right and the Good*, Oxford: Clarendon Press, 1930.
50. Searle, J., *Speech Acts: an Essay in the Philosophy of Language*, Cambridge (UK): Cambridge University Press, 1969.
51. Smiley, T. J., "The Logical Basis of Ethics", *Acta Philosophica Fennica*, **16**, 1963, 237–246.
52. Spohn, W., "An Analysis of Hansson's Dyadic Deontic Logic", *Journal of Philosophical Logic*, **4**, 1975, 237–252.
53. Stenius, E., "The Principles of a Logic of Normative Systems", *Acta Philosophica Fennica*, **16**, 1963, 247–260.
54. von Wright, G., *Logical Studies*, London: Routledge and Kegan, 1957.
55. von Wright, G. H., "Deontic Logic", *Mind*, **60**, 1951, 1–15.
56. von Wright, G. H., "A Note on Deontic Logic and Derived Obligation", *Mind*, **65**, 1956, 507–509.
57. von Wright, G. H., "A New System of Deontic Logic", *Danish Yearbook of Philosophy*, **1**, 1961, 173–182, reprinted in [26] 105–115.
58. von Wright, G. H., "A Correction to a New System of Deontic Logic", *Danish Yearbook of Philosophy*, **2**, 1962, 103–107, reprinted in [26] 115–119.
59. von Wright, G. H., *Norm and Action*, London: Routledge & Kegan Paul, 1963.
60. von Wright, G. H., *An Essay in Deontic Logic and the General Theory of Action*, Amsterdam: North Holland, 1968.
61. von Wright, G. H., "Norms, Truth and Logic", in: von Wright, G. H. (ed.), *Practical Reason: Philosophical Papers vol. I*, Oxford: Blackwell, 1983, 130–209.
62. Ziemba, Z., "Deontic Syllogistics", *Studia Logica*, **28**, 1971, 139–159.

Conflicting Imperatives and Dyadic Deontic Logic*

Jörg Hansen

Institut für Philosophie
Universität Leipzig
Beethovenstraße 15, D-04107 Leipzig
jhansen@uni-leipzig.de

Abstract. Often a set of imperatives or norms seems satisfiable from the outset, but conflicts arise when ways to fulfill all are ruled out by unfortunate circumstances. Semantic methods to handle normative conflicts were devised by B. van Fraassen and J. F. Horty, but these are not sensitive to circumstances. The present paper extends these resolution mechanisms to circumstantial inputs, defines dyadic deontic operators accordingly, and provides a sound and (weakly) complete axiomatic system for such deontic semantics.

1 The Question of Normative Conflicts

Do moral conflicts exist? The orthodox belief in the 1950's was that such conflicts only exist at first glance – the seemingly conflicting obligations arising from the application of merely incomplete principles. Instead, what is actually obligatory must be determined by a careful moral deliberation that involves considering and weighing all relevant facts and reasons, and cannot produce conflicting outcomes. Among the first that came to reject this view were E. J. Lemmon [27] and B. Williams [43]: Lemmon observed that in cases of true moral dilemma, one does not know the very facts needed to determine which obligation might outweigh the other. Williams argued *in reductio* that if, in case of conflicting oughts, there is just one thing one 'actually' ought to do, then feelings of regret about having not acted as one should have are out of place and one should not mind getting into similar situations again. To avoid the derivation of the ought of a contradiction from two oughts of equal weight but with contradictory contents, Williams argued that deontic logic should give up the agglomeration principle

(C) $OA \wedge OB \rightarrow O(A \wedge B)$.

Lemmon had no such qualms: he advocated dropping the Kantian Principle 'ought implies can'

(KP) $OA \rightarrow \Diamond A$,

thus allowing for obligations to bring about the impossible, and concluded:

[*] I thank David Makinson for inspiring remarks, and Lou Goble and Leon van der Torre for comments on an earlier version presented at ΔEON 2004. The paper is part of a project begun in [16] to relate deontic logic to reasoning about imperatives.

> "I should like to see a proper discussion of the arguments that go to resolve moral dilemmas, because I do not believe that this is an area of total irrationality, though I do not believe that a traditional logical approach (the logic of imperatives, deontic logic, and whatnot) will do either."

Regarding commands and legal norms, G.H. von Wright ([45] ch.7), like H. Kelsen ([23] p.211) at the time, excluded the coexistence of conflicting norms from the same source: The giving of two conflicting norms is the expression of an irrational will; it is a performative self-contradiction and as such a pure fact that fails to create a norm. E. Stenius [40] and later C.E. Alchourrón and E. Bulygin [1] rejected this view: A system of norms that is impossible to obey might be unreasonable and its norm-giver blameworthy, but its existence does not constitute a logical contradiction – conflicts are ubiquitous in systems of positive law and logic cannot deny this fact. In his later theory, von Wright [49] concedes that existing normative systems may or may not be contradiction-free, and reformulates deontic principles as meta-norms for consistent norm-giving. Kelsen [24] later came to view logic as inapplicable to law.

2 Van Fraassen's Proposal and Horty's Variation

2.1 Van Fraassen's Operator O^F

Not taking sides, *pro* or *contra* the existence of genuine normative conflicts, but arguing that the view in favor seems at least tenable, B. van Fraassen [11] took up the burden of finding plausible logical semantics that could accommodate conflicting obligations. The intended semantics should accept the possible truth of two deontic sentences OA, $O\neg A$ without committing the norm-subject to the absurd by making $O(A \wedge \neg A)$ true, for van Fraassen wanted to keep the Kantian Principle. Given the existence of certain imperatives in force, i.e. imperatives that are left as valid, relevant, not overridden etc. by some unspecified deliberation process, van Fraassen's idea was to make these imperatives part of the logical model, and to describe something as obligatory if it serves to satisfy some, not necessarily all, imperatives. Formally, let $\mathscr{I}$ be the set of imperatives in force, $\mathbf{B}$ be the set of possible states of affairs, and $i^+ \subseteq \mathbf{B}$ be the possible states of affairs where the imperative $i \in \mathscr{I}$ is considered fulfilled. Let $\|A\| \subseteq \mathbf{B}$ be the set of possible states of affairs where the indicative sentence A is considered true. Finally, let $\mathit{score}(v)$ be the set of all imperatives that are fulfilled in the state of affairs v: $\mathit{score}(v) = \{i \in \mathscr{I} \mid v \in i^+\}$. Van Fraassen then defines:

$$\text{[Df-F]} \quad O^F A \text{ is true} \quad \text{iff} \quad \exists v \in \|A\| : \forall v' \in \|\neg A\| : \mathit{score}(v) \nsubseteq \mathit{score}(v')$$

So A is obligatory if and only if (iff) there is some score that can be achieved when A is true, which is not included in any score that could be achieved when $\neg A$ is true. In other words, A is obligatory iff there are imperatives that can only be (collectively) satisfied when A is true, but not when A is false.

By slightly changing the viewpoint, van Fraassen's proposal might also be described in the following way: let I be a set not of imperatives, but of indicative sentences in the language $\mathscr{L}_{BL}$ of some basic logic *BL*. The motivation is that I contains one sentence A for each imperative i in force that is true in exactly those states of affairs in which the imperative is fulfilled, i.e. $\|A\| = i^+$. *BL* is assumed to be compact and the turnstile in $\Gamma \vdash_{BL} A$ means a classical consequence relation that characterizes *BL*, $\Gamma \subseteq \mathscr{L}_{BL}$, $A \in \mathscr{L}_{BL}$. Let the *remainder set* $\Gamma \bot\ A$ be the set of all maximal subsets that do not derive A, i.e. of all $\Gamma' \subseteq \Gamma$ such that (i) $\Gamma' \nvdash_{BL} A$, and (ii) there is no Γ'' such that $\Gamma' \subset \Gamma'' \subseteq \Gamma$ and $\Gamma'' \nvdash A$. Then Df-F is equivalent to Df-F*($\top$ means an arbitrary tautology):[1]

[Df-F*] $O^F A$ is true iff $\exists I' \in I \bot \neg\top : I' \vdash_{BL} A$

So A is obligatory iff it is derivable from a maximally consistent subset of I. So something is obligatory if it is required for doing 'the most': if it is necessitated by a strategy to fulfill so many imperatives that no one who satisfies these as well could satisfy more. While a parallel operation for belief change is known as 'credulous reasoning', to call it 'orthodox' might seem more appropriate: the agent is not released from any of her obligations as long as they are fulfillable, even if this fulfilment is at the expense of violating other norms.

To see how van Fraassen's semantics work, first let $I = \{A, B\}$, where A and B are supposedly contingent and independent. There are no conflicts, I is consistent and $O^F A$, $O^F B$ and $O^F(A \wedge B)$ are all true since I derives A, B and $A \wedge B$. Thus agglomeration of contents is permitted so long as the underlying imperatives do not conflict. For the case of conflict, change I into $\{A\wedge C, B\wedge\neg C\}$; $O^F A$ and $O^F B$ are true since A and B derive from the maximally consistent sets $\{A \wedge C\}$ and $\{B \wedge \neg C\}$, but $O^F(C \wedge \neg C)$ is false since no consistent subset derives $C \wedge \neg C$. The same is true for $O^F(A \wedge B)$ though $\{A \wedge B\}$ is consistent: the truth of $A \wedge B$ is not necessary for maximal norm satisfaction.

An axiomatic system DL^F that is (weakly) complete with regard to van Fraassen's semantics is defined by the following axiom-schemes, in addition to *BL*-instances and *modus ponens* (cf. [17] sec. 5, $\bot$ means an arbitrary contradiction, and the index 'F' here and below indicates that the deontic operators occurring in the axiom scheme are thus indexed):

(M^F) $O^F(A \wedge B) \rightarrow (O^F A \wedge O^F B)$
(P^F) $\neg O^F \bot$
(N^F) $O^F \top$
(Ext^F) If $\vdash_{BL} A \leftrightarrow B$ then $\vdash_{DF} O^F A \leftrightarrow O^F B$

To van Fraassen's own puzzlement, the cases where agglomeration remains permissible do not seem axiomatizable: object language does not reveal whether particular A and B of some $O^F A$ and $O^F B$ are derived from the demands of imperatives that do not conflict and so $O^F(A \wedge B)$ should be supported.[2]

[1] Cf. Horty's [20] Theorem 2.

[2] So agglomeration requires some consistency check of the underlying imperatives' contents. Van der Torre and Tan [41], [42] proposed a two-phase deontic logic, where

2.2 The Skeptical Operator O^S

The invalidation of the agglomeration principle by van Fraassen's semantics did not make them popular (cf. Donagan [8] p. 298). Moreover, let a P^F-operator expressing permission be defined in the usual way as $P^F A =_{def} \neg O^F \neg A$, and consider again $I = \{A \wedge C, B \wedge \neg C\}$: then $O^F(A \wedge C)$ is true, and so is $O^F \neg C$. Applying (M^F) and (Ext^F), $O^F \neg(A \wedge C)$ must be true, hence $P^F(A \wedge C)$ is false. So not even the obligatory is always permitted, which seems strange (cf. Jacquette [22]).

In reaction to the dismissal of the agglomeration principle, Donagan [8] and Brink [7] have claimed that even if there could be a normative demand for A and a conflicting demand for B, with $\vdash_{BL} A \rightarrow \neg B$, it need not follow that the norm-subject has an obligation to realize A *and* an obligation to realize B. Rather, there should just be a disjunctive obligation to realize A *or* B. Given competing normative standards of equal weight, the strategy of this reasoning is not to trust a single standard, but to consider obligatory only what *all* standards demand. Let I be as before. Varying van Fraassen's truth definition, Horty [20] has formalized this 'skeptical' ought as follows:[3]

[Df-S] $O^S A$ is true iff $\forall I' \in I \bot \neg \top : I' \vdash_{BL} A$

So $O^S A$ is true iff A is derivable from all maximally consistent subsets of I. 'skeptical' is the term used in the epistemic-oriented literature, yet 'legalist' also seems fitting, since a norm violation is never pronounced as obligatory even if it is inevitable. This does not let the agent off the hook: by doing what is obligatory in this sense, a maximum of norms will, by necessity, get satisfied.

Let again $I = \{A \wedge C, B \wedge \neg C\}$. $O^S A$ and $O^S B$ are false and $O^S(A \vee B)$ is true: just $A \vee B$, but neither A nor B are derived by both of the two consistent subsets $\{A \wedge C\}$ and $\{B \wedge \neg C\}$. $P^S(A \wedge C)$ is also true: A, C were assumed to be contingent and independent, so the maximally consistent subset $\{A \wedge C\} \subseteq I$ does not derive $\neg(A \wedge C)$. So what is O^F-obligatory is at least P^S-permitted.

A complete axiomatic system DL^S is defined by the axiom-schemes (M^S), (C^S), (P^S), (N^S), and (Ext^S), together with *BL*-instances and *modus ponens*. Since the truth definitions for $O^F A$ and $O^S A$ merely depend on a set I and *BL*, mixed expressions such as $O^F A \wedge \neg O^S A$ are meaningful and may be accepted as well-formed. Then

(C^{SF}) $O^S A \wedge O^F B \rightarrow O^F(A \wedge B)$

is valid, and the mixed system $DL^{\{F,S\}}$ – containing the axiom schemes for DL^F, DL^S, the axiom scheme (C^{SF}), all instances of *BL*-theorems and *modus ponens* – is sound and (weakly) complete (cf. [17] sec. 6).

'consistent aggregation' must take place before weakening. For the present imperative semantics, I suggested a bimodal approach in [17] with an operator O^2 that 'more directly talks about the imperatives'. For comparisons and a new proposal cf. [15].

[3] More in parallel to van Fraassen's original definition, one may equivalently define:

[Df-S*] $O^S A$ is true iff $\forall v \in \|\neg A\| : \exists v' \in \|A\| : \mathit{score}(v) \subset \mathit{score}(v')$

3 Predicaments and Dyadic Deontic Logic

Arguing for the possibility of moral conflicts, R. Barcan Marcus [33] gave the following example:

> "Under the single principle of promise keeping, I might make two promises in all good faith and reason that they will not conflict, but then they do, as a result of circumstances that were unpredictable and beyond my control."

Note that there is no conflict at the outset: any dilemma could have been averted by not promising anything. Moreover, there might have been some point in time at which keeping both promises was possible: having 500 $ with me and another 1000 $ in the office, I promise Raoul and Johnny 500 $ each on Saturday with every intention of paying them on Monday, only to find out that the office had been burglarized over the weekend. Donagan [8] argues that this is not a genuine conflict, because three resolving principles apply: (i) one must not make promises one cannot or must not keep, (ii) all promises are made with the implicit condition that they are void if they cannot or must not be kept, (iii) one must not make promises when one does not believe that the other party has fully understood (ii). But as I well knew beforehand, neither Raoul nor Johnny are going to let me off the hook, regardless of what may happen at the office. According to (iii), I was wrong to make the promises. So am I entitled to break them (both)? – We have here what G.H. von Wright [48] terms a 'predicament': a situation from which there is no permitted way out, but to which there also is no permitted inlet. The normative order is consistent, it is only through one's own fault that one finds oneself in a predicament.[4] Von Wright then asks:

> "The man in a predicament will, of necessity, react in some way or other, either do something or remain passive. Even though every reaction of his will be a sin, is it not reasonable to think that there is yet something he ought to do rather than anything else? To deny this would be to admit that it makes, deontically, no difference what he does. But is this reasonable? (...) If all our choices are between forbidden things, our duty is to choose the least bad thing." ([48] p. 80)

Sub-ideal demands are usually represented by a dyadic deontic sentence $O(A/C)$, meaning that it ought to be that A given C is true. By accepting all instances

[4] That predicaments *only* arise from an agent's own faults, and not through misfortune or the wrongdoings of others, is a view von Wright and Donagan ascribe to Thomas Aquinas, but this does not seem quite correct: In the discussion of oaths (*Summa Theologica* II.II Qu. 89 art. 7 ad 2), Thomas considers the objection that it would sometimes be contrary to virtue, or an obstacle to it, if one were to fulfill what one has sworn to do – so oaths need not always be binding. In answering, Thomas distinguishes oaths that are unlawful from the outset, where a man sinned in swearing, and oaths that could be lawful at the outset but lead to an evil result through some new and unforeseen emergency: fulfillment of such oaths is unlawful.

of (DD-$\top$) $O(A/\top) \rightarrow P(A/\top)$ as a logical truth in [48], von Wright dismisses an inconsistent normative system as 'conceptual absurdity': if A is obligatory on tautological conditions (i.e. unconditionally obligatory), then there cannot be a likewise unconditional obligation to the contrary. Although von Wright originally used the stronger (DD) $O(A/C) \rightarrow P(A/C)$ for arbitrary C (axiom A1 of the 'old system' in [44], and axiom B1 of the 'new system' in [46]), he later turned against it, arguing that while morality makes no conflicting claims, it is not a logical impossibility that conflicting promises can give rise to predicaments.[5] Dyadic operators seem essential for even making this distinction.[6]

Turning object language oughts into a special sort of conditionals does not necessarily imply a change in the formalization of the background imperatives: consider the set $I = \{(\neg C \vee A), (\neg C \vee \neg A)\}$, corresponding to background imperatives in the usual way. I is also its single maximally consistent subset, which derives $\neg C$, so $O^F \neg C$ and $O^S \neg C$ are both true. But a single standard is no longer available once C becomes true: the imperatives have not all been fulfilled (otherwise one would not be in condition C), and any maximal set of imperatives that is consistent with the given circumstances cannot contain all. So the proposal is to call A obligatory in case C iff A is necessary for doing 'the most' that can be achieved, given the truth of C. Formally:

[Df-DF] $O^F(A/C)$ iff $\exists I' \in I \bot \neg C : I' \cup \{C\} \vdash_{BL} A$

According to this definition, $O^F(A/C)$ is true iff there is some set, among the maximal subsets of I consistent with C, that together with C derives A. This is obviously a conservative extension of the definition given for the unconditional case, so we may define $O^F A =_{def} O^F(A/\top)$.

If a cautious, disjunctive approach were appropriate for cases of conflict, then it would be hard to see why predicaments should be treated differently: that conflicts must be accounted for at the outset, but analogues of Buridan's ass cannot be brought about by fate or unpredictable human nature, would hardly be plausible. Distrusting any single standard, such an approach would accept, given the circumstances C, only what is necessary by any standard that could still be met – no crying over spilled milk. Formally:

[Df-DS] $O^S(A/C)$ iff $\forall I' \in I \bot \neg C : I' \cup \{C\} \vdash_{BL} A$

According to this definition, $O^S(A/C)$ is true iff all the maximal subsets of I consistent with C derive A, given the truth of C. This is again a conservative extension of the unconditional case, so one may define $O^S A =_{def} O^S(A/\top)$.

After a comparison of the above definitions with similar approaches namely in the study of nonmonotonic reasoning, I will give an axiomatic dyadic deontic system $DDL^{\{F,S\}}$, which I prove to be sound and (only) weakly complete with respect to the above semantics.

[5] Cf. [47], [48] pp. 36, 81, 89

[6] Below, I extend the treatment of conflicts to the area of predicaments, and do not follow von Wright in ruling out conflicts. However, this can easily be done by axiomatically adding (DD-$\top$) to the system presented below.

4 Comparisons

Though the truth definitions introduced in the preceding section naturally extend the proposals of van Fraassen and Horty for dealing with normative conflicts to the dyadic context and the related problem of predicaments, and though their resolution mechanisms are not exactly new (cf. below), there has not been much discussion of these concepts in the deontic logic literature. Notably, Horty's own dyadic operator in [21] is defined with respect to (simply) maximally non-conflicting sets of prima facie oughts, and it is disregarded that their joint demands may now be inconsistent with the situation. However, the more general literature on nonmonotonic reasoning includes a range of parallel concepts.

Regarding O^S, the most obvious parallel is Kratzer and Lewis's premise semantics in [25] and [30] which has a set of formulas H (the premises) to define counterfactuals in much the same way as the set I is used here in the definition of deontic conditionals. Considering Kratzer's definition, and setting aside the world-relativity of H, let $\mathscr{S}_{H,C} = \{H' \subseteq H \mid H' \nvdash_{BL} \neg C\}$ be the set of all subsets of H that are, according to some basic logic BL, consistent with C. Then the counterfactual conditional $\Box\!\!\rightarrow$ is defined in the following way:

$$H \models C \Box\!\!\rightarrow A \quad \text{iff} \quad \forall H' \in \mathscr{S}_{H,C} : \exists H'' \in \mathscr{S}_{H,C} : H' \subseteq H'', H'' \cup \{C\} \vdash_{BL} A$$

In other words, $C \Box\!\!\rightarrow A$ is true iff each set in $\mathscr{S}_{H,C}$ has a superset in $\mathscr{S}_{H,C}$ that implies $C \rightarrow A$.[7] The truth definition is tailored for a basic logic that may fail compactness and so accommodates sets H with ever-larger C-consistent subsets, but no maximal ones. Here, BL was assumed to be compact, and we obtain:

Observation 1 (Relation to premise semantics)
For any set $I \subseteq \mathscr{L}_{PL}$: $I \models O^S(A/C)$ *iff* $I \models C \Box\!\!\rightarrow A$

Proof. Left-to-right: Suppose that $C \Box\!\!\rightarrow A$ is false, so there is some $I' \in \mathscr{S}_{I,C}$: $\forall I'' \in \mathscr{S}_{I,C}$: if $I' \subseteq I''$ then $I'' \cup \{C\} \nvdash_{PL} A$. $I' \nvdash_{PL} \neg C$, so by definition there is some $I'' \in I\bot\neg C$ such that $I' \subseteq I''$. So $I'' \cup \{C\} \nvdash_{PL} A$, so $\neg\forall I' \in I\bot\neg C$: $I' \cup \{C\} \vdash_{PL} A$ and so $O^S(A/C)$ is false. *Right-to-left:* Suppose $O^S(A/C)$ is false, so $\exists I' \in I\bot\neg C : I' \cup \{C\} \nvdash A$. Then $I' \in \mathscr{S}_{I,C}$, and by definition of $I\bot\neg C$ there is no other $I'' \in \mathscr{S}_{I,C} : I' \subseteq I''$, so $\forall I'' \in \mathscr{S}_{I,C}$: if $I' \subseteq I''$ then $I'' \cup \{C\} \nvdash A$, and $C \Box\!\!\rightarrow A$ is false.

Then, the definition of O^S parallels that of a consequence relation associated to a Poole system without constraints [35]: This has two sets Γ, Δ of formulas, the facts and the defaults. A scenario is a set $\Delta' \cup \Gamma$ such that $\Delta' \subseteq \Delta$ and $\Delta' \cup \Gamma \nvdash_{BL} \bot$. A 'maximal scenario' is one where $\Delta' \in E(\Gamma)$, $E(\Gamma)$ being the

[7] Lewis's [30] variation requires this property only of non-empty $H' \in \mathscr{S}_{H,C}$. This corresponds to replacing $I' \cup \{C\} \vdash_{BL} A$ in the truth definition for O^S with $\exists B_1, ..., B_n \in I' :\vdash_{BL} (B_1 \wedge ... \wedge B_n \wedge C) \rightarrow A$ to produce a regular instead of a normal operator. – Lewis notes (p. 233) that for deontic conditionals, the premises of the premise semantics might be understood to be "something that ought to hold", so he is to be credited for the imperative semantics employed here.

set of all $\mathit{\Delta}' \subseteq \mathit{\Delta}$ such that $\mathit{\Delta}' \cup \{\mathit{\Gamma}\} \nvdash_{BL} \bot$, and for all $\mathit{\Delta}'' \subseteq \mathit{\Delta}$, if $\mathit{\Delta}' \subset \mathit{\Delta}''$ then $\mathit{\Delta}'' \cup \{\mathit{\Gamma}\} \vdash_{PL} \bot$. *Cn* being *BL*-consequence, a prediction A from the facts $\mathit{\Gamma}$ and the defaults $\mathit{\Delta}$ is then defined as:

$$\mathit{\Gamma} \mid\!\sim_{skept(\mathit{\Delta})} A \quad =_{def} \quad A \in \bigcap_{\mathit{\Delta}' \in E(\mathit{\Gamma})} Cn(\mathit{\Delta}' \cup \mathit{\Gamma})$$

So A is predicted from $\mathit{\Gamma}$ and $\mathit{\Delta}$ if all maximal scenarios derive A. – Likewise, a ‘credulous prediction’ can be defined as

$$\mathit{\Gamma} \mid\!\sim_{cred(\mathit{\Delta})} A \quad =_{def} \quad A \in \bigcup_{\mathit{\Delta}' \in E(\mathit{\Gamma})} Cn(\mathit{\Delta}' \cup \mathit{\Gamma})$$

(cf. Brass [5] for the analogy and notation). The following is then immediate:

Observation 2 (Relation to Poole systems)
For any set $I \subseteq \mathscr{L}_{BL}$*:* $I \models O^S(A/C)$ *iff* $\{C\} \mid\!\sim_{skept(I)} A$,
$I \models O^F(A/C)$ *iff* $\{C\} \mid\!\sim_{cred(I)} A$.

Regarding the O^F-operator, it is perhaps not quite as obvious that its corresponding P^F-operator is closely related to the ‘X-logics’ of Siegel and Forget [38], [10]: The consequence relation $\mid\!\sim_X$ of these logics holds between a set of formulas $\mathit{\Gamma}$ and a formula A *modulo* a set X of formulas, where the definition is

$$\mathit{\Gamma} \mid\!\sim_X A \quad \textit{iff} \quad Cn(\mathit{\Gamma} \cup \{A\}) \cap X \;=\; Cn(\mathit{\Gamma}) \cap X$$

As Makinson [31] pointed out, X can be understood as a set of ‘bad’ propositions that one is to avoid. So $\mathit{\Gamma} \mid\!\sim_X A$ is true iff A can be realized together with $\mathit{\Gamma}$ without increasing the set of ‘bad’ proposition above those that were already true given $\mathit{\Gamma}$. Here we have a set I of ‘desired’ propositions, so a statement seems ‘bad’ if it asserts that some desired proposition be false, e.g. $\neg A$ is true for some $A \in I$, or that at least one $A_1, ..., A_n \in I$ is false, i.e. $\neg(A_1 \wedge ... \wedge \neg A_n)$ is true. Let $I^{\curlywedge} = \{\neg \bigwedge\{A_1, ..., A_n\} \mid \{A_1, ..., A_n\} \subseteq I, 1 \leq n \leq card(I)\}$ be the ‘bad set’ corresponding to I. We then obtain:

Observation 3 (Relation to X-logics)
For any set $I \subseteq \mathscr{L}_{PL}$, $X = I^{\curlywedge}$*:* $I \models P^F(A/C)$ *iff* $\{C\} \mid\!\sim_X A$.

Proof. Right-to-left: Suppose $\{C\} \not\mid\!\sim_X A$, so $Cn(C \wedge A) \cap I^{\curlywedge} \;\neq\; Cn(C) \cap I^{\curlywedge}$, so by monotony of *Cn* there is a $B^{\curlywedge} \in Cn(C \wedge A) \cap I^{\curlywedge}$ such that $B^{\curlywedge} \notin Cn(C) \cap I^{\curlywedge}$. By definition, $B^{\curlywedge} = \neg(B_1 \wedge ... \wedge B_n)$ for some $B_1, ..., B_n \in I$. The first fact provides $\{C \wedge A\} \vdash_{BL} \neg(B_1 \wedge ... \wedge B_n)$ and by contraposition $\{B_1, ..., B_n\} \vdash_{BL} C \rightarrow \neg A$. From the second fact we obtain $\{C\} \nvdash_{BL} \neg(B_1 \wedge ... \wedge B_n)$, so by contraposition $\{B_1, ..., B_n\} \nvdash_{BL} \neg C$, so $\{b_1, ..., B_n\} \subseteq I'$ for some maxi-consistent $I' \in I \bot \neg C$, and so there is some $I' \in I \bot \neg C : I' \cup \{C\} \vdash_{BL} \neg A$, so $I \models O^F(\neg A/C)$ and by definition $I \nvDash P^F(A/C)$. *Left-to-right:* Suppose $I \nvDash P^F(A/C)$, so there is a $I' \in I \bot \neg C : I' \cup \{C\} \vdash_{BL} \neg A$. By compactness of *BL* there are $\{B_1, ..., B_n\} \subseteq I'$ with $\{B_1, ..., B_n\} \vdash_{BL} C \rightarrow \neg A$, and $\{B_1, ..., B_n\} \nvdash_{BL} \neg C$ since they are in I'. By definition there is a $B^{\curlywedge} \in I^{\curlywedge}$ such that $B^{\curlywedge} = \neg(B_1 \wedge ... \wedge B_n)$. By the first fact $\{C \wedge A\} \vdash_{BL} B^{\curlywedge}$, so $B^{\curlywedge} \in Cn(C \wedge A) \cap I^{\curlywedge}$, and by the second $\{C\} \nvdash_{BL} B^{\curlywedge}$, so $B^{\curlywedge} \notin Cn(C) \cap I^{\curlywedge}$. So $Cn(C \wedge A) \cap I^{\curlywedge} \;\neq\; Cn(C) \cap I^{\curlywedge}$, and $\{C\} \not\mid\!\sim_X A$.

A unified treatment of both, skeptical and credulous consequence can be found in Bochman's 'epistemic states'-semantics in [2], [3], [4]. Epistemic states, equivalent to the cumulative models in [26], are triples $\mathbb{E} = \langle S, \prec, \ell \rangle$, where S is a set of objects (belief states), $\prec$ some asymmetric 'preference' relation on S, and ℓ a labeling function that assigns each state $s \in S$ a deductively closed theory. $min\, S' = \{s \in S' \mid \forall t \in S', t \neq s : t \not\prec s\}$ is the set of minimal states in $S' \subseteq S$. $\langle A \rangle = \{s \in S \mid \neg A \notin \ell(s)\}$ is the set of belief states consistent with A. For each A, $\langle A \rangle$ must be $\prec$-*smooth*, i.e. for any $s \in \langle A \rangle$, either $s \in min\langle A \rangle$ or there is some $t \in min\langle A \rangle$ with $t \prec s$. With BL as basic logic, Bochman's definitions for skeptical and credulous consequence relations $\mathrel{|\!\sim}$ and $\mathrel{|\!\approx}$ are:

$$
\begin{array}{lll}
A \mathrel{|\!\sim}_{\mathbb{E}} B & \textit{iff} & \forall s \in min\langle A \rangle : \ell(s) \vdash_{PL} A \rightarrow B \\
A \mathrel{|\!\approx}_{\mathbb{E}} B & \textit{iff} & \langle A \rangle = \varnothing \text{ or } \exists s \in min\langle A \rangle : \ell(s) \vdash_{PL} A \rightarrow B
\end{array}
$$

Observation 4 (Relation to Bochman's epistemic states)
Let $I \subseteq \mathscr{L}_{PL}$, and let the corresponding 'epistemic state' $\mathbb{E}_I = \langle S, \prec, \ell \rangle$ be such that (i) $S = \mathscr{P}(I)$, (ii) $\ell(s) = Cn(s)$, and (iii) $s \prec t$ iff $t \subsetneq s$. Then

$$
\begin{array}{lll}
I \models O^S(A/C) & \textit{iff} & C \mathrel{|\!\sim}_{\mathbb{E}_I} A, \\
I \models O^F(A/C) & \textit{iff} & \nvdash_{PL} \neg C \textit{ and } C \mathrel{|\!\approx}_{\mathbb{E}_I} A.
\end{array}
$$

Proof. I prove first (a) $I' \in min\langle A \rangle$ iff $I' \in I \bot \neg A$, (b) $\mathbb{E}_I$ is an epistemic state, (c) $\langle A \rangle = \varnothing$ iff $\vdash_{BL} \neg A$: For (a), by definition $\langle A \rangle = \{I' \subseteq I \mid I' \nvdash_{BL} \neg A\}$, so $\langle A \rangle$ is the set of subsets of I consistent with A. $I' \in min\langle A \rangle$ means that for any $I'' \in min\langle A \rangle$, $I' \neq I''$: $I'' \not\prec I'$. By definition for any $I'' \in \langle A \rangle$, $I' \neq I''$: $I' \nsubseteq I''$. This means there is no $I'' \in I$ consistent with A such that $I' \subseteq I''$, which means $I' \in I \bot \neg A$. For (b), if $I' \subseteq I$ is in $\langle A \rangle$, i.e. it is consistent with A, and $I' \notin I \bot \neg A$ then by definition of $I \bot \neg A$ there is some $I'' \in I \bot \neg A$ such that $I' \subset I''$, so there is some $I'' \in min\langle A \rangle$ with $I'' \prec I'$. So $\mathbb{E}_I$ is smooth, hence it is an epistemic state. For (c), $\langle A \rangle = \varnothing$ iff $\{I' \in \mathscr{P}(I) \mid I' \nvdash_{BL} \neg A\} = \varnothing$ iff $\varnothing \vdash_{BL} \neg A$ holds by monotony of BL. – Putting together, we get: $I \models O^S(A/C)$ iff $\forall I' \in I \bot \neg C : I' \cup \{C\} \vdash_{BL} A$ iff $\forall I' \in min\langle C \rangle : I' \vdash_{BL} C \rightarrow A$ iff $C \mathrel{|\!\sim}_{\mathbb{E}_I} A$. Likewise: $I \models O^F(A/C)$ iff $\exists I' \in I \bot \neg C : I' \cup \{C\} \vdash_{BL} A$ iff $\exists I' \in min\langle C \rangle : I' \vdash_{BL} C \rightarrow A$ iff $\langle C \rangle \neq \varnothing$ and [$\langle C \rangle = \varnothing$ or $\exists I' \in min\langle C \rangle : I' \vdash_{BL} C \rightarrow A$] iff $\nvdash_{BL} \neg C$ and $C \mathrel{|\!\approx}_{\mathbb{E}_I} A$.

A final parallel brings us back to deontic logic, namely the multiplex preference semantics of Goble in [12], [13] and [14], where a multitude of preference relations enables definitions like 'all-best' (universally preferred) and 'some-best' (existentially preferred), which are then used in definitions of deontic operators. That, in the finite case, such semantics corresponds closely to the present account will be explicated in sec. 6. Regarding meta-theory, for a somewhat more general semantic setting the skeptical consequence relation was axiomatized by Kraus, Lehmann and Magidor [26], and the credulous consequence relation by Bochman [2]. However, a completeness proof for a system that includes both[8] seems to be missing so far and this is what I shall now turn to.

[8] One might add a third (monadic) deontic modality O^2 that 'more directly talks about the imperatives' to axiomatize consistent agglomeration, but I must leave the details to future study (cf. [17] sec. 6 for the the resulting monadic system $DL^{\{2,F,S\}}$).

5 The Dyadic Deontic Logic $\boldsymbol{DDL^{\{F,S\}}}$

Let the *basic logic* be propositional logic PL: The alphabet has proposition letters $Prop = \{p_1, p_2, ...\}$, operators '$\neg$', '$\wedge$', '$\vee$', '$\rightarrow$', '$\leftrightarrow$' and parentheses '(', ')'. The language $\mathscr{L}_{PL}$ is defined as usual. $\bigwedge$, $\bigvee$ in front of a set of sentences means their conjunction and disjunction, and e.g. $\bigwedge_{i=1}^{n} A_i$ further abbreviates $\bigwedge\{A_i, ..., A_n\}$. Semantically, valuation functions $v : Prop \rightarrow \{1, 0\}$ define the truth of sentences $A \in \mathscr{L}_{PL}$ as usual (written $v \models A$), $\mathbf{B}$ is the set of all such valuations, and $\|A\|$ is the extension $\{v \in \mathbf{B} \mid v \models A\}$ of A. PL is a sound and complete axiomatic system, and $\vdash_{PL} A$ means that A is provable in PL.

The alphabet of the *language* $\mathscr{L}_{DDL^{\{F,S\}}}$ additionally has the operators 'O^F', 'O^S', and the auxiliary '/'. $DDL^{\{F,S\}}$ is then the smallest set such that

a) for all $A, C \in \mathscr{L}_{PL}$, $O^F(A/C)$ and $O^S(A/C) \in DDL^{\{F,S\}}$,

b) if $A, B \in DDL^{\{F,S\}}$, so are $\neg A$, $(A \wedge B)$, $(A \vee B)$, $(A \rightarrow B)$, $(A \leftrightarrow B)$.

Outer parentheses will be mostly omitted. We define $P^*(A/C) =_{def} \neg O^*(\neg A/C)$, where * is F or S. For simplification we do not permit mixed expressions and nested deontic operators like $p_1 \wedge O^S(p_2/p_1)$, $P^S(O^F(p_2/p_2)/p_1)$.

For $DDL^{\{F,S\}}$-*semantics*, the truth of $DDL^{\{F,S\}}$-sentences is defined with respect to a set $I \subseteq \mathscr{L}_{PL}$ (Boolean operators being as usual):

$$\begin{array}{lll} I \models O^F(A/C) & \text{iff} & \exists I' \in I\bot\neg C : \ I' \cup \{C\} \vdash_{PL} A \\ I \models O^S(A/C) & \text{iff} & \forall I' \in I\bot\neg C : \ I' \cup \{C\} \vdash_{PL} A \end{array}$$

If $I \models A$, A is called $DDL^{\{F,S\}}$-*satisfiable*, and $DDL^{\{F,S\}}$-*valid* if $I \models A$ for all $I \subseteq \mathscr{L}_{PL}$ (we write $\models_{DDL^{\{F,S\}}} A$).

Consider the following *axiom-schemes* (* is the uniform index F or S):

(CExt*)	If $\vdash_{PL} C \rightarrow (A \leftrightarrow B)$ then $\vdash_{DDL^{\{F,S\}}} O^*(A/C) \leftrightarrow O^*(B/C)$		
(ExtC*)	If $\vdash_{PL} C \leftrightarrow D$ then $\vdash_{DDL^{\{F,S\}}} O^*(A/C) \leftrightarrow O^*(A/D)$		
(DM*)	$O^*(A \wedge B/C) \rightarrow (O^*(A/C) \wedge O^*(B/C))$		
(DC^S)	$O^S(A/C) \wedge O^S(B/C) \rightarrow O^S(A \wedge B/C)$		
(DC^{SF})	$O^S(A/C) \wedge O^F(B/C) \rightarrow O^F(A \wedge B/C)$		
(DN^S)	$O^S(\top/C)$	(DN-R*)	If $\nvdash_{PL} \neg C$ then $\vdash_{DDL^{\{F,S\}}} O^*(\top/C)$
(DP^F)	$P^F(\top/C)$	(DP-R*)	If $\nvdash_{PL} \neg C$ then $\vdash_{DDL^{\{F,S\}}} P^*(\top/C)$
(Cond*)	$O^*(A/C \wedge D) \rightarrow O^*(D \rightarrow A/C)$		
(CCMon*)	$O^*(A \wedge D/C) \rightarrow O^*(A/C \wedge D)$		
(RMon^F)	$P^F(D/C) \rightarrow (O^F(A/C) \rightarrow O^F(A/C \wedge D))$		
(RMon^{FSS})	$P^F(D/C) \rightarrow (O^S(A/C) \rightarrow O^S(A/C \wedge D))$		
(RMon^{SSF})	$P^S(D/C) \rightarrow (O^S(A/C) \rightarrow O^F(A/C \wedge D))$		

The *system* $DDL^{\{F,S\}}$ is then the set such that (i) all $\mathscr{L}_{DDL^{\{F,S\}}}$-instances of PL-tautologies are in $DDL^{\{F,S\}}$, (ii) all $\mathscr{L}_{PL}$-instances of the above axiom schemes are in $DDL^{\{F,S\}}$, and (iii) $DDL^{\{F,S\}}$ is closed under *modus ponens*. If $A \in DDL^{\{F,S\}}$ we write $\vdash_{DDL^{\{F,S\}}} A$ and call A *provable* in $DDL^{\{F,S\}}$. $\Gamma \subseteq \mathscr{L}_{DDL^{\{F,S\}}}$ is $DDL^{\{F,S\}}$-*inconsistent* iff there are $A_1,..., A_n$ in Γ, $n \geq 1$, with $\vdash_{DDL^{\{F,S\}}} (A_1 \wedge ... \wedge A_n) \rightarrow \bot$, otherwise Γ is $DDL^{\{F,S\}}$-*consistent*. $A \in \mathscr{L}_{DDL^{\{F,S\}}}$ is $DDL^{\{F,S\}}$-*derivable* from $\Gamma \subseteq \mathscr{L}_{DDL^{\{F,S\}}}$ (written $\Gamma \vdash_{DDL^{\{F,S\}}} A$) iff $\Gamma \cup \{\neg A\}$ is $DDL^{\{F,S\}}$-inconsistent.

Theorem 1 (*DDL*$^{\{F,S\}}$-theorems).
The following are DDL$^{\{F,S\}}$-derivable (is a uniform index or as indicated):*

(Ref^S)	$O^S(A/A)$ | (Ref-R*) If $\nvdash_{PL} \neg A$ then $\vdash_{DDL^{\{F,S\}}} O^*(A/A)$	
(RW*)	If $\vdash_{PL} A \to B$ then $\vdash_{DDL^{\{F,S\}}} O^*(A/C) \to O^*(B/C)$	
(Pres*)	$O^*(\bot/C) \to (O^*(A/D) \to O^*(A \wedge \neg C/D))$	
(CMon*)	$O^*(D/C) \to (O^*(A/C) \to O^*(A/C \wedge D))$	SSS, SFF, FSF
(Cut*)	$O^*(D/C) \to (O^*(A/C \wedge D) \to O^*(A/C))$	SSS, SFF, FSF
(Or*)	$(O^*(A/C) \wedge O^*(A/D)) \to O^*(A/C \vee D)$	SSS, SFF, FSF
(DR*)	$O^*(A/C \vee D) \to (O^*(A/C) \vee O^*(A/D))$	FFF, SFS, SSF
(FH*)	$P^*(C/D) \to (O^*(A/C \vee D) \to O^*(A/C))$	FFF, FSS, SSF
(FH+*)	$P^*(A \to C/C \vee D) \to (O^*(A/C \vee D) \to O^*(A/C))$	FFF, FSS, SSF
(Trans*)	$P^*(A/A \vee B) \wedge P^*(B/B \vee C) \to P^*(A/A \vee C)$	FFF, FSS, SFS
(P-Loop^F)	$P^F(A_2/A_1) \wedge ... \wedge P^F(A_n/A_{n-1}) \wedge P^F(A_1/A_n) \to P^F(A_n/A_1)$	
(Loop^S)	$O^S(A_2/A_1) \wedge ... \wedge O^S(A_n/A_{n-1}) \wedge O^S(A_1/A_n) \to O^S(A_n/A_1)$	

Proof. All easy and left to the reader.

Regarding axioms and theorems, (CExt*) is a contextual extensionality rule for consequents, and (ExtC*) an extensionality rule for antecedents. (DM*) and (DC*) are dyadic versions of their monadic analogues. The O^S-axioms are like those of Kraus, Lehmann and Magidor [26], but (Cond^S) and (CCMon^S) equivalently replace (Or^S) and (CMon^S), and (DP-R^S) is added. The O^F-axioms are those of Bochman [2], where his (Pres^F) is strengthened to (DP^F). Instead of (Cond^F), (CCMon^F) and (RMon^F), Goble [14] more elegantly employs (Trans^F) and (DR^F), which is equivalent given (DP-R^F). The 'mixed schemes' are again Bochman's. Instead of (DC ^{SF}), (RMon^{FSS}), and (RMon^{SSF}), Goble has

(DK^{SF}) $O^S(A \to B/C) \to (O^F(A/C) \to O^F(B/C))$,

(Trans^{FSS}) and (Trans^{SFS}), which is again equivalent. The names are from the study of nonmonotonic logics, namely *reflexivity*, *right weakening*, *preservation*, (*conjunctive*) *cautious monotony*, *conditionalization* and *disjunctive reasoning.* (Ref*) is Hansson's [18] th. 2, (CCMon*) Rescher's [36] th. 4.4, (Or*) is the right-to-left version of von Wright's (B3) in [46], and (DR*) Hansson's th. 13. Føllesdal and Hilpinen [9] introduced the strong version (FH*) of (RMon*) (their th. 77). (FH+*) is even stronger: its displayed versions could replace (CCMon*) and (RMon**), **$=F, FSS, SSF$. (Trans*) is transitivity of weak preference given by $A \preccurlyeq B =_{def} P^*(A/A \vee B)$ (Lewis [28] p. 54). Spohn [39] introduced (P-Loop^F) to define the relevant equivalence classes in his completeness proof of Hansson's *DSDL3*, and its O-form was rediscovered by Kraus, Lehmann and Magidor [26] who put it to the same use. Note that (CCMon*) is (P-Cond*), and (Cut*) is (P-RMon*), where the deontic operators are swapped in the P-versions.

Theorem 2. *DDL$^{\{F,S\}}$ is sound.*

Proof. The validity of (DM*), (DC^S), (DC^{SF}), (CExt*), and (ExtC*) is immediate. (DN^S), (DP^F) are valid since any subset of $\mathscr{L}_{PL}$ derives $\top$, and any maximally consistent subset is consistent. If $\nvdash_{PL} \neg C$ then at least $\varnothing$ is in $I\bot\neg C$, hence $I\bot\neg C \neq \varnothing$ and then both (DN-R^F) and (DP-R^S) hold likewise.

(Cond^F) Assume $O^F(A/C \wedge D)$, so there is an $I' \in I\bot\neg(C \wedge D)$ such that $I' \cup \{C \wedge D\} \vdash_{PL} A$ and $I' \cup \{C\} \vdash_{PL} D \rightarrow A$. Since $I' \nvdash_{PL} \neg(C \wedge D)$, also $I' \nvdash_{PL} \neg C$, so by maximality there is an $I'' \in I\bot\neg C$ such that $I' \subseteq I''$, so there is an $I'' \in I\bot\neg C : I'' \cup \{C\} \vdash_{PL} D \rightarrow A$, so $O^F(D \rightarrow A/C)$.

(Cond^S) Assume $O^S(A/C\wedge D)$. So for all $I' \in I\bot\neg(C\wedge D) : I'\cup\{C\wedge D\} \vdash_{PL} A$. If there is an $I'' \in I\bot\neg C : I'' \cup \{C\} \nvdash_{PL} D \rightarrow A$ then $I'' \cup \{C\} \nvdash_{PL} \neg D$ and $I'' \nvdash_{PL} \neg(C \wedge D)$. By maximality $\exists I' \in I\bot\neg(C \wedge D) : I'' \subseteq I'$. Since $I' \nvdash_{PL} \neg(C \wedge D)$, $I' \nvdash_{PL} \neg C$, so there is an $I''' \in I\bot\neg C : I' \subseteq I'''$. Then $I'' \subseteq I'''$ and by maximality of $I'' \in I\bot\neg C$, $I'' = I'''$ and hence $I'' = I'$. So I'' is in $I\bot\neg(C\wedge D)$ and $I''\cup\{C\wedge D\} \nvdash_{PL} A$, but this violates the assumption. So for all $I'' \in I\bot\neg C : I'' \cup \{C\} \vdash_{PL} D \rightarrow A$, and $O^S(D \rightarrow A/C)$.

(CCMon^F) Assume $O^F(A \wedge D/C)$, so $\exists I' \in I\bot\neg C : I' \cup \{C\} \vdash_{PL} A \wedge D$. Then $I'\cup\{C\} \nvdash_{PL} \neg D$, for otherwise $I' \vdash_{PL} \neg C$ which is excluded by the definition of $I\bot\neg C$. So $I' \nvdash_{PL} \neg(C \wedge D)$, by maximality $\exists I'' \in I\bot\neg(C \wedge D) : I' \subseteq I''$ and $I'' \cup \{C\} \vdash_{PL} A$, $I'' \cup \{C \wedge D\} \vdash_{PL} A$. Hence $O^F(A/C \wedge D)$.

(CCMon^S) Assume $O^S(A \wedge D/C)$, so for all $I' \in I\bot\neg C : I' \cup \{C\} \vdash_{PL} A \wedge D$, and so $I' \cup \{C\} \nvdash_{PL} \neg D$, for otherwise $I' \vdash_{PL} \neg C$ contrary to the definition of $I\bot\neg C$, and so for all $I' \in I\bot\neg C : I' \nvdash_{PL} \neg(C\wedge D)$. Suppose $I'' \in I\bot\neg(C\wedge D)$, so $I'' \nvdash_{PL} \neg(C \wedge D)$ and $I'' \nvdash_{PL} \neg C$. By maximality $\exists I' \in I\bot\neg C$ such that $I'' \subseteq I'$. In turn $I' \nvdash_{PL} \neg(C \wedge D)$ as just proved, so $\exists I''' \in I\bot\neg(C \wedge D)$ such that $I' \subseteq I'''$. But then $I'' = I'''$ by maximality of $I'' \in I\bot\neg(C \wedge D)$, so $I'' = I' \in I\bot\neg C$ and $I'' \cup \{C\} \vdash_{PL} A$ as assumed. So $I'' \cup \{C \wedge D\} \vdash_{PL} A$ for any $I'' \in I\bot\neg(C \wedge D)$. So $O^S(A/C \wedge D)$ is true.

(RMon^F) Assume $O^F(A/C)$, so $\exists I' \in I\bot\neg C : I'\cup\{C\} \vdash_{PL} A$. If $P^F(D/C)$ then $\forall I' \in I\bot\neg C : I'\cup\{C\} \nvdash_{PL} \neg D$. So $I' \nvdash_{PL} \neg(C\wedge D)$. So by maximality $\exists I'' \in I\bot\neg(C\wedge D) : I' \subseteq I''$, so $I''\cup\{C\} \vdash_{PL} A$, by monotony $I''\cup\{C\wedge D\} \vdash_{PL} A$, so $O^F(A/C \wedge D)$ is true.

(RMon^{FSS}) Assume $O^S(A/C)$, so $\forall I' \in I\bot\neg C : I'\cup\{C\} \vdash_{PL} A$, and $P^F(D/C)$, so $\forall I' \in I\bot\neg C : I' \cup \{C\} \nvdash_{PL} \neg D$, so $I' \nvdash_{PL} \neg(C \wedge D)$. Suppose $I'' \in I\bot\neg(C \wedge D)$, so also $I'' \nvdash_{PL} \neg C$ and by maximality $\exists I' \in I\bot\neg C : I'' \subseteq I'$. We have $I' \cup \{C\} \vdash_{PL} A$, so $\exists B_1, ..., B_n \in I' : \{B_1, ..., B_n\} \cup \{C\} \vdash_{PL} A$ by PL-compactness. If $I'' \cup \{C\} \nvdash_{PL} A$ then $\{B_1, ..., B_n\} \nsubseteq I''$, by maximality $I'' \cup \{B_1, ..., B_n\} \vdash_{PL} \neg(C \wedge D)$, but $I'' \cup \{B_1, ..., B_n\} \subseteq I'$, so $I' \cup \{C\} \vdash_{PL} \neg(C\wedge D)$ contrary to the assumption. So $I''\cup\{C\} \vdash_{PL} A$, $I''\cup\{C\wedge D\} \vdash_{PL} A$ for any $I'' \in I\bot\neg(C \wedge D)$. So $O^S(A/C \wedge D)$ is true.

(RMon^{SSF}) Assume $O^S(A/C)$, so $\forall I' \in I\bot\neg C : I'\cup\{C\} \vdash_{PL} A$, and $P^S(D/C)$, so $\exists I' \in I\bot\neg C : I' \cup \{C\} \nvdash_{PL} \neg D$. So $I' \nvdash_{PL} \neg(C \wedge D)$. So by maximality $\exists I'' \in I\bot\neg(C \wedge D) : I' \subseteq I''$, so $I'' \cup \{C\} \vdash_{PL} A$, so $I'' \cup \{C \wedge D\} \vdash_{PL} A$, so $O^F(A/C \wedge D)$ is true.

Theorem 3. *DDL*$^{\{F,S\}}$*-semantics are not compact.*

Proof. In [17], I provided a counterexample to the compactness of semantics that only employ the monadic deontic operator O^F. Since $O^F A$ can be defined as $O^F(A/\top)$, this also refutes the compactness of *DDL*$^{\{F,S\}}$ and of the subsystem containing only the dyadic operator O^F. The following counterexample is expressed in terms of the dyadic operators O^S only, which also refutes the compactness of the subsystem containing only this operator: let

$$\begin{array}{rll} \Gamma = & \{O^S(p_2/\top)\} & \\ & \cup\ \{P^S(\neg p_2/p_1)\} & \cup\ \bigcup_{i=3}^{\infty}\{O^S(p_i/p_1)\} \\ & \cup\ \{P^S(\neg p_2/\neg p_1)\} & \cup\ \bigcup_{i=3}^{\infty}\{O^S(p_i/\neg p_1)\} \\ & \cup\ \{P^S(\neg p_2/p_1 \leftrightarrow p_2)\} & \cup\ \bigcup_{i=3}^{\infty}\{O^S(p_i/p_1 \leftrightarrow p_2)\} \\ & \cup\ \{P^S(\neg p_2/p_1 \leftrightarrow \neg p_2)\} & \cup\ \bigcup_{i=3}^{\infty}\{O^S(p_i/p_1 \leftrightarrow \neg p_2)\} \end{array}$$

Γ is finitely $DDL^{\{F,S\}}$-satisfiable: let n be the greatest index of any proposition letter occurring in some finite $\Gamma_f \subseteq \Gamma$. Then

$$I_f = \{\ p_{n+1} \wedge (p_1 \rightarrow \neg p_2)\ ,\ \neg p_{n+1} \wedge (\neg p_1 \rightarrow \neg p_2)\ ,\ p_2, p_3, ..., p_n\}$$

satisfies Γ_f. For easy verification, I list the relevant sets of maximal subsets:

$$\begin{array}{rl} I_f \bot \neg\top & = \left\{ \begin{array}{l} \{p_{n+1} \wedge (p_1 \rightarrow \neg p_2), p_2, p_3, ..., p_n)\}, \\ \{\neg p_{n+1} \wedge (\neg p_1 \rightarrow \neg p_2), p_2, p_3, ..., p_n)\} \end{array} \right\} \\ \begin{array}{r} I_f \bot \neg p_1 \\ I_f - \bot \neg(p_1 \leftrightarrow p_2) \end{array} & = \left\{ \begin{array}{l} \{p_{n+1} \wedge (p_1 \rightarrow \neg p_2), p_3, ..., p_n)\}, \\ \{\neg p_{n+1} \wedge (\neg p_1 \rightarrow \neg p_2), p_2, p_3, ..., p_n)\} \end{array} \right\} \\ \begin{array}{r} I_f \bot p_1 \\ I_f \bot \neg(p_1 \leftrightarrow \neg p_2) \end{array} & = \left\{ \begin{array}{l} \{\ p_{n+1} \wedge (p_1 \rightarrow \neg p_2),\ p_2, p_3, ..., p_n)\ \}, \\ \{\ \neg p_{n+1} \wedge (\neg p_1 \rightarrow \neg p_2),\ p_3, ..., p_n)\} \end{array} \right\} \end{array}$$

However, Γ is not $DDL^{\{F,S\}}$-satisfiable: suppose $I \subseteq \mathscr{L}_{PL}$ satisfies Γ, and let $A \in \{p_1, \neg p_1, p_1 \leftrightarrow p_2, p_1 \leftrightarrow \neg p_2\}$. Observe that

(i) There are $I_1, I_2 \in I \bot \neg\top$ such that $I_1 \vdash_{PL} p_1 \wedge p_i$, $I_2 \vdash_{PL} \neg p_1 \wedge p_i$, $i \geq 2$.
Proof: From $O^S(p_2/\top), P^S(\neg p_2/\neg p_1) \in \Gamma$ and the validity of (RMon^{FSS}), follows $O^F(p_1/\top)$, i.e. there is an $I_1 \in I \bot \neg\top : I_1 \vdash_{PL} p_1$. Likewise from $O^S(p_2/\top)$ and $P^S(\neg p_2/p_1) \in \Gamma$, it follows that there is an $I_2 \in I \bot \neg\top : I_2 \vdash_{PL} \neg p_1$. To satisfy $O^S(p_2/\top)$ it is necessary that for all $I' \in I \bot \neg\top : I' \vdash_{PL} p_2$, and from $O^S(p_i/p_1), O^S(p_i/\neg p_1) \in \Gamma$, and the validity of (Or^S), it is obtained that for all $I' \in I \bot \neg\top : I' \vdash_{PL} p_i$, $i \geq 3$.

(ii) For each A, there is an $I_A \in I \bot \neg A : I_A \cup \{A\} \vdash_{PL} \neg p_2$.
Proof: Let $A \in \{p_1, p_1 \leftrightarrow p_2\}$. Then by observation (i) $I_1 \in I \bot \neg A$. Since $I_1 \vdash_{PL} p_2$, to satisfy $P^S(\neg p_2/A) \in \Gamma$ there is an $I_A \in I \bot \neg A$ such that $I_A \cup I_1 \vdash_{PL} \neg A$. So $I_A \cup \{A\} \vdash_{PL} \neg(p_1 \wedge p_2 \wedge ... \wedge p_n)$ for some n. If $n \geq 3$, then $I_A \cup \{A\} \vdash_{PL} \neg(p_1 \wedge p_2 \wedge ... \wedge p_{n-1})$, since $I_A \cup \{A\} \vdash_{PL} p_n$ is necessary for $O^S(p_n/A) \in \Gamma$. So $I_A \cup \{A\} \vdash_{PL} \neg(p_1 \wedge p_2)$, so $I_A \cup \{A\} \vdash_{PL} \neg p_2$. Likewise, the proof for $A \in \{\neg p_1, p_1 \leftrightarrow \neg p_2\}$ is obtained from $I_2 \in I \bot \neg A$.

(iii) If $A \in \{p_1, p_1 \leftrightarrow \neg p_2\}$ then $I_A \cup \{p_1, \neg p_2, p_3, p_4, ...\} \nvdash_{PL} \bot$.
If $A \in \{\neg p_1, p_1 \leftrightarrow p_2\}$ then $I_A \cup \{\neg p_1, \neg p_2, p_3, p_4, ...\} \nvdash_{PL} \bot$.
Proof: Suppose $A \in \{p_1, p_1 \leftrightarrow \neg p_2\}$ and $I_A \cup \{p_1, \neg p_2, p_3, p_4, ...\} \vdash_{PL} \bot$. Then $I_A \cup \{A, \neg p_2, p_3, p_4, ...\} \vdash_{PL} \bot$. So $I_A \cup \{A\} \vdash_{PL} \neg(\neg p_2 \wedge p_3 \wedge p_4 \wedge ... \wedge p_n)$ for some n. But also $I_A \cup \{A\} \vdash_{PL} \neg p_2 \wedge p_3 \wedge p_4 \wedge ... \wedge p_n$ by observation (ii) and from the fact that I satisfies $O^S(p_i/A) \in \Gamma$, $3 \leq i \leq n$. So $I_A \vdash_{PL} \neg A$, but this contradicts $I_A \in I \bot \neg A$. The proof for $A \in \{\neg p_1, p_1 \leftrightarrow p_2\}$ and the set $I_A \cup \{\neg p_1, \neg p_2, p_3, p_4, ...\}$ is done likewise.

It follows that $I_{p_1} \cup I_{(p_1 \leftrightarrow \neg p_2)} \nvdash_{PL} \bot$ and $I_{\neg p_1} \cup I_{(p_1 \leftrightarrow p_2)} \nvdash_{PL} \bot$. This is most easily seen by appealing to *PL*-semantics: some $v \in \mathbf{B}$ satisfies $\{p_1, \neg p_2, p_3, p_4, ...\}$ and by (iii) all elements of I_{p_1} as well as all of $I_{(p_1 \leftrightarrow \neg p_2)}$, so their union is satisfiable and therefore consistent (likewise for $\{\neg p_1, \neg p_2, p_3, p_4, ...\}$ and $I_{\neg p_1} \cup I_{(p_1 \leftrightarrow p_2)}$). From (ii) it follows that

$$I_{p_1} \cup I_{(p_1 \leftrightarrow \neg p_2)} \vdash_{PL} (p_1 \rightarrow \neg p_2) \wedge ((p_1 \leftrightarrow \neg p_2) \rightarrow \neg p_2)$$
$$I_{\neg p_1} \cup I_{(p_1 \leftrightarrow p_2)} \vdash_{PL} (\neg p_1 \rightarrow \neg p_2) \wedge ((p_1 \leftrightarrow p_2) \rightarrow \neg p_2)$$

But the conclusions are tautologically equivalent to $\neg p_2$, so there are consistent subsets of I that derive $\neg p_2$, and $I \nvDash O^S(p_2/\top)$, although $O^S(p_2/\top) \in \Gamma$.

Theorem 4. *$DDL^{\{F,S\}}$ is weakly complete.*

Proof. The proof follows the completeness proof of Spohn [39] for B. Hansson's [18] preference-based dyadic deontic logic *DSDL3*. Since parts of this proof will be reused in the next section for logics that might not include unrestricted (DN*) or (DP*), I will avoid their use up to the last step of this proof.

A: Preliminaries

We must prove that if $\models_{DDL^{\{F,S\}}} A$ then $\vdash_{DDL^{\{F,S\}}} A$ for any $A \in \mathscr{L}_{PL}$. We assume $\nvdash_{DDL^{\{F,S\}}} A$ so $\neg A$ is $DDL^{\{F,S\}}$-consistent. We build a disjunctive normal form of $\neg A$ and obtain a disjunction of conjunctions, where each conjunct is $O^*(B/C)$ or $\neg O^*(B/C)$. One disjunct must be $DDL^{\{F,S\}}$-consistent. Let δ be that disjunct. Let the δ-restricted language $\mathscr{L}^{\delta}_{PL}$ be the *PL*-sentences that contain only proposition letters occurring in δ. Let $r(\mathscr{L}^{\delta}_{PL})$ be 2^{2^n} mutually non-equivalent representatives of $\mathscr{L}^{\delta}_{PL}$, where n is the number of proposition letters in δ. By writing *PL*-sentences (including $\top$ and $\bot$), we now mean their unique representatives in $r(\mathscr{L}^{\delta}_{PL})$. We construct a set Δ with the following properties:

(a) Any conjunct of δ is in Δ.
(b) For all $B, C \in r(\mathscr{L}^{\delta}_{PL})$:
  - either $P^F(B/C)$ or $O^F(\neg B/C) \in \Delta$, and
  - either $P^S(B/C)$ or $O^S(\neg B/C) \in \Delta$.

(c) Δ is $DDL^{\{F,S\}}$-consistent.

It then suffices to find a set $I \subseteq \mathscr{L}_{PL}$ that makes true all $B \in \Delta$.

B: Identifying the deontic bases

We identify syntactically what Hansson called the *deontic basis* in an extension $\|C\|$ (Spohn [39] writes $\widetilde{C}$). Monadic deontic logic has just one basis, dyadic deontic logic usually has one basis for any C, and here there may be several bases, which expresses some conflict or predicament in case C.

Definition 1. *For any $C \neq \bot$, $C \in r(\mathscr{L}^{\delta}_{PL})$, let*

- $\mathcal{O}^S_C = \bigwedge\{A \in r(\mathscr{L}^{\delta}_{PL}) \mid O^S(A/C) \in \Delta\}$,
- $\mathbb{O}^F_C = min\,\{A \in r(\mathscr{L}^{\delta}_{PL}) \mid O^F(A/C) \in \Delta\}$.

where $min\,\Gamma = \{A \in \Gamma \mid \forall B \in \Gamma$, if $\vdash_{PL} B \rightarrow A$ then $\vdash_{PL} B \leftrightarrow A\}$, $\Gamma \subseteq \mathscr{L}_{PL}$.

From (DC^S), (RW*), and $DDL^{\{F,S\}}$-consistency of Δ we obtain, for any $C \neq \bot$:

(B1) $O^S(A/C) \in \Delta$ iff $\vdash_{PL} \mathcal{O}^S_C \rightarrow A$

(B2) $O^F(A/C) \in \Delta$ iff $\exists \mathcal{O} \in \mathbb{O}^F_C : \vdash_{PL} \mathcal{O} \rightarrow A$

C: Identifying the relevant class of domains

We identify the most general circumstances $\mathcal{C}_A$ where A is P^F-permitted. To the same effect, Spohn employs equivalence classes $[A]^{\approx}$ defined using (P-Loop^F): $A \approx B$ iff B is in some $\{B_1, ..., B_n\} \subseteq r(\mathscr{L}^\delta_{PL})$ with $P^F(B_1/A)$, $P^F(B_2/B_1)$, ..., $P^F(B_n/B_{n-1})$, $P^F(A/B_n) \in \Delta$. The set of all such classes is then $\{[\mathcal{C}]^{\approx} \mid \mathcal{C} \in \mathbb{C}\}$.

Definition 2. *For all $A \in r(\mathscr{L}^\delta_{PL})$, let*

$$\mathbb{C}_A = max\,\{C \in r(\mathscr{L}^\delta_{PL}) \mid P^F(A/C) \in \Delta\},$$
$$\mathbb{C} = \textstyle\bigcup_{A \in r(\mathscr{L}^\delta_{PL})} \mathbb{C}_A,$$

where $max\,\Gamma = \{A \in \Gamma \mid \forall B \in \Gamma : \text{if} \vdash_{PL} A \rightarrow B, \text{then} \vdash_{PL} B \leftrightarrow A\}$.

(C1) If $P^F(A/D) \in \Delta$, then there is a $\mathcal{C} \in \mathbb{C}_A$ such that $\vdash_{PL} D \rightarrow \mathcal{C}$.
Proof: Immediate from definition of $\mathbb{C}_A$ and finitude of $r(\mathscr{L}^\delta_{PL})$.

(C2) For all $\mathcal{C} \in \mathbb{C}$: $\mathbb{C}_{\mathcal{C}} = \{\mathcal{C}\}$.
Proof: By definition $\mathcal{C} \in \mathbb{C}_A$ for some $A \in r(\mathscr{L}^\delta_{PL})$, so by definition $P^F(A/\mathcal{C}) \in \Delta$, and $P^F(\mathcal{C}/\mathcal{C})$ due to (CExt^F). Suppose $\mathcal{C}' \in \mathbb{C}_{\mathcal{C}}$: then by definition $P^F(\mathcal{C}/\mathcal{C}') \in \Delta$. With (FH^F) we get $P^F(A/\mathcal{C} \vee \mathcal{C}'), P^F(\mathcal{C}/\mathcal{C} \vee \mathcal{C}') \in \Delta$, so $\mathcal{C} = (\mathcal{C} \vee \mathcal{C}') = \mathcal{C}'$ follows from maximality of $\mathcal{C}, \mathcal{C}'$.

(C3) For all $\mathcal{C} \in \mathbb{C}$, if $P^F(\mathcal{C}/D) \in \Delta$, then $\vdash_{PL} D \rightarrow \mathcal{C}$.
Proof: By definition $\mathcal{C} \in \mathbb{C}_A$ for some $A \in r(\mathscr{L}^\delta_{PL})$, so by definition $P^F(A/\mathcal{C}) \in \Delta$. If $P^F(\mathcal{C}/D) \in \Delta$, then we get $P^F(A/\mathcal{C} \vee D) \in \Delta$ with (FH^F). So $\mathcal{C} = (\mathcal{C} \vee D)$ by maximality, hence $\vdash_{PL} D \rightarrow \mathcal{C}$.

(C4) For all $A \neq \bot$: $\mathbb{C}_A = \{\mathcal{C}_A\}$ for some $\mathcal{C}_A \in \mathbb{C}_A$ and $\vdash_{PL} A \rightarrow \mathcal{C}_A$.
Proof: If $A \neq \bot$ then $P^F(A/A) \in \Delta$ due to (DP-R^F) and (CExt^F), so there is some $\mathcal{C} \in \mathbb{C}_A$ such that $\vdash_{PL} A \rightarrow \mathcal{C}$ by (C1). Assume $\mathcal{C}' \in \mathbb{C}_A$: By definition $P^F(A/\mathcal{C})$, $P^F(A/\mathcal{C}') \in \Delta$, so we get $P^F(\mathcal{C}/\mathcal{C}') \in \Delta$ with (RW^F), and $P^F(A/\mathcal{C} \vee \mathcal{C}') \in \Delta$ with (FH^F), so $\mathcal{C} = (\mathcal{C} \vee \mathcal{C}') = \mathcal{C}'$ by maximality. So $\mathcal{C}$ is the desired $\mathcal{C}_A$.

(C5) For all $A \neq \bot$: $\vdash_{PL} \mathcal{O}^S_A \leftrightarrow (A \wedge \mathcal{O}^S_{\mathcal{C}_A})$.
Proof: $\vdash_{PL} A \rightarrow \mathcal{C}_A$ due to (C4), and by (B2) $O^S(\mathcal{O}^S_A/A) \in \Delta$, so with (Cond^S) we obtain $O^S(A \rightarrow \mathcal{O}^S_A/\mathcal{C}_A) \in \Delta$ and thus the right-to-left direction $\vdash_{PL} (A \wedge \mathcal{O}^S_{\mathcal{C}_A}) \rightarrow \mathcal{O}^S_A$. For the opposite, $\vdash_{PL} \mathcal{O}^S_A \rightarrow A$ follows from (CExt^S), by definitions and (C4) $P^F(A/\mathcal{C}_A)$, $O^S(\mathcal{O}^S_{\mathcal{C}_A}/\mathcal{C}_A) \in \Delta$, so we get $O^S(\mathcal{O}^S_{\mathcal{C}_A}/A) \in \Delta$ with (RMon^{FSS}). So $\vdash_{PL} \mathcal{O}^S_A \rightarrow (A \wedge \mathcal{O}^S_{\mathcal{C}_A})$.

(C6) For all $A \neq \bot$, $\mathcal{O}_A \in \mathbb{O}^F_A$: $\exists \mathcal{O}_{\mathcal{C}_A} \in \mathbb{O}^F_{\mathcal{C}_A} : \vdash_{PL} \mathcal{O}_A \leftrightarrow (A \wedge \mathcal{O}_{\mathcal{C}_A})$.

Proof: Let $\mathcal{O}_A \in \mathbb{O}^F_A$, so $O^F(\mathcal{O}_A/A) \in \Delta$. $\vdash_{PL} A \rightarrow \mathcal{C}_A$ and (Cond^F) derive $O^F(A \rightarrow \mathcal{O}_A/\mathcal{C}_A) \in \Delta$, so $\exists \mathcal{O}_{\mathcal{C}_A} \in \mathbb{O}^F_{\mathcal{C}_A}$ with $\vdash_{PL} (A \wedge \mathcal{O}_{\mathcal{C}_A}) \rightarrow \mathcal{O}_A$. If $P^F(\neg\mathcal{O}_{\mathcal{C}_A}/A) \in \Delta$, then from $O^F(\mathcal{O}_{\mathcal{C}_A}/\mathcal{C}_A) \in \Delta$ and (RMon^F) we get $O^F(\neg A/\mathcal{C}_A) \in \Delta$, but by definition $P^F(A/\mathcal{C}_A) \in \Delta$. So $O^F(\mathcal{O}_{\mathcal{C}_A}/A) \in \Delta$, and $O^F(A \wedge \mathcal{O}_{\mathcal{C}_A}/A) \in \Delta$ by (CExt). Since $\vdash_{PL} (A \wedge \mathcal{O}_{\mathcal{C}_A}) \rightarrow \mathcal{O}_A$ we obtain $\vdash_{PL} \mathcal{O}_A \leftrightarrow (A \wedge \mathcal{O}_{\mathcal{C}_A})$ by minimality of $\mathcal{O}_A$.

(C7) For all $\mathcal{C} \neq \bot$, $\mathcal{C} \in \mathbb{C}$: If $\{\mathcal{C} \rightarrow \mathcal{O}_\mathcal{C}\} \cup \{D\} \nvdash \bot$, then $O^F(\mathcal{C} \rightarrow \mathcal{O}_\mathcal{C}/D) \in \Delta$.

Proof: Assume $\{\mathcal{C} \rightarrow \mathcal{O}_\mathcal{C}\} \cup \{D\} \nvdash \bot$. If $O^F(\neg\mathcal{C}/D) \in \Delta$, then the conclusion is trivial. Otherwise $P^F(\mathcal{C}/D) \in \Delta$, so $\vdash_{PL} D \rightarrow \mathcal{C}$ by (C3). For r.a.a. suppose $P^F(\neg\,\mathcal{O}_\mathcal{C}/D) \in \Delta$. With $O^F(\mathcal{O}_\mathcal{C}/\mathcal{C}) \in \Delta$ we obtain $O^F(\mathcal{O}_\mathcal{C} \wedge \neg D/\mathcal{C}) \in \Delta$ by (FH+^{F}), and $\vdash_{PL} \mathcal{O}_\mathcal{C} \rightarrow \neg D$ by minimality of $\mathcal{O}_\mathcal{C}$. But then $\vdash_{PL} D \rightarrow (\mathcal{C} \wedge \neg\mathcal{O}_\mathcal{C})$, which refutes the assumption. Hence $O^F(\mathcal{O}_\mathcal{C}/D) \in \Delta$ and $O^F(\mathcal{C} \rightarrow \mathcal{O}_\mathcal{C}/D) \in \Delta$ by use of (CExt^F).

D: Identifying the multiple system of spheres

If this were ‘ordinary’ dyadic deontic logic with agglomeration and so just one basis $\mathcal{O}_C$ for any C, we would be almost done: like Spohn [39] orders his equivalence classes $[C]^\approx$ by a relation *before*, $\mathbb{C}$ could be ordered into $\langle \mathcal{C}_1, ..., \mathcal{C}_n \rangle$ with $\mathcal{C}_1 = \top$, and $\mathcal{C}_{i+1} = \mathcal{C}_i \wedge \neg\mathcal{O}_{\mathcal{C}_i}$ until this equals $\bot$. $\langle S_1, ..., S_n \rangle$ with $S_i = (\mathcal{C}_i \wedge \neg\mathcal{C}_{i+1})$, $1 \leq i < n$, is then the ‘system of spheres’. Here this method fails since no $\mathcal{C} \in \mathbb{C}$ is guaranteed to have a single basis. But as it turns out, $\mathbb{C}$ has the structure of a ‘multiple’ system of spheres that is similarly identified.

(D1) $\{\top\} \subseteq \mathbb{C}$

Proof: $P^F(\top/\top) \in \Delta$ by (DP-R^F), and $\vdash_{PL} C \rightarrow \top$ for any $P^F(\top/C) \in \Delta$, so $\top \in \mathbb{C}_\top$, $\top \in \mathbb{C}$.

(D2) For all $\mathcal{C} \in \mathbb{C}$, $\mathcal{O} \in \mathbb{O}^F_\mathcal{C}$: If $\mathcal{C} \wedge \neg\mathcal{O} \neq \bot$, then $\mathcal{C} \wedge \neg\mathcal{O} \in \mathbb{C}$.

Proof: If $\mathcal{C} \wedge \neg\mathcal{O} \neq \bot$ then $P^F(\top/\mathcal{C} \wedge \neg\mathcal{O}) \in \Delta$ by (DP-R^F), $\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}} \in \mathbb{C}^F$. We prove $\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}} = \mathcal{C} \wedge \neg\mathcal{O}$: $\vdash_{PL} (\mathcal{C} \wedge \neg\mathcal{O}) \rightarrow \mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}}$ is immediate from (CExt^F) and (C1). If $\nvdash_{PL} \mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}} \rightarrow (\mathcal{C}\wedge\neg\mathcal{O})$ then $\{\mathcal{C} \rightarrow \mathcal{O}\} \nvdash_{PL} \neg\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}}$, so $O^F(\mathcal{C} \rightarrow \mathcal{O}/\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}}) \in \Delta$ follows from (C7). $P^F(\mathcal{C} \wedge \neg\mathcal{O}/\mathcal{C}_{\mathcal{C}\wedge\neg\mathcal{O}}) \in \Delta$ by definition, so Δ is $DDL^{\{F,S\}}$-inconsistent, but we assumed otherwise.

(D3) For all $\mathcal{C} \in \mathbb{C}$: If $\vdash_{PL} \mathcal{C} \rightarrow D$, $\mathcal{C} \neq D$, then $\exists\mathcal{O} \in \mathbb{O}^F_D$: $\vdash_{PL} \mathcal{C} \rightarrow (D \wedge \neg\mathcal{O})$.

Proof: Either $P^F(\mathcal{C}/D) \in \Delta$, so $\vdash_{PL} D \rightarrow \mathcal{C}$, $\mathcal{C} = D$ (C3). Or $O^F(\neg\mathcal{C}/D) \in \Delta$, so $\vdash_{PL} \mathcal{O} \rightarrow \neg\mathcal{C}$ for some $\mathcal{O} \in \mathbb{O}^F_D$ and $\vdash_{PL} \mathcal{C} \rightarrow (D \wedge \neg\mathcal{O})$.

(D4) For all $D \in r(\mathscr{L}^\delta_{PL})$, $\mathcal{O} \in \mathbb{O}^F_D \cup \{\mathcal{O}^S_D\}$: If $D \neq \bot$, then $D \neq (D \wedge \neg\mathcal{O})$.

Proof: If $D = (D \wedge \neg\mathcal{O})$, then $\vdash_{PL} D \rightarrow \neg\mathcal{O}$. But also $\vdash_{PL} \mathcal{O} \rightarrow D$ due to (CExt*), so $\mathcal{O} = \bot$ and $O^*(\bot/D) \in \Delta$ by (B1-2). So $D = \bot$ by (DP-R*).

(D5) Let $\mathbb{D}$ be such that (i) $\top \in \mathbb{D}$, and (ii) if $D \in \mathbb{D}$, $\mathcal{O} \in \mathbb{O}^F_D$ and $(D \wedge \neg\mathcal{O}) \neq \bot$, then $(D \wedge \neg\mathcal{O}) \in \mathbb{D}$. Then $\mathbb{D} = \mathbb{C} \backslash \{\bot\}$.

Proof: $\mathbb{D} \subseteq \mathbb{C}$ is immediate from (D1), (D2). As for $\mathbb{C} \subseteq \mathbb{D}$, for each $\mathcal{C} \in \mathbb{C}$, $\mathcal{C} \neq \bot$, there is some $D \in \mathbb{D}$ such that (a) $\vdash_{PL} \mathcal{C} \rightarrow D$, and (b) for no $\mathcal{O} \in \mathbb{O}^F_D$:$\vdash_{PL} \mathcal{C} \rightarrow (D \wedge \neg\mathcal{O})$. (a) is guaranteed by $\top \in \mathbb{D}$, and (b) follows from (D3), (D4) and finiteness of $r(\mathscr{L}^\delta_{PL})$. So $\mathcal{C} = D$ by (D3).

E: Canonical construction and coincidence lemma

Definition 3 (Canonical Construction). *For all $\mathcal{C} \in \mathbb{C} \cup \{\bot\}$, $D \in r(\mathscr{L}^{\delta}_{PL})$:*

- *F-Succ*$(D) = \{D \wedge \neg\mathcal{O} \mid \mathcal{O} \in \mathbb{O}^F_D\}$,
- *F-Chain*$(\mathcal{C})$ *be the set of sequences* $\langle D_1, ..., D_n\rangle$, $1 \leq n$, *where* $D_1 = \top$, $D_{i+1} \in$ *F-Succ*(D_i), $D_i \neq D_{i+1}$ *for any* $1 \leq i < n$, *and* $D_n = \mathcal{C}$,
- *S-Chain*$(\mathcal{C}, D)$ *be the set of sequences* $\langle D_1, ..., D_k, D_{k+1}, ..., D_n\rangle$, $1 \leq k < n$, *where* $\langle D_1, ..., D_k\rangle \in$ *F-Chain*$(\mathcal{C})$, $D_{i+1} = D_i \wedge \neg\mathcal{O}^S_{D_i}$, $D_i \neq D_{i+1}$ *for any* $k \leq i < n$, *and* $D_n = D$.

For any $\mathcal{C} \in \mathbb{C}\backslash\{\bot\}$, $\mathcal{C}' \in$ *F-Succ*$(\mathcal{C})$, *let*

- $\pi : \mathbb{C} \cup \{\bot\} \rightarrow [Prop\backslash\mathscr{L}^{\delta}_{PL}]$ *be a function that associates a unique proposition letter not occurring in* δ *with each element of* $\mathbb{C} \cup \{\bot\}$,
- $\phi(\mathcal{C}, \mathcal{C}') = \pi(\mathcal{C}') \wedge \bigwedge\{\neg\pi(\mathcal{C}'') \mid \mathcal{C}'' \in$ *F-Succ*$(\mathcal{C}), \mathcal{C}' \neq \mathcal{C}''\}$,
- $\sigma(\mathcal{C}) = \bigwedge\{\neg\pi(\mathcal{C}') \mid \mathcal{C}' \in$ *F-Succ*$(\mathcal{C})\}$.

For any $\mathcal{C} \in \mathbb{C} \cup \{\bot\} \setminus \{\top\}$, $\langle \mathcal{C}_1, ..., \mathcal{C}_n\rangle \in$ *F-Chain*$(\mathcal{C})$, *let*

- $i^F[\langle \mathcal{C}_1, ..., \mathcal{C}_n\rangle] = \neg\, \mathcal{C} \wedge \bigwedge_{i=1}^{n-1} \phi(\mathcal{C}_i, \mathcal{C}_{i+1})$.

For any $\mathcal{C} \in \mathbb{C}$, $\langle D_1, ..., D_k, D_{k+1}, ..., D_n\rangle \in$ *S-Chain*$(\mathcal{C}, D)$, $D_k = \mathcal{C}$, *let*

- $i^S[\langle D_1, ..., D_k, D_{k+1}, ..., D_n\rangle] = \neg D \wedge \begin{cases} \sigma(\mathcal{C}) \wedge \bigwedge_{i=1}^{k-1} \phi(\mathcal{C}_i, \mathcal{C}_{i+1}) & \textit{if } \mathcal{C} \neq \top, \\ \sigma(\mathcal{C}) & \textit{otherwise.} \end{cases}$

Let I^F *be the set of all such* $i^F[\langle \mathcal{C}_1, ..., \mathcal{C}_n\rangle]$, *and likewise* I^S *be the set of all such* $i^S[\langle D_1, ..., D_k, D_{k+1}, ..., D_n\rangle]$. *Then finally* $I = I^F \cup I^S$.

The definition provides the construction of the canonical set I to make all of Δ true. *F-Succ*$(\mathcal{C})$ is the set of immediate 'contrary-to-duty' successors $\mathcal{C}'$ of $\mathcal{C}$, i.e. $\exists\mathcal{O} \in \mathbb{O}^F_{\mathcal{C}}$ with $\mathcal{C}' = \mathcal{C} \wedge \neg\mathcal{O}$. (D2) showed each $\mathcal{C} \in \mathbb{C}$ to be such a successor of (a successor of ...) $\top$, and *F-Chain*$(\mathcal{C})$ is the set of all such chains beginning with $\top$ and ending with $\mathcal{C}$. ϕ is used to make any two $i^F[ch(\mathcal{C}')]$, $i^F[ch(\mathcal{C}'')]$, $\mathcal{C}' \neq \mathcal{C}''$ being successors of (successors of...) $\mathcal{C}$, inconsistent with each other and with any $i^S[ch(\mathcal{C}, D)]$ via σ. Since $\mathbb{C}$ is finite, so is the number of proposition letters introduced by π, I^F, I^S and I. – Regarding the sequences used to construct I, I use $ch(\mathcal{C})$ for $\langle D_1, ..., D_n\rangle \in$ *F-Chain*$(\mathcal{C})$ with $\mathcal{C} \in \mathbb{C} \cup \{\bot\} \setminus \{\top\}$, $ch(\mathcal{C}, D)$ for $\langle D_1, ..., D_n\rangle \in$ *S-Chain*$(\mathcal{C}, D)$ with $\mathcal{C} \in \mathbb{C}$, $D \in r(\mathscr{L}^{\delta}_{PL})$, and ch, ch' etc. for any sequence for which either holds. – We obtain:

(E1) For all ch=$\langle D_1, ..., D_n\rangle$, $\vdash_{PL} D_{i+1} \rightarrow D_i$ and $\nvdash_{PL} D_i \rightarrow D_{i+1}$, $1 \leq i < n$.

(E2) If $\{i^F[ch(\mathcal{C})], i^F[ch(\mathcal{C}')]\} \nvdash_{PL} \bot$, then $ch(\mathcal{C})$ is a segment of $ch(\mathcal{C}')$ or vice versa.

(E3) If $\{i^S[ch(\mathcal{C}, D)], i^S[ch(\mathcal{C}', D')]\} \nvdash_{PL} \bot$, then $\mathcal{C} = \mathcal{C}'$ and $ch(\mathcal{C}, D)$ is a segment of $ch(\mathcal{C}', D')$ or vice versa.

(E4) If $\{i^F[ch(\mathcal{C})], i^S[ch(\mathcal{C}', D)]\} \nvdash_{PL} \bot$, then $ch(\mathcal{C}) = \langle \mathcal{C}_1, ..., \mathcal{C}_i\rangle$ is a segment of $ch(\mathcal{C}', D) = \langle D_1, ..., D_k, D_{k+1}, ..., D_n\rangle$, where $D_k = \mathcal{C}'$ and $1 \leq i \leq k < n$.

(E5) No $i^F[ch(\mathcal{C})]$ or $i^S[ch(\mathcal{C}, D)] \in I$ is a contradiction.

Proof. (E1) is immediate from (D4). (E2-4) are immediate from the definitions of ϕ and σ. For (E5), first note that each $i \in I$ consists of a $r(\mathscr{L}^{\delta}_{PL})$-conjunct and a $[\mathscr{L}_{PL}\backslash\mathscr{L}^{\delta}_{PL}]$-conjunct. Since no proposition letter occurring in one occurs in the other, if i is a contradiction, then so must be one of its conjuncts. Regarding the $r(\mathscr{L}^{\delta}_{PL})$-conjunct, for any $i^F[ch(\mathcal{C})]$ it is $\neg\mathcal{C}$ which must be consistent since $\mathcal{C} = \top$ is excluded. For any $i^S[ch(\mathcal{C}, D)]$ the $r(\mathscr{L}^{\delta}_{PL})$-conjunct is $\neg D$, and $ch(\mathcal{C}, D) = \langle D_1, ..., D_n\rangle$ with $D_1 = \top$, $D_n = D$ and $n \neq 1$, so by (E1) $D = \top$ is excluded. For the $[\mathscr{L}_{PL}\backslash\mathscr{L}^{\delta}_{PL}]$-conjunct of any $i^F[ch(\mathcal{C})]$, $ch(\mathcal{C}) = \langle \mathcal{C}_1, ..., \mathcal{C}_n\rangle \in F\text{-}Chain(\mathcal{C})$, it is $\bigwedge_{i=1}^{n-1} \phi(\mathcal{C}_i, \mathcal{C}_{i+1})$, a conjunction of conjunctions of non-negated and negated proposition letters, which cannot be a contradiction:

- No conjunct $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$, $1 \leq i < n$, is a contradiction: For any $\mathcal{C}' \neq \mathcal{C}'' \in \mathbb{C}$, $\pi(\mathcal{C}') \neq \pi(\mathcal{C}'')$, and no $\pi(\mathcal{C}')$ occurs negated and non-negated in $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$.
- If $\pi(\mathcal{C}')$ occurs non-negated in $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$ and negated in $\phi(\mathcal{C}_j, \mathcal{C}_{j+1})$, $i < j$, then $\mathcal{C}' = \mathcal{C}_{i+1}$ and $\mathcal{C}' \in F\text{-}Succ(\mathcal{C}_j)$. So there is a $ch(\mathcal{C}') \in F\text{-}Chain(\mathcal{C}')$, $ch(\mathcal{C}') = \langle \mathcal{C}_1, ..., \mathcal{C}_i, \mathcal{C}', ..., \mathcal{C}_j, \mathcal{C}'\rangle$, which violates (E1).
- If $\pi(\mathcal{C}')$ occurs negated in $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$ and non-negated in $\phi(\mathcal{C}_j, \mathcal{C}_{j+1})$, $i < j$, then $\mathcal{C}' \in F\text{-}Succ(\mathcal{C}_i)$ and $\mathcal{C}' = \mathcal{C}_{j+1}$. $\vdash_{PL} \mathcal{C}_{j+1} \rightarrow \mathcal{C}_{i+1}$, so $\vdash_{PL} \mathcal{C}' \rightarrow \mathcal{C}_{i+1}$. So there are $\mathcal{O}_1, \mathcal{O}_2 \in \mathbb{O}^F_{\mathcal{C}_i}$ with $\mathcal{C}' = \mathcal{C}_i \wedge \neg\mathcal{O}_1$, $\mathcal{C}_{i+1} = \mathcal{C}_i \wedge \neg\mathcal{O}_2$, and $\vdash_{PL} (\mathcal{C}_i \wedge \neg\mathcal{O}_1) \rightarrow (\mathcal{C}_i \wedge \neg\mathcal{O}_2)$. Then $\vdash_{PL} \mathcal{O}_2 \rightarrow (\mathcal{C}_i \rightarrow \mathcal{O}_1)$, and with (CExt^F) $\vdash_{PL} \mathcal{O}_2 \rightarrow \mathcal{O}_1$. By minimality $\mathcal{O}_2 = \mathcal{O}_1$ and $\mathcal{C}' = \mathcal{C}_{i+1}$, but $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$ left $\pi(\mathcal{C}_{i+1})$ non-negated.

For the $[\mathscr{L}_{PL}\backslash\mathscr{L}^{\delta}_{PL}]$-conjunct of $i^S[ch(\mathcal{C}, D)]$, the case that $\pi(\mathcal{C}')$ occurs non-negated in $\phi(\mathcal{C}_i, \mathcal{C}_{i+1})$ and negated in $\sigma(\mathcal{C})$ is done like the second case above.

For any $B \in r(\mathscr{L}^{\delta}_{PL})$, $I' \in I\bot\neg B$, $I' \neq \varnothing$:

(E6) there is some designated $i^{\mathrm{x}} \in I'$, $i^{\mathrm{x}} = i^F[ch]$ or $i^{\mathrm{x}} = i^S[ch]$, such that for all $i^F[ch']$, $i^S[ch']$ in I', ch is a segment of ch',

(E7) the $r(\mathscr{L}^{\delta}_{PL})$-conjunct of i^{x} *PL*-derives the $r(\mathscr{L}^{\delta}_{PL})$-conjunct of any $i \in I'$,

(E8) i^{x} is $i^F[ch]$ or $i^S[ch]$ with $ch = \langle D_1, ..., D_n\rangle$ such that $\vdash_{PL} B \rightarrow D_{n-1}$.

Proof. (E6) is immediate from (E2-4) and finiteness of I'. For (E7), the $r(\mathscr{L}^{\delta}_{PL})$-conjunct of i is $\neg D_n$, where D_n is the last member of some $ch = \langle D_1, ..., D_n\rangle$ such that $i = i^F[ch]$ or $i = i^S[ch]$. By (E6) the $r(\mathscr{L}^{\delta}_{PL})$-conjunct of i^{x} is $\neg D_k$ for some $1 \leq k \leq n$, and by (E1) $\vdash_{PL} \neg D_k \rightarrow \neg D_n$. For (E8), note that D_{n-1} exixts as $ch = \langle\top\rangle$ is excluded by the construction. If $D_{n-1} = \top$, then $\vdash_{PL} B \rightarrow D_{n-1}$ is trivial. Otherwise there must be some $ch' = \langle D_1, ..., D_{n-1}\rangle$ such that $i^*(ch') \in I$, * being F or S. By (E6), $i^*(ch')$ cannot be in I', though the $r(\mathscr{L}^{\delta}_{PL})$-conjunct $\neg D_{n-1}$ derives the $r(\mathscr{L}^{\delta}_{PL})$-conjunct $\neg D_n$ of i^{x} by (E1) and hence that of any other $i' \in I'$ due to (E7), while its $[\mathscr{L}_{PL}\backslash\mathscr{L}^{\delta}_{PL}]$-conjunct is derived by that of i^{x}. So it must be that $\{\neg D_{n-1}\} \cup \{B\} \vdash_{PL} \bot$, and $\vdash_{PL} B \rightarrow D_{n-1}$.

Lemma 1 (Coincidence Lemma). *For all $A, B \in r(\mathscr{L}^{\delta}_{PL})$:*
$I \models O^F(A/B)$ *iff* $O^F(A/B) \in \varDelta$
$I \models O^S(A/B)$ *iff* $O^F(A/B) \in \varDelta$

Proof. Coincidence for O^F:

Right-to-Left: Assume $O^F(A/B) \in \Delta$, so some $\mathcal{O}_B \in \mathbb{O}^F_B$ derives A. By (C6) $\vdash_{PL} ((\mathcal{C}_B \to \mathcal{O}_{\mathcal{C}_B}) \wedge B) \leftrightarrow \mathcal{O}_B$ for some $\mathcal{C}_B \in \mathbb{C}$, $\mathcal{O}_{\mathcal{C}_B} \in \mathbb{O}^F_{\mathcal{C}_B}$. By (D2), $\mathcal{C}_B \wedge \neg\mathcal{O}_{\mathcal{C}_B} \in \mathbb{C} \cup \{\bot\}$, so $i^F[ch] \in I$ for some $ch \in F\text{-}Chain(\mathcal{C}_B \wedge \neg\mathcal{O}_{\mathcal{C}_B})$. If $\{i^F[ch]\}$ *PL*-derives $\neg B$, only its $r(\mathscr{L}^{\delta}_{PL})$-conjunct $\mathcal{C}_B \to \mathcal{O}_{\mathcal{C}_B}$ can be relevant, since the $[\mathscr{L}_{PL} \backslash \mathscr{L}^{\delta}_{PL}]$-conjunct is consistent (E5) and has no proposition letter in common with $\neg B$. If $\{\mathcal{C}_B \to \mathcal{O}_{\mathcal{C}_B}, B\} \vdash_{PL} \bot$, then $\mathcal{O}_B = \bot$ which contradicts $P^F(\top/B) \in \Delta$ by (DP^F) and $DDL^{\{F,S\}}$-consistency of Δ. So $\{i^F[ch]\} \nvdash_{PL} \neg B$, so for some $I' \in I\bot\neg B$: $I' \cup \{B\} \vdash_{PL} \mathcal{C}_B \to \mathcal{O}_{\mathcal{C}_B}$, hence $I' \cup \{B\} \vdash_{PL} A$.

Left-to-Right: Assume $O^F(A/B) \notin \Delta$, so $P^F(A/B) \in \Delta$ and for r.a.a. suppose $\exists I' \in I\bot\neg B$: $I' \cup \{B\} \vdash_{PL} \neg A$. Suppose $I' \neq \varnothing$, so let i^{x} be the designated member of I' and $\neg D$ its $r(\mathscr{L}^{\delta}_{PL})$-conjunct. Then $\{\neg D\} \cup \{B\} \vdash_{PL} \neg A$ as the $r(\mathscr{L}^{\delta}_{PL})$-conjuncts of any $i \in I$ are *PL*-derived by $\neg D$ (E7), and the $[\mathscr{L}_{PL} \backslash \mathscr{L}^{\delta}_{PL}]$-conjuncts are not relevant for a derivation of $\neg A \in r(\mathscr{L}^{\delta}_{PL})$. i^{x} is $i^F[ch]$ or $i^S[ch]$ for some $ch = \langle D_1, ..., D_n\rangle$ with $\neg D = \neg D_n = (D_{n-1} \to \mathcal{O})$, $\mathcal{O} \in \mathbb{O}^F_{D_{n-1}} \cup \{\mathcal{O}^S_{D_{n-1}}\}$. So $\{\mathcal{O}\} \vdash_{PL} B \to \neg A$, so $O^F(B \to \neg A/D_{n-1}) \in \Delta$ or $O^S(B \to \neg A/D_{n-1}) \in \Delta$ by (B1), (B2). From $P^F(A/B) \in \Delta$ we get $P^F(\neg(B \to \neg A)/B) \in \Delta$ with (CExt^F). By (E8) $\vdash_{PL} B \to D_{n-1}$, so with (FH+^{F}) or (FH+^{SSF}) we obtain $O^F((B \to \neg A) \wedge \neg B/D_{n-1}) \in \Delta$ or $O^S((B \to \neg A) \wedge \neg B/D_{n-1}) \in \Delta$ respectively. But then $\vdash \mathcal{O} \to \neg B$ follows from minimality of $\mathcal{O}$. So $\{B\} \vdash_{PL} D_{n-1} \wedge \neg\mathcal{O}$, i.e. $\{B\} \vdash_{PL} D$, and since $\{i^{\mathrm{x}}\} \vdash_{PL} \neg D$ we get $i^{\mathrm{x}} \notin I'$. So there is no designated member of I', so by (E6) $I' = \varnothing$. Then $\{B\} \vdash_{PL} \neg A$. With $P^F(A/B) \in \Delta$ and (CExt^F) we get $P^F(\bot/B)$, so by (DN-R^F) $B = \bot$ and $I\bot\neg B = \varnothing$, completing the r.a.a.

Coincidence for O^S:

Right-to-Left: Assume $O^S(A/B) \in \Delta$ and for r.a.a. suppose that there is some $I' \in I\bot\neg B$, : $I' \cup \{B\} \nvdash_{PL} A$. Assume $I' \neq \varnothing$, so let i^{x} be the designated member of I', and $\neg D$ its $r(\mathscr{L}^{\delta}_{PL})$-conjunct. i^{x} is $i^F[ch]$ or $i^S[ch]$ for some $ch = \langle D_1, ..., D_n\rangle$ with $\neg D = \neg D_n = (D_{n-1} \to \mathcal{O})$. Either $\mathcal{O} \in \mathbb{O}^F_{D_{n-1}}$, then $\vdash_{PL} \mathcal{O} \to \mathcal{O}^S_{D_{n-1}}$ follows from (DC^{SF}), or trivially if $\mathcal{O} = \mathcal{O}^S_{D_{n-1}}$. By (E8) $\vdash_{PL} B \to D_{n-1}$, so by (Cond^S) $\vdash_{PL} \mathcal{O}^S_{D_{n-1}} \to (B \to \mathcal{O}^S_B)$, and also $I' \cup \{B\} \vdash_{PL} \mathcal{O}$. Chaining the results, we get $I' \cup \{B\} \vdash_{PL} \mathcal{O}^S_B$, and $I' \cup \{B\} \vdash_{PL} A$ by definition of $\mathcal{O}^S_B$, contrary to what was assumed. So $I' = \varnothing$. For any $B \in r(\mathscr{L}^{\delta}_{PL})$, $B \neq \bot$, we have $i^S[ch(\mathcal{C}_B, (\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))] \in I^S$ by (C4) and definition of I^S. If $I' = \varnothing$, then $\{i^S[ch(\mathcal{C}_B, (\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))]\} \cup \{B\} \vdash_{PL} \bot$, and $\{\mathcal{C}_B \to \mathcal{O}^S_{\mathcal{C}_B}\} \cup \{B\} \vdash_{PL} \bot$ since only the $r(\mathscr{L}^{\delta}_{PL})$-conjunct is relevant. By (C4) we have $\vdash_{PL} B \to \mathcal{C}_B$, so $\{B \wedge \mathcal{O}^S_{\mathcal{C}_B}\} \vdash_{PL} \bot$, and by (C5) $\mathcal{O}^S_B = \bot$. From (DP-R^S) we get $B = \bot$, so $I\bot\neg B = \varnothing$, which completes the r.a.a.

Left-to-Right: Assume $O^S(A/B) \notin \Delta$. $B \neq \bot$ due to (DN^S) and (CExt^S), so $i^S[ch(\mathcal{C}_B, (\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))] \in I'$ for some $I' \in I\bot\neg B$ (otherwise again $B = \bot$). If $i^F[ch(\mathcal{C}')] \in I'$, then $ch(\mathcal{C}')$ is a segment of $ch(\mathcal{C}_B, (\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))$ by (E4), so $\vdash_{PL} \neg\mathcal{C}' \to \neg\mathcal{C}_B$ and as $\vdash_{PL} B \to \mathcal{C}_B$ also $\vdash_{PL} \neg\mathcal{C}' \to \neg B$. So $i^F[ch(\mathcal{C}')] \notin I'$ and $I' \cap I^F = \varnothing$. If $i^S[ch(\mathcal{C}', D')] \in I'$, then $ch(\mathcal{C}_B, (\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))$ is a segment of $ch(\mathcal{C}', D')$ by (E3), so its $r(\mathscr{L}^{\delta}_{PL})$-conjunct $\neg D'$ is derived by that of $i^S[ch(\mathcal{C}_B, (\mathcal{C}_B \wedge \neg\mathcal{O}^S_{\mathcal{C}_B}))]$. The $[\mathscr{L}_{PL}\backslash\mathscr{L}^\delta_{PL}]$-conjuncts are not relevant, so if $I' \cup B \vdash_{PL} A$, then $\{\mathcal{C}_B \rightarrow \mathcal{O}^S_{\mathcal{C}_B})\} \cup \{B\} \vdash_{PL} A$. Since $\vdash_{PL} B \rightarrow \mathcal{C}_B$, by (C4) then $\{B \wedge \mathcal{O}^S_{\mathcal{C}_B})\} \vdash_{PL} A$, and by (C5) $\vdash_{PL} \mathcal{O}^S_B \rightarrow A$, so $O^S(A/B) \in \Delta$ by (B1), contrary to the assumption. So $I' \cup B \nvdash_{PL} A$, so not for all $I' \in I\bot\neg B : I' \cup \{B\} \vdash_{PL} A$, so $I \models O^S(A/B)$ is false.

6 A Link to Multiplex Preference Semantics

In the preceding section, the completeness theorem was proved by identifying a multiple system of spheres. This multiple system of spheres can just as well be used to construct a multitude of preference relations, which – as originated with Goble [13], [14] – can then in turn be used to define the deontic operators: let $\mathbb{P}$ be a non-empty set of preference relations $P \subseteq \mathbf{B} \times \mathbf{B}$ such that each P is transitive, connected, and satisfies the 'limit assumption' :

(**LA**) If $\|A\| \neq \varnothing$ then $best_P(\|A\|) \neq \varnothing$

where $best_P(\|A\|) = \{v \in \|A\| \mid \forall v' \in \|A\| : vPv'\}$. For Hansson-type operators, let $\mathscr{L}_{DDL^{\{F+,S\}}}$ be like $\mathscr{L}_{DDL^{\{F,S\}}}$ except that O^{F^+} replaces O^F, and let the truth definitions for the deontic operators read:

$$\begin{array}{lll} \mathbb{P} \models O^{F^+}(A/C) & \text{iff} & \exists P \in \mathbb{P} : best_P(\|C\|) \subseteq \|A\| \\ \mathbb{P} \models O^S(A/C) & \text{iff} & \forall P \in \mathbb{P} : best_P(\|C\|) \subseteq \|A\| \end{array}$$

Likewise, for Lewis-type operators, let $\mathscr{L}_{DDL^{\{F,S-\}}}$ be like $\mathscr{L}_{DDL^{\{F,S\}}}$ except that O^{S^-} replaces O^S, and the truth definitions now read:

$$\begin{array}{lll} \mathbb{P} \models O^F(A/C) & \text{iff} & \exists P \in \mathbb{P} : \exists v \in \|C \wedge A\|) : \forall v' \in \|C \wedge \neg A\| : \text{not } v'Pv \\ \mathbb{P} \models O^{S^-}(A/C) & \text{iff} & \forall P \in \mathbb{P} : \exists v \in \|C \wedge A\|) : \forall v' \in \|C \wedge \neg A\| : \text{not } v'Pv \end{array}$$

The *axiomatic system* $DDL^{\{F^+,S\}}$ is like $DDL^{\{F,S\}}$ except that (DN^F) replaces (DN-R^F)and (DP-R^F) replaces (DP^F). Similarly, $DDL^{\{F,S^-\}}$ is like $DDL^{\{F,S\}}$ except that (DN-R^S) replaces (DN^S) and (DP^S) replaces (DP-R^S). So all systems only differ on the 'mind-boggling' [29] question whether everything or nothing is obligatory in impossible circumstances. $DDL^{\{F^+,S\}}$ and $DDL^{\{F,S^-\}}$ are sound (cf. [14], also Arrow's axiom: if $best_P(\|C \vee D\|) \cap \|C\| \neq \varnothing$, then $best_P(\|C\|) = best_P(\|C \vee D\|) \cap \|C\|$, is helpful). I have no counterexample to compactness, so the semantics might just be compact. Weak completeness is easily obtained from the previous constructions, but seems not to have been stated before, so I shall give the proof in full.

Theorem 5 (Completeness of *DDL*$^{\{F^+,S\}}$ and *DDL*$^{\{F,S^-\}}$).
The systems $DDL^{\{F^+,S\}}$ and $DDL^{\{F,S^-\}}$ are weakly complete with respect to the above multiplex preference semantics.

Proof. In proving $DDL^{\{F,S\}}$-completeness, up till the coincidence lemma no use was made use of unrestricted (DP^F) and (DN^S) missing in $DDL^{\{F^+,S\}}$ and $DDL^{\{F,S^-\}}$ respectively. So we can reuse and continue that construction with all pertaining lemmas in the canonical construction for $DDL^{\{F^+,S\}}$ and $DDL^{\{F,S^-\}}$, with the implicit understanding that for $\mathit{DDL}^{\{F^+,S\}}$ the index meant is F^+ rather than F, and for $\mathit{DDL}^{\{F,S^-\}}$ the index meant is S^- instead of S.

Let $\mathit{F\text{-}Chain}(\bot)$, $\mathit{S\text{-}Chain}(\mathcal{C},\bot)$, be defined as before, $\mathcal{C} \in \mathbb{C}$. We only consider $\mathit{ch} = \langle D_1, ..., D_n \rangle$ that are in such a set. Let $\mathcal{O}^{ch}_{D_i} = D_i \wedge \neg D_{i+1}$, for any $1 \leq i < n$. Note that $D_{i+1} = D_i \wedge \neg \mathcal{O}$ for some $\mathcal{O} \in \mathbb{O}^F_{D_i} \cup \{\mathcal{O}^S_{D_i}\}$, and $\vdash \mathcal{O} \rightarrow D_i$ by (CExt*), so $\vdash_{PL} \mathcal{O}^{ch}_{D_i} \leftrightarrow \mathcal{O}$. For any $1 \leq i < j \leq n$:

(S1) $\|D_j\| \subseteq \|D_i\|$
(S2) $\|D_j\| \cap \|\mathcal{O}^{ch}_{D_i}\| = \varnothing$ (and $\|\mathcal{O}^{ch}_{D_j}\| \cap \|\mathcal{O}^{ch}_{D_i}\| = \varnothing$ with CExt*, $j \neq n$)
(S3) $n \neq \infty$
(S4) $\|\mathcal{O}^{ch}_{D_1}\| \cup ... \cup \|\mathcal{O}^{ch}_{D_{n-1}}\| = \mathbf{B}$

Proof. (S1) and (S2) are immediate from (E1) and the definition of ch. n is finite (S3) since $r(\mathscr{L}^{\delta}_{PL})$ is finite and repetitions in ch are excluded. For (S4), $\mathbf{B} \backslash \|\mathcal{O}^{ch}_{D_1}\| \cup ... \cup \|\mathcal{O}^{ch}_{D_n}\| = \|D_{n-1} \wedge \neg \mathcal{O}^{ch}_{D_{n-1}}\| = \|D_{n-1} \wedge (D_{n-1} \rightarrow D_n)\|$, and $D_n = \bot$ by definition.

For any $\mathit{ch} = \langle D_1, ..., D_n \rangle$, $v, v' \in \mathbf{B}$, let P_{ch} be such that

$$vP_{ch}v' \quad \textit{iff} \quad v \in \|\mathcal{O}^{ch}_{D_i}\|,\ v' \in \|\mathcal{O}^{ch}_{D_j}\|,\ i \leq j < n$$

By (S2) and (S4), each v must belong to exactly one sphere. The index of each $\mathcal{C}_i$ is transitive and connected, so P_{ch} is as well. **LA** holds due to (S3) and (S4).

Let $\mathit{ch} = \langle D_1, ..., D_n \rangle$ be as described above. Let D_i be its "smallest A-permitting sphere", i.e. a D_i with $\|\mathcal{O}^{ch}_{D_i}\| \cap \|A\| \neq \varnothing$ and $\forall j$, $1 \leq j < i < n$: $\|\mathcal{O}^{ch}_{D_j}\| \cap \|A\| = \varnothing$ (we write D_A for D_i). We then obtain, for any ch and $A \neq \bot$:

(S5) There is a $D_A \in \mathit{ch}$,
(S6) $\|A\| \subseteq \|D_A\|$, and
(S7) $\mathit{best}_{P_{ch}}(\|A\|) = \|\mathcal{O}^{ch}_{D_A}\| \cap \|A\|$.

Proof. (S5) is immediate from (S1), (S4) and $D_1 = \top$. For (S6) let $D_A = D_i \in \mathit{ch}$, $1 \leq i < n$: If $\|A\| \not\subseteq \|D_A\|$, then $A \wedge \neg D_A \neq \bot$, so by (S5) there is a $D_{A \wedge \neg D_A} = D_j \in \mathit{ch}$. If $D_A = D_{A \wedge \neg D_A}$, then $\|\mathcal{O}^{ch}_{D_A}\| = \|\mathcal{O}^{ch}_{D_{A \wedge \neg D_A}}\|$, by construction $\vdash_{PL} \mathcal{O}^{ch}_{D_A} \rightarrow D_A$, so $\|\mathcal{O}^{ch}_{D_{A \wedge \neg D_A}}\| \cap \|A \wedge \neg D_A\| = \varnothing$ contrary to the definition of $D_{A \wedge \neg D_A}$. If $i < j$, then $\|D_{A \wedge \neg D_A}\| \subseteq \|D_A\|$, but then $\|D_{A \wedge \neg D_A}\| \cap \|A \wedge \neg D_A\| = \varnothing$ and again $\|\mathcal{O}^{ch}_{D_{A \wedge \neg D_A}}\| \cap \|A \wedge \neg D_A\| = \varnothing$. So $j < i$, but then $\|\mathcal{O}^{ch}_{D_A}\| \cap \|A \wedge \neg D_A\| \neq \varnothing$ implies $\|\mathcal{O}^{ch}_{D_A}\| \cap \|A\| \neq \varnothing$, so D_A was not the smallest A-permitting sphere. (S7) then follows from the definitions of $\mathcal{O}^{ch}_{D_A}$ and P_{ch}.

Finally, let

$$\mathbb{P} = \{\ P_{ch} \mid \mathit{ch} \in \mathit{F\text{-}Chain}(\bot) \cup \bigcup_{\mathcal{C} \in \mathbb{C}} \mathit{S\text{-}Chain}(\mathcal{C},\bot)\ \}$$

Note that $\mathbb{P} \neq \varnothing$: by (D1) $\top \in \mathbb{C}$, so even if $\top \wedge \mathcal{O}_{\top} = \bot$ for all $\mathcal{O} \in \mathbb{O}^F_{\top} \cup \mathcal{O}^S_{\top}$, then $\langle \top, \bot \rangle \in \mathit{F\text{-}Chain}(\bot)$, $\langle \top, \bot \rangle \in \mathit{S\text{-}Chain}(\top, \bot)$, hence $P_{\langle \top, \bot \rangle} = \mathbf{B} \times \mathbf{B} \in \mathbb{P}$. So $\mathbb{P}$ is as required.

The next lemma holds for all $A \in r(\mathscr{L}^{\delta}_{PL})$, $A \neq \bot$ and $P \in \mathbb{P}$ and saves us from having to do separate proofs for the two systems:

(S8) $\mathit{best}_P(\|A\|) \subseteq \|B\|$ iff $\exists v \in \|A \wedge B\| : \forall v' \in \|A \wedge \neg B\| : \text{not } v'Pv$.

Proof: Assume $\mathit{best}_P(\|A\|) \subseteq \|B\|$: $A \neq \bot$, so $\mathit{best}_P(\|A\|) \neq \varnothing$ due to (**LA**). So $\exists v \in \mathit{best}_P(\|A\|)$ s.t. $v \in \|A \wedge B\|$. Suppose $v'Pv$ for some $v' \in \|A \wedge \neg B\|$: so $v' \in \|A\|$ and $v' \in \mathit{best}_P(\|A\|)$ by transitivity of P and definition of *best*. But then $\mathit{best}_P(\|A\|) \cap \|\neg B\| \neq \varnothing$, contradicting the assumption. Assume $\exists v \in \|A \wedge B\|) :$ $\forall v' \in \|A \wedge \neg B\| : \text{not } v'Pv$, and for r.a.a. suppose that $\mathit{best}_P(\|A\|) \cap \|\neg B\| \neq \varnothing$: So $\exists v' \in \mathit{best}_P(\|A\|) \cap \|\neg B\|$. Then $v' \in \|A \wedge \neg B\|$ by definition of *best*, and not $v'Pv$, as assumed. But $v \in \|A \wedge B\|$, so $v \in \|A\|$, so since $v' \in \mathit{best}_P(\|A\|)$ we have $v'Pv$ by definition of *best*, which completes the r.a.a.

Lemma 2 (Coincidence Lemma). *For all* $A, B \in r(\mathscr{L}^{\delta}_{PL})$*:*
$\mathbb{P} \models O^F(A/B)$ *iff* $O^F(A/B) \in \mathit{\Delta}$
$\mathbb{P} \models O^S(A/B)$ *iff* $O^F(A/B) \in \mathit{\Delta}$

Proof. Coincidence for O^F*:*

Case $B = \bot$**:** If $B = \bot$, then in the case of $\mathit{DDL}^{\{F^+,S\}}$, $O^{F+}(\top/\bot) \in \mathit{\Delta}$ holds due to (DN$^{F^+}$), so $O^{F^+}(A/\bot) \in \mathit{\Delta}$ due to (CExt$^{F^+}$). Also, if $B = \bot$, then for any $P \in \mathbb{P}$, $\mathit{best}_P(\|B\|) = \varnothing$. $\mathbb{P} \neq \varnothing$, so $\exists P \in \mathbb{P} : \mathit{best}_P(\|B\|) \subseteq \|A\|$ holds for any A, and both sides of the iff-clause are true, and so is the iff-clause. – In the case of $\mathit{DDL}^{\{F,S^-\}}$, if $B = \bot$, then $P^F(\top/\bot) \in \mathit{\Delta}$ due to (DP^F), so $P^F(\neg A/\bot) \in \mathit{\Delta}$ due to (CExt^F), and so by definition of $\mathit{\Delta}$, $O^F(A/\bot) \notin \mathit{\Delta}$. Also, if $B = \bot$, then $\|A \wedge B\| = \varnothing$, so for any P it is false that there is some $v \in \|A \wedge B\|$ such that $\forall v' \in \|A \wedge \neg B\| : \text{not } v'Pv$. So both sides of the iff-clause are false, and the iff-clause true.

Case $B \neq \bot$**:** *Right-to-left:* $O^F(A/B) \in \mathit{\Delta}$, so $\exists \mathcal{O}_B \in \mathbb{O}^F_B$: $\vdash_{PL} \mathcal{O}_B \rightarrow A$, so by (C6) $\exists \mathcal{O}_{\mathcal{C}_B} \in \mathbb{O}^F_{\mathcal{C}_B}$: $\mathcal{O}_{\mathcal{C}_B} \wedge B = \mathcal{O}_B$. Since $\mathcal{C}_B \in \mathbb{C}$, there is a $ch \in$ *F-Chain*$(\bot)$ such that $ch = \langle D_1, ..., D_i, D_{i+1}, ..., D_n \rangle$, $1 \leq i < n$, $D_1 = \top$, $D_i = \mathcal{C}_B$, $D_{i+1} = \mathcal{C} \wedge \neg \mathcal{O}_{\mathcal{C}_B}$, and $D_n = \bot$. By definition $\mathcal{O}^{ch}_{D_i} = \mathcal{O}_{\mathcal{C}_B}$. For any $1 \leq j < i$, $\|\mathcal{O}^{ch}_{D_j}\| \cap \|\mathcal{C}_B\| = \varnothing$ due to (S2), so also $\|\mathcal{O}^{ch}_{D_j}\| \cap \|B\| = \varnothing$, and $\|\mathcal{O}^{ch}_{D_i}\| \cap \|B\| \neq \varnothing$, for otherwise $\mathcal{O}_{\mathcal{C}_B} \wedge B = \bot = \mathcal{O}_B$, by (B2) $O^F(\bot/B) \in \mathit{\Delta}$, and by (DP-R^F) $B = \bot$, contrary to what was assumed. So $D_i = D_B$, and $\mathit{best}_{P_{ch}}(\|B\|) = \|\mathcal{O}^{ch}_{D_B}\| \cap \|B\| = \|\mathcal{O}_{\mathcal{C}_B} \wedge B\| = \|\mathcal{O}_B\|$, and so $\mathit{best}_{P_{ch}}(\|B\|) \subseteq \|A\|$ and by (S8) $\mathbb{P} \models O^F(A/B)$.

Left-to-right: Suppose $\mathbb{P} \models O^F(A/B)$, so $\exists P \in \mathbb{P} : \mathit{best}_P(\|B\|) \subseteq \|A\|$ by (S8). By construction there is some $ch = \langle D_1, ..., D_n \rangle$ such that $P = P_{ch}$, $ch \in$ *F-Chain*$(\bot)$ or $ch \in$ *S-Chain*$(\mathcal{C}, \bot)$ for some $\mathcal{C} \in \mathbb{C}$. As $B \neq \bot$, by (S5) there is some 'smallest B-permitting sphere' D_B in ch, with $\mathit{best}_P(\|B\|) = \|\mathcal{O}^{ch}_{D_B}\| \cap \|B\|$ and $\|\mathcal{O}^{ch}_{D_B}\| \cap \|B\| \neq \varnothing$. Either $D_B \in \mathbb{C}$ and $\mathcal{O}^{ch}_{D_B} \in \mathbb{O}^F_{D_B}$: then $\{D_B \rightarrow \mathcal{O}^{ch}_{D_B}\} \nvdash_{PL} \neg B$, so $O^F(D_B \rightarrow \mathcal{O}^{ch}_{D_B}/B) \in \mathit{\Delta}$ by (C7), $\vdash_{PL} B \rightarrow D_B$ by (S6), so $O^F(\mathcal{O}^{ch}_{D_B} \wedge B/B) \in \mathit{\Delta}$ with (CExt^F) and $O^F(A/B) \in \mathit{\Delta}$ by (RW^F). Or $\mathcal{O}^{ch}_{D_B} = \mathcal{O}^S_{D_B}$, so $O^S(\mathcal{O}^{ch}_{D_B}/D_B) \in \mathit{\Delta}$, and $O^S(B \rightarrow A/D_B) \in \mathit{\Delta}$ by (RW^S). Assume $O^F(A/B) \notin \mathit{\Delta}$, then $P^F(\neg A/B) \in \mathit{\Delta}$ by definition of $\mathit{\Delta}$, so $P^F(\neg(B \rightarrow A)/B) \in \mathit{\Delta}$ by (CExt^F), and so with (FH+^{SSF}) we obtain $O^S((B \rightarrow A) \wedge \neg B/D_B) \in \mathit{\Delta}$, so by definition $\vdash_{PL} \mathcal{O}^S_{D_A} \rightarrow \neg B$, but then $\|\mathcal{O}^{ch}_{D_A}\| \cap \|B\| = \varnothing$, contrary to what was assumed. So $O^F(A/B) \in \mathit{\Delta}$.

Coincidence for O^S:

Case $B = \bot$**:** If $B = \bot$, then in the case of $DDL^{\{F^+,S\}}$, $O^S(\top/\bot) \in \varDelta$ holds due to (DN^S), so $O^S(A/\bot) \in \varDelta$ due to (CExt^S). Also, if $B = \bot$, then for any $P \in \mathbb{P}$, $best_P(\|B\|) = \varnothing$, so for all $P \in \mathbb{P} : best_P(\|B\|) \subseteq \|A\|$ holds for any A, and both sides of the iff-clause are true, as is the iff-clause. – In the case of $DDL^{\{F,S^-\}}$, if $B = \bot$, then $P^S(\top/\bot) \in \varDelta$ due to (DP^S), so $P^S(\neg A/\bot) \in \varDelta$ due to (CExt^S), and by definition of $\varDelta$, $O^S(A/\bot) \notin \varDelta$. Also, if $B = \bot$, then $\|A \wedge B\| = \varnothing$, and since $\mathbb{P} \neq \varnothing$ there is some P for which it is false that $\exists v \in \|A \wedge B\| : \forall v' \in \|A \wedge \neg B\|$: not $v'Pv$, so it is not true for all P. So both sides of the iff-clause are false, and the clause true.

Case $B \neq \bot$**:** *Right-to-left:* Assume $O^S(A/B) \in \varDelta$, and for r.a.a. $\mathbb{P} \not\models O^S(A/B)$, so by (S8) $\exists P \in \mathbb{P}$: $best_P(\|B\|) \cap \|\neg A\| \neq \varnothing$. By construction there is some $ch = \langle D_1, ..., D_n\rangle$ such that $P = P_{ch}$, $ch \in$ *F-Chain*$(\bot)$ or $ch \in$ *S-Chain*$(\mathcal{C}, \bot)$ for some $\mathcal{C} \in \mathbb{C}$. Since $B \neq \bot$, by (S5) there is some 'smallest B-permitting sphere' D_B in ch, with $best_P(\|B\|) = \|\mathcal{O}^{ch}_{D_B}\| \cap \|B\|$ and $\|\mathcal{O}^{ch}_{D_B}\| \cap \|B\| \neq \varnothing$. Either $D_B \in \mathbb{C}$ and $\mathcal{O}^{ch}_{D_B} \in \mathbb{O}^F_{D_B}$, or $\mathcal{O}^{ch}_{D_B} = \mathcal{O}^S_{D_B}$: In both cases, since $\vdash_{PL} B \rightarrow D_B$ by (S6), we have $O^S(B \rightarrow A/D_B) \in \varDelta$ from $O^S(A/B) \in \varDelta$ and (Cond^S), and $\{\mathcal{O}^{ch}_{D_B}\} \vdash_{PL} B \rightarrow A$ either by (DC^{SF}) and minimality of $\mathcal{O}^{ch}_{D_B} \in \mathbb{O}^F_{D_B}$, or by definition of $\mathcal{O}^S_{D_B}$. So $\|\mathcal{O}^{ch}_{D_B}\| \cap \|B\| = best_P(\|B\|) \subseteq \|A\|$, which completes the r.a.a.

Left-to-right: Suppose $\mathbb{P} \models O^S(A/B)$, $O^S(A/B) \notin \varDelta$, so $P^S(\neg A/B) \in \varDelta$. $B \neq \bot$, so $\mathcal{C}_B \in \mathbb{C}$ by (C4), so let $ch = \langle D_1, ..., D_n\rangle$ be in *S-Chain*$(\mathcal{C}_B, \bot)$. Then $D_B = \mathcal{C}_B$: Suppose $D_B = D_i$ and $\mathcal{C}_B = D_j$ for $j < i$. By construction of ch, $\mathcal{O}^{ch}_{D_j} = \mathcal{O}^S_{\mathcal{C}_B}$, and $\mathcal{O}^S_{\mathcal{C}_B} \wedge B \neq \bot$ for otherwise $\mathcal{O}^S_B = \bot$ by (C6), which with (DP-R^S) derives $B = \bot$, but this was excluded. So $\|\mathcal{O}^S_{\mathcal{C}_B}\| \cap \|B\| \neq \varnothing$, so D_B is not the 'smallest B-permitting sphere'. Suppose $D_B = D_i$ and $\mathcal{C}_B = D_j$ for $j > i$, then $\|\mathcal{C}_B\| \cap \|\mathcal{O}^{ch}_{D_B}\| = \varnothing$ by (S2), as $\vdash_{PL} B \rightarrow \mathcal{C}_B$ by (C4^F), $\|B\| \cap \|\mathcal{O}^{ch}_{D_B}\} = \varnothing$, and again D_B is not the 'smallest B-permitting sphere'. So $D_B = \mathcal{C}_B$ and by construction of ch, $\mathcal{O}^{ch}_{D_B} = \mathcal{O}^S_{\mathcal{C}_B}$. Since $P^S(\neg A/B) \in \varDelta$, $\nvdash_{PL} \mathcal{O}^S_B \rightarrow A$ follows from (B1) and construction of $\varDelta$. With (C5), (S7) we get $\|\mathcal{O}^S_B\| = \|\mathcal{O}^S_{\mathcal{C}_B} \wedge B\| = \|\mathcal{O}^{ch}_{D_B}\| \cap \|B\| = best_{P_{ch}}(\|B\|)$. So $best_{P_{ch}}(\|B\|) \nsubseteq \|\neg A\|$ and by (S8) $\mathbb{P} \not\models O^S(A/B)$, contradicting the assumption.

7 The Puzzle Is Still Incomplete

From a complete picture of dyadic deontic reasoning about conflicting imperatives, at least two pieces are still missing. The first is that an imperative itself may be conditional in a way irreducible to a material implication in its content: e.g. if I'm to throw rice as the wedding party leaves the church, but Huey, Dewey and Louie have stolen the bag, blocking the doors won't garner me any praise. It has been argued that such conditional imperatives have two associated propositions, the antecedent and the consequent, and that obligations are only 'triggered', if the antecedents hold, thus providing the opportunity for norm satisfaction or violation. Secondly, even though weighing out the relevant factors may not always produce an unequivocal result, imperatives *can* be ordered by rank of the issuing authority or normative weight: e.g. finding the victim of an accident on my way to a crucial appointment, it seems clear what my obligations are and to not be the time for skeptical or credulous reasoning.

In an attempt to tackle these complexities, Horty [21] proposed the following definition of the imperatives 'binding' in some circumstances A:

$$\begin{aligned} \mathit{Binding}_{(\mathscr{I},<)}(A) =_{def} \{ i \in \mathscr{I} \mid\ & (1)\ i \in \mathit{Triggered}_{\mathscr{I}}(A), \\ & (2)\ \text{there is no } j \in \mathit{Triggered}_{\mathscr{I}}(A) \text{ such that} \\ & \quad (a)\ i < j, \text{ and} \\ & \quad (b)\ \{\mathit{consequent}(i), \mathit{consequent}(j)\} \vdash_{BL} \bot \ \} \end{aligned}$$

where $\mathit{Triggered}_{\mathscr{I}}(A) =_{def} \{i \in \mathscr{I} \mid A \vdash_{BL} \mathit{antecedent}(i)\}$, and $<$ is some strict partial order on $\mathscr{I}$, $i < j$ meaning that i ranks higher than j. The truth of the (skeptical or credulous) dyadic deontic formula $O^*(B/A)$ is then defined with respect to the set of consequents of the imperatives in $\mathit{Binding}_{(\mathscr{I},<)}(A)$.

Yet Horty's proposal is problematic for several reasons. First, the triggering condition does not capture all senses in which antecedents may hold. Consider the situation $(C \vee D)$ and let $\mathscr{I} = \langle C \Rightarrow A, D \Rightarrow A, !(\neg A \wedge B)\rangle$, where for short $A \Rightarrow B$ means the imperative with $\mathit{antecedent}(A)$ and $\mathit{consequent}(B)$, $!(\neg A \wedge B)$ is the unconditional imperative $\top \Rightarrow (\neg A \wedge B)$, and the sequence represents the ordering $<$. Though we do not know which imperative overrides the weakest imperative $!(\neg A \wedge B)$, we know for sure that it is overridden in these circumstances and so should not be included in $\mathit{Binding}_{(\mathscr{I},<)}(C \vee D)$, but with Horty's definition it is. I suggest that, for a better definition of the set $\mathit{Triggered}_{\mathscr{I}}(C \vee D)$, we need an operation like Makinson and van der Torre's [32] 'basic output', which is expressly tailored to process such disjunctive inputs (triggering conditions) intelligibly. But directly applying their construction seems difficult, since it would also close the set of the relevant imperatives' consequents under consequences, and this is hardly an option when conflicts are allowed.

Secondly, the inconsistency check seems both too rigid and not rigid enough. For the latter, let $\mathscr{I}_1 = \langle !(A \wedge \neg B), !(B \wedge C)\rangle$ and $\mathscr{I}_2 = \langle !((A \wedge \neg B) \vee D), !\neg D, !(B \wedge C)\rangle$: in both cases more important imperatives are in conflict with the weakest, but it is rejected only in the first. For the former, let $\mathscr{I} = \langle C \Rightarrow \neg D, C \Rightarrow (B \wedge D)\rangle$ and the situation be $(C \wedge D)$. $C \Rightarrow (B \wedge D)\rangle$ is not in $\mathit{Binding}_{(\mathscr{I},<)}(C \wedge D)$, its consequent contradicting that of a more important imperative. But this has become unfulfillable, which intuitively clears the way for obligatoriness of B. For a solution, I propose to leave inconsistency checks entirely to the (credulous or skeptical) reasoning strategy defined via sets consistent with the circumstances C: let each of these include a maximally C-consistent subset of the most important triggered imperatives' consequents, a maximal subset of the second most important triggered imperatives' consequents that can be C-consistently added to the former, etc. This is the incremental maximizing employed for belief revision by Brewka [6] and Nebel [34] (to work, $<$ must be well-founded). Drawing on a parallel result by Rott ([37] th. 7), my conjecture is that as long as conflicts between incomparable or equally important imperatives are allowed, the logic for accordingly defined deontic operators will still be $DDL^{\{F,S\}}$.

References

1. Alchourrón, C. E. and Bulygin, E., "Unvollständigkeit, Widersprüchlichkeit und Unbestimmtheit der Normenordnungen", in: Conte, A. G., Hilpinen, R. and von Wright, G. H. (eds.), *Deontische Logik und Semantik*, Wiesbaden: Athenaion, 1977, 20–32.
2. Bochman, A., "Credulous Nonmonotonic Inference", in: Dean, T. (ed.), *Proceedings of the 16th International Joint Conference on Artificial Intelligence, IJCAI '99, Stockholm, August 1999*, San Francisco: Morgan Kaufmann, 1999, 30 – 35.
3. Bochman, A., *A Logical Theory of Nonmonotonic Inference and Belief Change*, Berlin: Springer, 2001.
4. Bochman, A., "Brave Nonmonotonic Inference and Its Kinds", *Annals of Mathematics and Artificial Intelligence*, **39**, 2003, 101–121.
5. Brass, S., "On the Semantics of Supernormal Defaults", in: Bajcsy, R. (ed.), *Proceedings of the 13th International Joint Conference on Artificial Intelligence, IJCAI '93*, New York: Morgan Kaufmann, 1993, 578 – 583.
6. Brewka, G., "Belief Revision in a Framework for Default Reasoning", in: Fuhrmann, A. and Morreau, M. (eds.), *The Logic of Theory Change, Workshop, Konstanz, October 13-15, 1989*, LNAI 465, Berlin: Springer, 1991, 206–222.
7. Brink, D. O., "Moral Conflict and Its Structure", *Philosophical Review*, **103**, 1994, 215–247.
8. Donagan, A., "Consistency in Rationalist Moral Systems", *Journal of Philosophy*, **81**, 1984, 291–309.
9. Føllesdal, D. and Hilpinen, R., "Deontic Logic: An Introduction", in [19] 1–35.
10. Forget, L., Risch, V. and Siegel, P., "Preferential Logics are X-logics", *Journal of Logic and Computation*, **11**, 2001, 71–83.
11. van Fraassen, B., "Values and the Heart's Command", *Journal of Philosophy*, **70**, 1973, 5–19.
12. Goble, L., "Multiplex Semantics for Deontic Logics", *Nordic Journal of Philosophical Logic*, **5**, 2000, 113–134.
13. Goble, L., "Preference Semantics for Deontic Logics: Part I – Simple Models", *Logique & Analyse*, **46**, 2003, *forthcoming*.
14. Goble, L., "Preference Semantics for Deontic Logics: Part II – Multiplex Models", *Logique & Analyse*, **47**, 2004, *forthcoming*.
15. Goble, L., "A Logic for Deontic Dilemmas", *Journal of Applied Logic*, **3**, 2005, *in this volume*.
16. Hansen, J., "Sets, Sentences, and Some Logics about Imperatives", *Fundamenta Informaticae*, **48**, 2001, 205–226.
17. Hansen, J., "Problems and Results for Logics about Imperatives", *Journal of Applied Logic*, **2**, 2004, 39–61.
18. Hansson, B., "An Analysis of Some Deontic Logics", *Nôus*, **3**, 1969, 373–398, reprinted in [19] 121–147.
19. Hilpinen, R., *Deontic Logic: Introductory and Systematic Readings*, Dordrecht: Reidel, 1971.
20. Horty, J. F., "Nonmonotonic Foundations for Deontic Logic", in: Nute, D. (ed.), *Defeasible Deontic Logic*, Dordrecht: Kluwer, 1997, 17 – 44.
21. Horty, J. F., "Reasoning with Moral Conflicts", *Noûs*, **37**, 2003, 557–605.
22. Jacquette, D., "Moral Dilemmas, Disjunctive Obligations, and Kant's Principle that 'Ought' implies 'Can"', *Synthese*, **88**, 1991, 43–55.
23. Kelsen, H., *Reine Rechtslehre*, 2nd ed., Wien: Deuticke, 1960.

24. Kelsen, H., *Allgemeine Theorie der Normen*, Wien: Manz, 1979.
25. Kratzer, A., "Conditional Necessity and Possibility", in: Bäuerle, R. (ed.), *Semantics from Different Points of View*, Berlin: Springer, 1979, 116–147.
26. Kraus, S., Lehmann, D. and Magidor, M., "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics", *Artificial Intelligence*, **44**, 1990, 167–207.
27. Lemmon, E. J., "Moral Dilemmas", *Philosophical Review*, **71**, 1962, 139–158.
28. Lewis, D., *Counterfactuals*, Oxford: Basil Blackwell, 1973.
29. Lewis, D., "Semantic Analyses for Dyadic Deontic Logic", in: Stenlund, S. (ed.), *Logical Theory and Semantic Analysis*, Dordrecht: Reidel, 1974, 1 – 14.
30. Lewis, D., "Ordering Semantics and Premise Semantics for Counterfactuals", *Journal of Philosophical Logic*, **10**, 1981, 217–234.
31. Makinson, D., "Bridges between Classical and Nonmonotonic Logic", *Logic Journal of the IGPL*, **11**, 2003, 69–96.
32. Makinson, D. and van der Torre, L., "Input/Output Logics", *Journal of Philosophical Logic*, **29**, 2000, 383–408.
33. Marcus, R. B., "Moral Dilemmas and Consistency", *Journal of Philosophy*, **77**, 1980, 121–136.
34. Nebel, B., "Syntax-Based Approaches to Belief Revision", in: Gärdenfors, P. (ed.), *Belief Revision*, Cambridge: University Press, 1992, 52–88.
35. Poole, D., "A Logical Framework for Default Reasoning", *Artificial Intelligence*, **36**, 1988, 27–47.
36. Rescher, N., "An Axiom System for Deontic Logic", *Philosophical Studies*, **9**, 1958, 24–30.
37. Rott, H., "Belief Contraction in the Context of the General Theory of Rational Choice", *Journal of Symbolic Logic*, **58**, 1993, 1426–1450.
38. Siegel, P. and Forget, L., "A Representation Theorem for Preferential Logics", in: Aiello, L. C., Doyle, J. and Shapiro, S. C. (eds.), *Principles of Knowledge Representation and Reasoning (Proceedings of the 5th International Conference, KR 96)*, San Francisco: Morgan Kaufmann, 1996, 453–460.
39. Spohn, W., "An Analysis of Hansson's Dyadic Deontic Logic", *Journal of Philosophical Logic*, **4**, 1975, 237–252.
40. Stenius, E., "The Principles of a Logic of Normative Systems", *Acta Philosophica Fennica*, **16**, 1963, 247–260.
41. van der Torre, L., *Reasoning About Obligations*, Amsterdam: Thesis Publishers, 1997.
42. van der Torre, L. and Tan, Y.-H., "Two-Phase Deontic Logic", *Logique & Analyse*, **43**, 2000, 411–456.
43. Williams, B., "Ethical Consistency", *Proceedings of the Aristotelian Society, supp*, **39**, 1965, 103–124.
44. von Wright, G. H., "A Note on Deontic Logic and Derived Obligation", *Mind*, **65**, 1956, 507–509.
45. von Wright, G. H., *Norm and Action*, London: Routledge & Kegan Paul, 1963.
46. von Wright, G. H., "A New System of Deontic Logic", *Danish Yearbook of Philosophy*, **1**, 1964, 173–182, reprinted in [19] 105–115.
47. von Wright, G. H., "A Correction to a New System of Deontic Logic", *Danish Yearbook of Philosophy*, **2**, 1965, 103–107, reprinted in [19] 115–120.
48. von Wright, G. H., *An Essay in Deontic Logic and the General Theory of Action*, Amsterdam: North Holland, 1968.
49. von Wright, G. H., "Deontic Logic – as I See it", in: McNamara, P. and Prakken, H. (eds.), *Norms, Logics and Information Systems*, Amsterdam: IOS, 1999, 15 – 25.

Artificial Intelligence and Law (2006)
DOI 10.1007/s10506-005-5081-x

Deontic logics for prioritized imperatives

JÖRG HANSEN
Institut für Philosophie, Universität Leipzig, Beethovenstraße 15, D-04107, Leipzig, Germany
e-mail: jhansen@uni-leipzig.de

Abstract. When a conflict of duties arises, a resolution is often sought by use of an ordering of priority or importance. This paper examines how such a conflict resolution works, compares mechanisms that have been proposed in the literature, and gives preference to one developed by Brewka and Nebel. I distinguish between two cases – that some conflicts may remain unresolved, and that a priority ordering can be determined that resolves all – and provide semantics and axiomatic systems for accordingly defined dyadic deontic operators.

Keywords: deontic logic, logic of imperatives, priorities

1. Introduction

W. D. Ross (1930) argued that whenever there appears to be a conflict of duties, through careful study of all aspects of the situation one will arrive at the conclusion – or rather: the considered opinion – that one of these duties is "more pressing" than others, and this duty is then one's duty *sans phrase*, whereas the others were *prima facie* only. Ross gives the following example:

EXAMPLE (The road accident). *"If I have promised to meet a friend at a particular time for some trivial purpose, I should certainly think myself justified in breaking my engagement if by doing so I could prevent a serious accident or bring relief to the victims of one."*

There are two conflicting obligations: to keep the promise, and to prevent the accident or help its victims. The second takes priority: it is in these circumstances "more of a duty" than keeping the appointment.

While in the example the determination of the priority ordering seems to rely on a comparison of the outcomes of satisfying or violating the conflicting duties under considerations of utility and possible harm, in the case of legal obligations or individual imperatives the ordering can often be directly obtained from the norm's position in a normative hierarchy or the rank of the issuing authority. These factors may also relate to each other, e.g. when the decision of a commander in the field overrules that of her superior due to some unforeseen danger or opportunity. I will leave aside the question of how a particular ordering is determined, and also not address Ross's notoriously problematic distinction between *a prima facie* duty and a duty *sans phrase*. What interests me here is rather how a conflict resolution based on (established) priorities works, i.e. what the resolution mechanism looks like, or should look like, when an ordering of priority or importance of possibly conflicting norms is assumed. Section 2 introduces the formal framework and explains how it is used to define deontic operators. After pointing out counterintuitive results of a conflict resolution based on a method by Horty (2003), I show in Section 3 that a method developed for the resolution of inconsistencies in prioritized theory bases by Brewka (1989, 1991) and Nebel (1991, 1992) fares better (Section 3). A broader comparison includes ordering based mechanisms by Alchourrón and Makinson (1981), Gärdenfors (1984), Alchourrón (1986), and variants (Section 4). Section 5 explores what a priority ordering must be like to resolve *all* possible conflicts, and provides a sound and weakly complete axiomatic system (which readers might find familiar) for a corresponding dyadic deontic operator. All formal proofs are delegated to the Appendix. Section 6 concludes.

2. Imperative semantics and deontic logic

When a conflict between norms is resolved by an appeal to some priority ordering, I assume that what is thus conceived as ordered are the norms themselves, though their ordering may reflect a ranking of their sources, or an axiological order of the states realized when fulfilling the norms. So for a logical analysis, some formal representation of norms is required. I only consider unconditional imperatives,[1] like "Invite Jones to dinner!", and I is a set of such imperatives. To each imperative corresponds a descriptive sentence like "You invite Jones to dinner," which – grammatically similar, but in the indicative, not the imperative mood – describes what must be the case if and only if (iff) the imperative is satisfied. Any such descriptive sentence is assumed to have a formalization in the language of a basic logic, which I let be propositional logic PL.[2] A function $f\colon\ I \rightarrow \mathscr{L}_{PL}$ assigns every imperative in I the $\mathscr{L}_{PL}$-formalization of its corresponding descriptive sentence, and the tuple $\langle I, f\rangle$ is called a *basic imperative structure*. I write $!A$ for an i in I with $f(i) = A$, and use the superscripted i^f, Γ^f instead of $f(i)$, $f(\Gamma)$ for better readability. In analogue to the usual concept of remainders, let $I \curlywedge A$ be the maximal sets of imperatives such that the sets of corresponding descriptive sentences do not derive A (I also call these "A-remainders" of I), i.e. $I \curlywedge A$ contains all $\Gamma \subseteq I$ such that (i) $\Gamma^f \nvdash_{PL} A$ and (ii) there is no $\Delta \subseteq I : \Gamma \subset \Delta$ and $\Delta^f \nvdash_{PL} A$.

In the 'imperativist tradition' of deontic logic, authors used such semantics to interpret deontic formulas, rather than employing the usual possible worlds semantics.[3] Let deontic formulas be those of a language $\mathscr{L}_{DL}$, based on an alphabet like the one for $\mathscr{L}_{PL}$, except that it additionally contains the operator symbol 'O', whereby OA formalizes the (true or false) statement that what A describes is obligatory. $\mathscr{L}_{DL}$ is then the smallest set such that

(a) for all $A \in \mathscr{L}_{PL}, OA \in \mathscr{L}_{DL}$,
(b) if $A, B \in \mathscr{L}_{DL}$, so are $\neg A, (A \wedge B), (A \vee B), (A \rightarrow B), (A \leftrightarrow B)$.

Interpretations of Boolean operators being as usual, the truth definition

(td-1) $\langle I, f \rangle \models OA$ iff $I^f \vdash_{PL} A$.

defines a normal modal logic, i.e. the set of $\mathscr{L}_{DL}$-sentences defined as true for all tuples $\langle I, f \rangle$ equals the axiomatically defined set that contains all $\mathscr{L}_{DL}$-instances into tautologies, furthermore all $\mathscr{L}_{PL}$-instances into

(Ext) If $\vdash_{PL} A \leftrightarrow B$, then $OA \leftrightarrow OB$ is in the set.
(M) $O(A \wedge B) \rightarrow (OA \wedge OB)$
(C) $(OA \wedge OB) \rightarrow O(A \wedge B)$
(N) $O\top$

and is closed under *modus ponens*. Furthermore, the above truth definition defines standard deontic logic *SDL*, which adds the "deontic" scheme (D):

(D) $OA \rightarrow PA$

iff $\langle I, f \rangle$ is required to be such that I^f is consistent (as usual, PA abbreviates $\neg O \neg A$). Requiring I^f to be consistent excludes conflicts between imperatives and is thus a severe and in this case unwanted restriction, for to show how conflicts are resolved they must first be semantically modeled. But if e.g. two imperatives $!p_1$ and $!(p_2 \wedge \neg p_1)$ can both be in I, then not only does (D) fail, but also (td-1) is not very useful, making OA true for any $A \in \mathscr{L}_{PL}$. Instead, the following definition for a "disjunctive" ought operator was put forward:[4]

(td-2) $\langle I, f \rangle \models OA$ iff $\forall \Gamma \in I \curlywedge \bot : \Gamma^f \vdash_{PL} A$.

So OA is true when all maximally consistent subsets of what the imperatives demand derive A. It is apparent that the definition tolerates conflicts and e.g. if $!p_1$ and $!(p_2 \wedge \neg p_1)$ are in I, then $O(p_1 \vee p_2)$, but not $O\bot$ is true. Moreover, this solution is easily adapted to the dyadic case and the related problem of dilemmas. Dyadic deontic logic uses a language $\mathscr{L}_{DDL}$ that employs the additional auxiliary sign "/" and is like $\mathscr{L}_{DL}$, except that clause (a) now reads

(a) for all $A, C \in \mathscr{L}_{PL}, O(A/C) \in \mathscr{L}_{DDL}$,

where $O(A/C)$ is read as "A is obligatory in the circumstances characterized by C". The truth definitions for dyadic deontic formulas should allow some influence of the circumstances, so e.g. if John must either not impregnate Suzy Mae or marry her, and she is in fact pregnant by him, then marrying her seems to be what he must do. But simply putting

(td-3) $\langle I, f\rangle \vDash O(A/C) \quad \text{iff} \quad I^f \cup \{C\} \vdash_{PL} A.$

will not suffice if subjects can get (themselves) into dilemmas, i.e. situations where the norms are collectively satisfiable at the outset, but due to misfortune or failure they cannot all be satisfied anymore. To handle such situations, and to e.g. prevent the derivation of $O(\bot/p_1)$ when I contains $!(\neg p_1 \vee p_2)$ and $!(\neg p_2 \wedge p_3)$, the truth definition for a "disjunctive" dyadic ought operator can be given as:

(td-4) $\langle I, f\rangle \models O(A/C) \quad \text{iff} \quad \forall \Gamma \in I \lambda \neg C : \Gamma^f \cup \{C\} \vdash_{PL} A.$

So A is obligatory in the situation described by C if A is what the imperatives in any $\neg C$-remainder demand, given C. With usual truth conditions for Boolean operators, this semantics has a sound and (weakly) complete axiom system PD defined as containing all $\mathscr{L}_{\mathrm{DDL}}$-instances into tautologies, all $\mathscr{L}_{PL}$-instances into

(CExt) If $\vdash_{PL} C \rightarrow (A \leftrightarrow B)$ then $\vdash_{PD} O(A/C) \leftrightarrow O(B/C)$

(ExtC) If $\vdash_{PL} C \leftrightarrow D$ then $\vdash_{PD} O(A/C) \leftrightarrow O(A/D)$

(DM) $O(A \wedge B/C) \rightarrow (O(A/C) \wedge O(B/C))$

(DC) $O(A/C) \wedge O(B/C) \rightarrow O(A \wedge B/C)$

(DN) $O(\top/C)$

(DD-R) If $\nvdash_{PL} \neg C$ then $\vdash_{PD} O(A/C) \rightarrow P(A/C)$

(Cond) $O(A/C \wedge D) \rightarrow O(D \rightarrow A/C)$

(CCMon) $O(A \wedge D/C) \rightarrow O(A/C \wedge D)$

and closed under *modus ponens*.[5] PD resembles the system P defined by Kraus et al. (1990) with the (restricted) dyadic "deontic" scheme (DD-R) added, hence the name.

For the present purposes, I define *a prioritized imperative structure* to be a tuple $\langle I, f, <\rangle$ that is like a basic imperative structure, except that it additionally includes an ordering relation $<$ on I, where the formal properties of this relation are for the moment left open. Unfortunately, authors disagree on the direction in which '$i_1 < i_2$' is to be read, if it means that i_1 takes priority over i_2 (reading $<$ like a preference relation), or that i_1 is less important than i_2 (reading $<$ like a utility function). I assume the former, and adapt differing definitions to this convention, so e.g. a tuple $i_1 < i_2 < i_3 \ldots$ is read like a list that starts with what is most important. For any ordering $<$ on some set Γ, I define $min_{<}\Gamma = \{i \in \Gamma \mid \forall i' \in \Gamma : \text{if } i' \neq i, \text{ then } i' \not< i\}$, so $min_{<}I$ is the set of the highest ranking, most important, etc. imperatives, and $max_{<}\Gamma = \{i \in \Gamma \mid \forall i' \in \Gamma : \text{if } i' \neq i, \text{ then } i \not< i'\}$, so $max_{<}I$ are the imperatives that come last, are least important, rank lowest, etc.

3. Reasoning with prioritized imperatives

As explained by Ross, in the case of a normative conflict one proceeds by examining the situation for clues to an ordering of the obligations involved, e.g. by considering the rank of the issuing authority, notions of urgency or a gross difference in the utilities of the outcomes. The example of the road accident illustrates that the disjunctive ought operator defined in the previous section, which pays no attention to priorities, produces inadequate results:

EXAMPLE (The road accident: disjunctive reasoning). *Let A be helping the accident victims, B keeping the promise, and T a conjunction of actual necessities, including the agent's present physical and psychical capabilities (I write $\bot$ for $\neg T$). An imperative interpretation produces $I = \{!A, !B\}$ and $\vdash_{PL} T \rightarrow (A \rightarrow \neg B)$ as helping causes me to miss the meeting. $I^f \vdash_{PL} \bot$, so* (td-1) *makes $O\bot$ true and the impossible obligatory, so it is not very useful. $I \curlywedge \bot = \{\{!A\}, \{!B\}\}$, so* (td-4) *makes $O(A \vee B/T)$ true but $O(A/T)$ false, as $\{B, T\} \nvdash_{PL} A$. So there is only a disjunctive obligation to help or proceed to the meeting. But intuitively, helping takes priority over anything else.*

The situation looks like a conflict: there exist requirements which cannot all be satisfied. But the conflict is avoided by (so far intuitively) giving priority to the norm of greater weight. Note that if symmetrical or incomparable obligations are not ruled out, then a demand that takes priority can not only dissolve a dilemma, but also create a conflict for an otherwise conflict-free situation:

EXAMPLE (The road accident II). *It is Tuesday afternoon, and like on all Tuesdays, Mirjam must fetch her grandmother from the day care center before it closes at 6:30. Today, Mirjam was also asked by her boss to bring the office mail to the post office after hours, which also closes at 6:30, but lies in the opposite direction. However, when she told of her other duty, she was allowed to leave early. Driving at 5:30 in the direction of the post office, Mirjam* *becomes involved in a traffic accident. The law requires her to stay at the accident site until the police have recorded it, which won't happen before 6:00. Then she can only get to one place, the post office or the day care center, on time. The law takes priority over her other duties, but a ranking of these is not obvious; in particular it is difficult to say which violation could have worse consequences, and Mirjam will have a hard time making up her mind.*

To formalize the reasoning about priorities when faced with conflicting demands, Horty (2003) proposed that the priority ordering is used to first determine a set of "binding imperatives" in the set of all imperatives:

DEFINITION 1 (Binding imperatives). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure. Then*

$$Binding = \{i \in I \mid \neg\exists j \in I : j < i \text{ and } \vdash_{PL} (i^f \wedge j^f) \rightarrow \bot\}.$$

So an imperative is "binding" if there is no higher ranking imperative with a materially inconsistent demand (cf. Horty 2003, p. 560). If it is also higher-ranking than any such imperative, Horty calls it "overriding". *Binding*, instead of *I*, is then used to define a disjunctive dyadic ought operator:

$$\text{(td-5)} \quad \langle I, f, < \rangle \models O(A/C) \quad \text{iff} \quad \forall \Gamma \in Binding \mathbin{⅄} \neg C : \Gamma^f \cup \{C\} \vdash_{PL} A$$

I examine how this definition[6] copes with the examples and variants:

EXAMPLE (The road accident: Horty's solution). *A is helping the accident victims, and B keeping the promise. So $I = \{!A, !B\}$ models the logical situation, where $\vdash_{PL} T \rightarrow (A \rightarrow \neg B)$, as the situation excludes both helping at the accident site and meeting my friend. The ordering is $!A < !B$, so $!A$ overrides $!B$ and Binding $= \{!A\}$. (td-5) makes $O(A/T), O(\neg B/T)$ and $P(\neg B/T)$ true, so helping is obligatory, keeping the appointment forbidden, and not keeping the appointment permitted, which is as it should be.*

EXAMPLE (The road accident II: Horty's solution). *Let A be Mirjam's staying at the accident site, B taking her boss's mail to the post office, and C fetching her grandmother from the day care center. $I = \{!A, !B, !C\}$ is the logical model of the normative situation. Mirjam left early enough to get to both the post office and the day care center in time, so $\nvdash_{PL} T \rightarrow (B \rightarrow \neg C)$, but $\vdash_{PL} T \rightarrow (A \rightarrow \neg(B \wedge C))$ as waiting excludes her accomplishing both. The legal obligation takes priority, so $!A < !B$ and $!A < !C$, while the ranking between $!B$ and $!C$ unclear. Binding $= I$, as the truth of A only excludes satisfying both $!B$ and $!C$, but getting to one place remains possible. Binding $\mathbin{⅄} \bot = \{\{!A, !B\}, \{!A, !C\}, \{!B, !C\}\}$, (td-5) making*

$O(A/T)$ false as $\{T, B, C\} \nvdash_{PL} A$, and only $O((A \wedge (B \vee C)) \vee (B \wedge C)/T)$ true, so it seems Mirjam can choose to wait or go on driving to her destinations, which is counterintuitive.

EXAMPLE (The road accident III, with Horty's solution). *Things are as in variant II, but suppose some grave danger arises from Mirjam's not being at the day care center before it closes, as the disturbed old lady will wander off on her own and may fall or get lost. Hence, fetching her is much more urgent than posting the letters, of which the important ones were very likely faxed beforehand. Fetching her grandma may be even more important than waiting for the police, but Mirjam is not sure about that and can do both anyway. So beside $!A < !B$ we have $!C < !B$, with the ranking between $!A$ and $!C$ unclear. Intuitively, Mirjam must stay until the police are finished and then fetch her grandma. But as it is two higher-ranking imperatives that exclude, if satisfied, the satisfaction of the lower-ranking one, still Binding = $\{!A, !B, !C\} = I$, making $O(A/T)$ false and even $P(B \wedge \neg C/T)$ true even though $!C$ ranks higher than $!B$.*

EXAMPLE (The road accident IV, with Horty's solution). *Mirjam did not dare ask her boss for permission to leave early, sneaking out at 5:45 instead, but that was too late to get to both places in time, i.e. $\vdash_{PL} T \rightarrow (C \rightarrow \neg B)$. Again, Mirjam gets involved in a traffic accident and is required to wait for the police. Fetching her grandmother takes priority over posting the mail, so $I = \{!A, !B, !C\}$, $!C < !B$ and $!A < !B$. Imagine the accident left her car a wreck, making it impossible to get to the day care center in time, but when the police finish around 6:15 she can still get to the post office. Let S be this situation (including T), so $\vdash_{PL} S \rightarrow \neg C$. Binding = $\{!A, !C\}$, as $!B$ is not reinstated when satisfying $!C$ is excluded, so Binding $\curlywedge \neg S = \{\{!A\}\}$, making $P(\neg B/S)$ true. But it is hard to see why Mirjam should not have to post the letters.*

EXAMPLE (The road accident V, with Horty's solution). *As in variant IV, Mirjam left too late to make it to both the post office and the day care center on time, so $\vdash_{PL} T \rightarrow (C \rightarrow \neg B)$. Again, fetching grandma takes priority over posting the letters, i.e. $!C < !B$. Suppose it is not the damage, but the time required by the police that makes it impossible to get to the day care center on time (it is too far from the accident site, while the post office is just a block away), so $\vdash_{PL} T \rightarrow (A \rightarrow \neg C)$. Making up her mind, Mirjam decides that the legal obligation to wait for the police probably takes priority over her familial duty, i.e. $!A < !C < !B$. Both $!B$ and $!C$ are overridden by higher ranking imperatives and so are not in Binding $\curlywedge \neg S = \{\{!A\}\}$, making $P(\neg B/S)$ true. But again it is hard to see why Mirjam is relieved from posting the letters.*

Thus Horty's set *Binding* solves simple cases, but is not adequate for complex hierarchies where more than two imperatives may be in conflict, and it makes life too easy when conflicting higher ranking imperatives become unfulfillable or are themselves overridden.[7] To overcome these difficulties when formalizing that disregard for a lower ranking imperative can (only) be excused by obedience to higher-ranking ones (and not *vice versa*), I suggest that neither remainder sets of I, nor of a fixed subset *Binding*, but an "incremental" maximizing strategy should be used. For a situation C, the relevant sets are constructed by first adding a maximal set of the most important imperatives such that their demands do not derive $\neg C$, then adding a maximal subset of the second most important imperatives that can be added without the corresponding demands now deriving $\neg C$, etc. Introduced by Rescher (1964, p. 50), such incremental maximizing was more rigorously defined and employed for the purpose of theory revision by Brewka (1989, 1991) and Nebel (1991, 1992). Both employ a strict partial order, i.e. $<$ is irreflexive and transitive. Nebel additionally assumes $<$ to be the asymmetric part of a complete preorder $\leq$, i.e. obtained from a reflexive, transitive and connected ordering $\leq$ *via* defining $i < j$ iff $i \leq j$ and $j \not\leq i$. Both agree that $<$ must be well-founded, i.e. infinite descending chains are excluded.[8] For any $<$, Brewka defines a *full prioritization* $\prec$ to be any (strict) well-order on the given set that preserves $<$, i.e. for all i, j: if $i < j$ then $i \prec j$. Clearly:

THEOREM 1 (Existence of full prioritizations). *For every well-founded strict partial order $<$ on a set Γ there is a full prioritization $\prec$, i.e. a strict well-order that is order-preserving with respect to $<$.*

Brewka then defines subsets of the set as 'preferred subtheories'. Calling them *preferred remainders* (they are not theories here), his definition translates thus:[9]

DEFINITION 2 (Brewka's preferred remainders). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure, where $<$ is a well-founded strict partial order on I. Then $\Gamma \subseteq I$ belongs to the preferred remainder set $I \Downarrow A$ iff (i) $\Gamma^f \nvdash_{PL} A$, and (ii) Γ is obtained from a full prioritization $\prec$ by defining*

$$S^A_{[\prec\downarrow i]} = \begin{cases} \bigcup_{j \prec i} S^A_{[\prec\downarrow j]} \cup \{i\} & \textit{if} \, [\bigcup_{j \prec i} S^A_{[\prec\downarrow j]}]^f \cup \{i^f\} \nvdash_{PL} A, \textit{and} \\ \bigcup_{j \prec i} S^A_{[\prec\downarrow j]} & \textit{otherwise,} \end{cases}$$

for any $i \in I$, and letting $S^A_{\prec} = \bigcup_{i \in I} S^A_{[\prec\downarrow i]}$ and $\Gamma = S^A_{\prec}$.

(i) bans the empty set from $I \Downarrow A$ for tautological A, and (ii) recursively defines $S_{[\prec\downarrow i]}$ to include all elements of some such set for a prior element j, adding i if possible without the corresponding set deriving A. Γ is the union of all such sets. I drop superscripts if the meaning is clear. The following is almost immediate:

THEOREM 2 (Preferred remainders are remainders). *Let $\langle I, f, <\rangle$ be a prioritized imperative structure, where $<$ is a well-founded strict partial order on I. Then $I \Downarrow A \subseteq I \curlywedge A$.*

As noted, Nebel's (1992) approach defines $<$ as the asymmetric part of a complete, well-founded preorder $\leq$. For each $i \in I$, the priority class $[i] = \{j \in I \mid i \leq j \text{ and } j \leq i\}$ contains all $j \in I$ of the same $\leq$-priority as i. A preference-ordering $\ll_N$ between all subsets Δ, Γ of I is then defined by letting

$$\Delta \ll_N \Gamma \quad \text{iff} \quad \exists i \in I : \forall j < i : \Gamma \cap [j] = \Delta \cap [j] \text{ and } \Gamma \cap [i] \subset \Delta \cap [i]$$

i.e. by preferring Δ over Γ iff both agree for all priority classes up to some $[i]$, of which Δ contains all elements of Γ plus more. Then choosing a maximally $\ll_N$-preferred set among all $\Gamma \subseteq I$ with $\Gamma^f \nvdash_{PL} A$ equals choosing from $I \Downarrow A$:

THEOREM 3 (Nebel's prioritized removals). *Let $\langle I, f, <\rangle$ be a prioritized imperative structure, where $<$ is the asymmetric part of a complete, well-founded preorder $\leq$ on I. Then $I \Downarrow A$ equals*

$$\{\Gamma \subseteq I \mid \Gamma^f \nvdash_{PL} A \textit{ and } \forall \Delta \subseteq I : \textit{ if } \Delta \ll_N \Gamma \textit{ then } \Delta^f \vdash_{PL} A\}.$$

There is an alternative, non-constructive definition of Brewka's preferred remainders, attributed to Ryan (1992) by Rintanen (1994) and also appearing in Sakama and Inoue (1996): Let $\langle I, f, <\rangle$ be a prioritized imperative structure, where $<$ is a well-founded strict partial order on I, and define $\mathbf{p}\ (I \curlywedge A)$ to be the set

$$\{\Gamma \in I \curlywedge A \mid \exists \prec : \forall \Delta \in I \curlywedge A \setminus \{\Gamma\} : \exists i \in \Gamma \setminus \Delta : \forall j \in \Delta \setminus \Gamma : i \prec j\}.$$

So some A-remainder Γ is in $\mathbf{p}\ (I \curlywedge A)$ iff for some full prioritization $\prec, \Gamma$ contains for any other A-remainder Δ some exclusive element that $\prec$-ranks higher than any element exclusively in Δ. The following holds.

THEOREM 4 (Preferred remainders, after Ryan and Sakama & Inoue). *Let $\langle I, f, <\rangle$ be a prioritized imperative structure, $<$ a well-founded strict partial order on I and $I \Downarrow A$ and $\mathbf{p}(I \curlywedge A)$ be as defined. Then $I \Downarrow A = \mathbf{p}(I \curlywedge A)$.*

Proposing use of Brewka's and Nebel's concept of preferred remainders, based on some strict partial, well-founded ordering $<$, as the resolution mechanism for conflicts between imperatives or dilemmas that arise in certain situations, a disjunctive ought operator can be defined parallel to (td-4) as follows:

(td-6) $\mathscr{I} \models O(A/C)$ iff $\forall \Gamma \in I \Downarrow \neg C : \Gamma^f \cup \{C\} \vdash_{PL} A$.

This truth definition fares better in dealing with the above examples:

EXAMPLE (The road accident: the Brewka/Nebel solution). *A is helping the accident victim, B keeping the promise, $I = \{!A,\ !B\}$ and $\vdash_{PL} T \rightarrow (A \rightarrow \neg B)$. The ordering is $!A\ <\ !B$, being its only full prioritization. The construction of $S_<$ includes $!A$ but rejects $!B$, so $I \Downarrow \bot = \{\{!A\}\}$ and (td-6) makes true $O(A/T), O(\neg B/T)$ and $P(\neg B/T)$, which is as it should be.*

EXAMPLE (The road accident II: the Brewka/Nebel solution). *$I = \{!A, !B, !C\}$. $\vdash_{PL} T \rightarrow (A \rightarrow \neg(B \wedge C))$, as waiting allows Mirjam to get to one place, the post office or the day care center in time, but not both. The ordering is $!A\ <\ !B$ and $!A\ <\ !C$, and for $!B$, $!C$ unclear. Its two full prioritizations are $!A \prec !B \prec !C$ or $!A \prec !C \prec !B$, producing $I \Downarrow \bot = \{\{!A,\ !B\},\ \{!A,\ !C\}\}$. Then (td-6) makes true $O(A \wedge (B \vee C)/T)$, so waiting and then going to one place, the post office or the day care center, is obligatory as it should be.*

EXAMPLE (The road accident III: the Brewka/Nebel solution). Still $\vdash_{PL} T \rightarrow (A \rightarrow \neg(B \wedge C))$, *i.e. waiting excludes getting to both places. Fetching her grandma now takes priority over going to the post office, so $!C\ <\ !B$ and $!A\ <\ !B$, this time the ranking between $!A$, $!C$ being unclear. The two full prioritizations are $!A \prec !C \prec !B$ and $!C \prec !A \prec !B$, so $I \Downarrow \bot = \{\{!A,\ !C\}\}$, and (td-6) makes $O(A \wedge C/T)$ true. So Mirjam must stay at the site until the police are finished with her and then go to fetch her grandmother, as it should be.*

EXAMPLE (The road accident IV: the Brewka/Nebel solution). *Getting to both the post office and the day care center on time was never possible, so $\vdash_{PL} T \rightarrow (C \rightarrow \neg B)$. The car being wrecked, the assumed situation S excludes getting to the day care center on time, so $\vdash_{PL} S \rightarrow \neg C$. Again $!C\ <\ !B$ and $!A\ <\ !B$, with the relation between $!A$ and $!C$ unclear, so $!A \prec !C \prec !B$ and $!C \prec !A \prec !B$ are the full prioritizations, producing $I \Downarrow \neg S = \{\{!A, !B\}\}$. Hence $O(A \wedge B/S)$, i.e. Mirjam must wait and then hurry to the post office, which is as it should be.*

EXAMPLE (The road accident V: the Brewka/Nebel solution). *Again, Mirjam cannot get to both the post office and the day care center on time, so $\vdash_{PL} T \rightarrow (C \rightarrow \neg B)$. Waiting for the police excludes getting to the day care center on time, i.e. $\vdash_{PL} T \rightarrow (A \rightarrow \neg C)$. Mirjam decides that the law* *overrides her familial duty, so* $!A < !C < !B$*, which, being its only full prioritization, yields* $I \Downarrow \bot = \{\{!A, !B\}\}$*. So Mirjam must wait and post the letters, as it should be.*

Let a semantics be called a *prioritized imperative semantics* iff it defines the truth of $\mathscr{L}_{DDL}$-sentences using (td-6), with respect to arbitrary prioritized imperative structures $\langle I, f, <\rangle$. Then it may be surprising – though Rott (1993, Theorem 7) already proved a similar result – that the logical properties of such a semantics are not different from that defining the deontic operator using (td-4), i.e. with respect to basic imperative structures and simple remainders, since the system *PD* remains sound and (only) weakly complete:

THEOREM 5 (Soundness, completeness of *PD*). *PD is sound and (only) weakly complete with respect to prioritized imperative semantics.*

4. Alternative resolution mechanisms

4.1. LEAST EXPOSURE AND ITS VARIANTS

Alchourrón and Makinson (1981) seem to have been the first to logically examine the idea of resolving contradictions in a body of norms, or contradictions that arise from such a body together with some set of true empirical facts, by imposing an order upon that body. The object of their study is a set of regulations that is partially ordered by a relation $\leq$, which does not necessarily stand for an ordering by priority or importance. Rather, $i \leq j$ means that j is as much exposed, or more exposed, to the risk of legislative derogation as i. If a conflict occurs between two parts of the code, or between the code and some empirical facts, the aim is to find a (possibly maximal) non-conflicting subset that is most secure from the changes which the law-giver will presumably enact upon learning of this situation. Their definition translates to the present framework as follows:

DEFINITION 3 (Alchourrón and Makinson's strict exposure). *Let* $\langle I, f, <\rangle$ *be a prioritized imperative structure, where* $<$ *is a strict partial order on* I *(like the asymmetric part of a partial order* $\leq$). Then for all $\Gamma, \Delta \subseteq I$:

$$\Gamma \ll_{AM} \Delta \quad \text{iff} \quad \Delta \neq \emptyset \text{ and } \forall i \in \Gamma : \exists j \in \Delta : i < j.$$

So a subset is strictly less exposed than some other if for any member of the first there is a member of the second which is strictly more exposed. To see how this approach compares to Brewka and Nebel's, consider three cases:

Case 1. Let $\langle I, f, <\rangle$ be $!p_1 < !(p_2 \wedge \neg p_3) < !p_3$, the imperative in the 'middle' conflicting with the lowest. Then we have

- $I \curlywedge \bot = \{\{!p_1, !(p_2 \wedge \neg p_3)\}, \{!p_1, !p_3\}\}$ and
- $I \Downarrow \bot = \{\{!p_1, !(p_2 \wedge \neg p_3)\}\}$.

The exposure criterion yields $\{!p_1, !(p_2 \wedge \neg p_3)\} \ll_{AM} \{!p_1, !p_3\}$: for each left member, a right member is strictly more exposed, namely $!p_3$.

Case 2. Let $\langle I, f, <\rangle$ be $!p_1 < !(\neg p_1 \wedge p_2) < !p_3$, so the "middle" now conflicts with the higher-ranking imperative. Then we have

- $I \curlywedge \bot = \{\{!p_1, !p_3\}, \{!(\neg p_1 \wedge p_2), !p_3\}\}$ and
- $I \Downarrow \bot = \{\{!p_1, !p_3\}\}$.

But $\{!p_1, !p_3\} \not\ll_{AM} \{!(\neg p_1 \wedge p_2), !p_3\}$: from the left set, $!p_3$ is not less exposed than $!(\neg p_1 \wedge p_2)$ from the right. The authors recognize that a conflict between higher-ranking norms excludes lower-ranking norms from a least exposed set and propose to use relevant logic for determining conflicts as a cure (Alchourrón and Makinson 1981, p. 139).

Case 3. Let $\langle I, f, <\rangle$ be $!p_1 < !(\neg p_1 \wedge p_2 \wedge \neg p_3) < !p_3$, the "middle" now in conflict with both ends of the hierarchy (by whatever logic). Then we have

- $I \curlywedge \bot = \{\{!p_1, !p_3\}, \{!(\neg p_1 \wedge p_2 \wedge \neg p_3)\}\}$ and
- $I \Downarrow \bot = \{\{!p_1, !p_3\}\}$.

Yet $\{!(\neg p_1 \wedge p_2 \wedge \neg p_3)\} \ll_{AM} \{!p_1, !p_3\}$: from the right set, $!p_3$ is more exposed than any left member. Mediocrity rules! But even if $!p_3$ is more exposed to legislative change, if that change came about and removed $!p_3$, the right set would still contain a member that ranks higher than any in the left.

Prakken, pursuing an argumentative approach, wants to employ Alchourrón and Makinson's criterion at the heart of his "rebuttal" mechanism used to determine justified arguments (derivations from facts and defaults. But the criterion he presents (Prakken 1997, p. 192) translates differently:

DEFINITION 4 (Prakken's criterion for hierarchical rebuttal). *Let $\langle I, f, <\rangle$ be a prioritized imperative structure, where $<$ is a strict partial order on I (like the asymmetric part of a partial order $\leq$). Then for all $\Gamma, \Delta \subseteq I$:*

$$\Gamma \ll_P \Delta \quad \text{iff} \quad \exists j \in \Delta : \forall i \in \Gamma : i < j.$$

So Δ can be improved by exchanging some member with any member of Γ. Giving the rationale here and in Prakken and Sartor (1997, p. 36)[10], the change in the order of the quantifiers seems intentional – yet it makes a difference: let I be $\{!p_1, !p_2, !(p_3 \wedge \neg p_1), !(p_4 \wedge \neg p_2)\}$ and $!p_1 < !(p_3 \wedge \neg p_1)$ and

$!p_2 < !(p_4 \wedge \neg p_2)$. E.g. p_1, p_2 may be primary targets and $p_3 \wedge \neg p_1$, $p_4 \wedge \neg p_2$ respective secondary ones, where reaching the secondary target includes failing to reach the (better) primary one. Reaching both primary targets seems best, and in fact $\{!p_1, !p_2\} \ll_{AM} \{!(p_3 \wedge \neg p_1), !(p_4 \wedge \neg p_2)\}$: for every member in the left set there is a lower-ranking one in the right. Also $I \Downarrow \bot = \{\{!p_1, p_2\}\}$, since all four full prioritizations yield this preferred remainder. But $\{!p_1, !p_2\} \not\ll_P \{!(p_3 \wedge \neg p_1), !(p_4 \wedge \neg p_2)\}$ as no member in the right set ranks lower than *all* in the left. So Prakken's criterion appears even less suited to our task than Alchourrón & Makinson's.[11]

Sartor (1991) used Alchourrón and Makinson's criterion for a "prevailing" relation between subsets *modulo* a rejected sentence A,[12] as follows:

DEFINITION 5 (Sartor's "prevailing" relation). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure, where $<$ is a strict partial order on I, and $\ll_{AM}$ be as defined above. Then for all $\Gamma, \Delta \subseteq I$:*

$$\Gamma \ll_S^A \Delta \quad \textit{iff} \ [\Gamma \cup \Delta]^f \vdash_{PL} A, \textit{ and } \forall \Delta' \subseteq \Delta \textit{ such that } [\Gamma \cup \Delta']^f \vdash_{PL} A : \\ \exists \Gamma' \subseteq \Gamma : [\Gamma' \cup \Delta']^f \vdash_{PL} A \textit{ and } \Gamma' \ll_{AM} \Delta'.$$

Finally **pref** $(I \curlywedge A) = min_{\ll_S^A}(I \curlywedge A)$.

To see how his definition works, consider first the "mediocrity rules" example: $\langle I, f, < \rangle$ is $!p_1 < !(\neg p_1 \wedge p_2 \wedge \neg p_3) < !p_3$. Then $\{!p_1, !p_3\} \ll_S^{\bot} \{!(\neg p_1 \wedge p_2 \wedge \neg p_3)\}$, as the only subset of the right set conflicting with the left set is the right set itself, and for this set some conflicting subset of the left set, namely $\{!p_1\}$, is strictly less exposed than the right set. So now the result is as it intuitively should be. Sartor's relation also handles the example against Prakken's criterion well: here $I = \{!p_1, !p_2, !(p_3 \wedge \neg p_1), !(p_4 \wedge \neg p_2)\}$, with $!p_1 < !(p_3 \wedge \neg p_1)$ and $!p_2 < !(p_4 \wedge \neg p_2)$. Then $\{!p_1, !p_2\} \ll \frac{\bot}{S} \{!(p_3 \wedge \neg p_1), !(p_4 \wedge \neg p_2)\}$, as the three subsets of the right set conflicting with the left set are $\{!(p_3 \wedge \neg p_1)\}$, $\{!(p_4 \wedge \neg p_2)\}$ and the right set itself, with which the following respective subsets of the left set are both in conflict and strictly less exposed: $\{!p_1\}$, $\{!p_2\}$, and the left set itself. So reaching the primary targets is best, as it should be. In fact, it can be proved that preferred remainders are always prevailing remainders, but the converse does not hold:

EXAMPLE (Counterexample to **pref**$(I \curlywedge A) \subseteq I \Downarrow A$). *Let $\langle I, f, < \rangle$ be such that I consists of*

$$\begin{array}{ll} i_1 : !(p_1 \wedge ((p_2 \vee p_3) \rightarrow q)) & i_3 : !p_2 \\ i_2 : !(\neg p_1 \wedge ((p_2 \wedge p_3) \rightarrow \neg q)) & i_4 : !(p_3 \wedge (p_1 \rightarrow \neg p_2)) \end{array}$$

and the ordering $i_1 < i_2 < [i_3, i_4]$. Intuitively, to satisfy i_1 takes priority, and then the only choice is the one between the equally ranking i_3 and i_4, and indeed, $I \Downarrow \bot = \{\{i_1, i_3\}, \{i_1, i_4\}\}$. But the remainder $\{i_2, i_3, i_4\}$, missing the most important imperative i_1, is also in $\mathbf{pref}(I \curlywedge \bot)$, as

$$(a) \quad \{i_1, i_3\} \not\ll^{\bot}_{S} \{i_2, i_3, i_4\}, \qquad (b) \quad \{i_1, i_4\} \not\ll^{\bot}_{S} \{i_2, i_3, i_4\}.$$

For (a), consider $\{i_4\}$, which is a subset of the right hand set: it conflicts with the left hand set, so for the relation to hold, a strictly less exposed subset of the left set must also conflict with $\{i_4\}$. The only such subset is the left set itself, but since for its member i_3 there is no strictly more exposed member in $\{i_4\}$, it is not strictly less exposed. The refutation of (b) works similarly using $\{i_3\}$.

So Sartor's definition also produces counterintuitive results where Brewka and Nebel's approach does not.[13]

4.2. UTILITY-REFLECTING PRIORITIES

Regarding the neighboring realm of epistemic logic, and the related problem of revising belief sets in the face of conflicting information, such information often finds the reasoner less willing to give up some beliefs than others. In an attempt to allocate this ordering of "epistemic importance" a rôle in determining which of the contradictory beliefs should be given up, Gärdenfors (1984) proposed the following: Let $\mathbf{K}$ be a belief set (set of descriptive sentences) that is the logical closure of some finite basis, and $\leq$ a relation (of epistemic importance) that is a complete preorder on this set, which additionally ranks logically equivalent beliefs equally. For any remainder $\Gamma \in \mathbf{K} \perp A$ there is then a "spanning sentence" S_Γ in Γ that derives any element in Γ. Then for any $\Gamma, \Delta \in \mathbf{K} \perp A$:

$$\Gamma \ll_G \Delta \quad \text{iff } S_\Gamma < S_\Delta$$

So a remainder is preferred to some other iff its spanning sentence is epistemically at least as important as that of the other. It is essential for the construction that $<$ is a complete ordering on $\mathbf{K}$, which due to logical closure includes the spanning sentence that is the "sum" of a remainder. But, the logical philosophers not being kings, a set of imperative-contents is rarely logically closed, which precludes a direct parallel. Yet, choosing subsets that "in sum" are the most important has an analogue if the ordering of the imperatives reflects not so much their importance or rank of the source, but a measure of "goodness" or utility of the outcome when satisfying the imperative. For this, let the (well-founded, strict partial) order $<_u$ correspond to a function $u : X \to \mathbb{R}$, with $I^f \subseteq X \subseteq \mathscr{L}_{PL}$, that assigns a real number to (at least) what the imperatives demand in a manner conversely respecting $<_u$, i.e. if $i <_u j$ then $u(i^f) > u(j^f)$. Then let

$$I\,{}^{\underline{u}}A = \{\Gamma \subseteq I | \Gamma^f \nvdash_{PL} A \text{ and } \forall \Delta \subseteq I: \text{ if} \sum u(\Delta^f) \geq \sum u(\Gamma^f) \text{ then } \Delta^f \vdash_{PL} A\}.$$

So an A-consistent subset is preferred to another if the good brought about by satisfying all of its demands sums up to a higher value than by doing so for the other set. $I\,{}^{\underline{u}}A$ includes the maximally preferred among such sets. Obviously $I\,{}^{\underline{u}}A \subseteq I \curlywedge A$ if u assigns just positive values. The O-operator is then defined by

$$\text{(td-7)} \quad \langle I, f, <_u \rangle \models O(A/C) \quad \text{iff} \quad \forall \Gamma \in I\,{}^{\underline{u}}\neg C : \Gamma^f \cup \{C\} \vdash_{PL} A.$$

If a set of imperatives that requires A to be true in the situation C constitutes a "reason" to call A obligatory in this situation, then $O(A/C)$ is true if the reasons for obligatoriness A have more collective weight (sum up to a higher value) than any such reasons for $\neg A$.[14] Well-foundedness of $<$, with the condition that u respects $<$, excludes infinitely increasing utilities and thus corresponds to the limit assumption in preference semantics, which avoids counterintuitive results but may be criticized as superficial (cf. Fehige 1994). Still, infinitely lower and lower ranking imperatives, whose satisfaction nevertheless produces some good, are not excluded, and thus not the scenario of McNamara (1995) where a bad act like killing your mum will eventually become permitted by (only then) piling up good deeds of small value. To protect imperatives from getting overruled by inferiors in this manner, one could assign to an imperative's satisfaction a utility that is absolutely higher than the sum of the utilities assigned to the satisfaction of any number of lower-ranking imperatives (think of the sequence 1, 0.5, 0.25, 0.125, ...). It is immediate that if $<_u$ is thus protected, i.e. for all $i \in I$, $u(i^f) \geq \sum u(\{j \in I \mid i <_u j\}^f)$, and u assigns just positive values, then $I\,{}^{\underline{u}}A \subseteq I \Downarrow A$.

4.3. SAFE CONTRACTION AND A MODIFICATION

What characterizes Brewka and Nebel's approaches is that they try to maximize the number of higher-ranking imperatives in a set that avoids a conflict. This intuition has a counterpart formulated by Alchourrón (1986):[15]

> "It is logical to believe that the reasonable way of overcoming a conflict of obligations is to leave aside the less important norms contributing to its creation."

Alchourrón's proposal is then to remove from the set of norms all those that are least-ranking in a minimal conflicting subset, thus removing the conflict as well. For the formal description of this "safe contraction", let $I \curlyvee A$ be the minimal sets of imperatives such that the corresponding sentences derive A (the "A-kernels" of I), i.e. the set of all $\Gamma \subseteq I$ such that (i) $\Gamma^f \vdash_{PL} A$ and

(ii) for all $\Delta \subseteq I$, if $\Delta \subset \Gamma$ then $\Delta^f \nvdash_{PL} A$. $<$ being a strict partial ordering on I, the following defines the set of all the least important imperatives of A-kernels of I:

$$\sigma(I \curlyvee A) =_{def} \{max_< \Gamma \mid \Gamma \in I \curlyvee A\}.$$

Due to PL-compactness, any $\Gamma \in I \curlyvee A$ is finite and so $max_< \Gamma \neq \emptyset$ if $\Gamma \neq \emptyset$. Then the set I/A of elements of I that are "safe" with respect to A is defined by

$$I/A =_{def} I \backslash \sigma(I \curlyvee A).$$

If C characterizes the situation, the dyadic deontic operator can then be defined as in (td-3), but using $I/\neg C$ instead of I:[16]

(td-8) $\langle I, f, < \rangle \models O(A/C)$ iff $[I/\neg C]^f \cup \{C\} \vdash_{PL} A$.

To see how this works, consider again the example of the road accident:

EXAMPLE (The road accident: safe contraction). $I = \{!A, !B\}$, *with* $!A < !B$, *as helping the victims of the accident is more important than proceeding to the appointment.* $\vdash_{PL} T \rightarrow (A \rightarrow \neg B)$, *because staying excludes meeting my friend.* $I \curlyvee \bot = \{\{!A, !B\}\}$, $max_< \{!A, !B\} = \{!B\} = \sigma(I \curlyvee \bot)$, *hence* $I/\bot = \{!A\}$. *Hence I must stay and help, as it should be.*

EXAMPLE (The road accident V: safe contraction). $I = \{!A, !B, !C\}$. *Mirjam left too late for both the post office and the day care center, so* $\vdash_{PL} T \rightarrow (C \rightarrow \neg B)$. *Waiting makes her too late for the day care center, so* $\vdash_{PL} T \rightarrow (A \rightarrow \neg C)$. *The legal requirement to stay and her duty to fetch her grandma are both more important than posting the letters, and Mirjam decides that the law also overrides her familiar duty, so* $!A < !C < !B$. *Intuitively Mirjam must wait and then hurry to the post office. But* $I \curlyvee \bot = \{!A, !C\}, \{!B, !C\}\}$, $max_< \{!A.!C\} = \{!C\}$, $max_< \{!B, !C\} = \{!B\}$. *So* $\sigma(I \curlyvee \bot) = \{!B, !C\}$ *and* $I/\bot = \{!A\}$. *Hence a solution by safe contraction only requires Mirjam to wait for the police, though she could still post the letters when the police are finished.*

The last example illustrates that safe contraction removes elements too liberally; even when it has already removed some element of a kernel, further elements get removed as well even though under the definition of a kernel removing one suffices: $!C$ was removed due to its conflict with $!A$, so there was no need to also remove $!B$ to avoid the conflict with $!C$. This makes life too easy for the norm subjects, and some moderation appears necessary. Following the idea that the removal mechanism should somehow be adjusted to the set's shrinking it brings about, a moderated version of safe contraction can be defined as follows:

DEFINITION 6 (Moderated safe contraction). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure, where $<$ is a well-founded strict partial order on I. For each $A \in \mathscr{L}_{PL}$, let $M_{[<,\alpha]} \subseteq I, 0 \leq \alpha \leq card(I)$, be*

$$M^A_{[<,\alpha]} =_{def} \bigcup_{\beta<\alpha} M^A_{[<,\beta]} \cup min_< \sigma([I \backslash \bigcup_{\beta<\alpha} M^A_{[<,\beta]}] \curlyvee A).$$

Finally $M^A_< =_{def} \bigcup^{card(I)}_{\alpha=1} M^A_{[<,\alpha]}$ *and* $I /\!/_< A = I \setminus M^A_<$.

So if X are the $<$-minimal elements in $\sigma(I \curlyvee A)$, then the moderated mechanism first puts X in the set $M_<$ of elements to be removed, then the $<$-minimal elements in $\sigma([I \backslash X] \curlyvee A)$, etc. Thus elements get removed in each step until there is no A-kernel left in I minus the last version of $M_<$, which also means that the cardinality of I suffices for the indices (I omit superscripts if the meaning is clear). To see how this works, consider again the above example:

EXAMPLE (The road accident V: moderated safe contraction). *$!C$ is the minimal element in $\sigma(I \curlyvee \bot) = \{!B, !C\}$ Removing $!C$ is unavoidable: the only other element $!A$ in the $\bot$-kernel in which $!C$ is maximal ranks before $!C$, so if $!A$ was maximal in some $\bot$-kernel, $!C$ would not be in $min_< \sigma(I \curlyvee \bot)$. So $!C$ is in $M_{[<,1]} = \{!C\}$, equalling $M_<$ since no $\bot$-kernel is left in $I \backslash \{!C\}$. Hence $I /\!/_< \bot = \{!A, !B\}$, so both obligations – to wait and post the letters – remain.*

If $<$ is not a well-order, the moderated method still removes too much: if it removes anything from a kernel, it removes all $<$-maximal members, but by definition of a kernel, one is enough. Instead, one might again consider the full prioritizations that preserve $<$, rather than $<$. The relation between moderated safe contraction and Brewka's method is then the following:

THEOREM 6 (Moderated safe contractions and preferred remainders). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure, where $<$ is a well-founded strict partial order on I. Then $I \Downarrow A = \{I /\!/_{\prec} A \mid \prec$ is some full prioritization of $<\}$.*

5. Uniquely prioritized imperatives

As demonstrated, the method proposed by Brewka and Nebel seems adequate for a resolution of normative conflicts by use of priority orderings of the underlying norms. Whether all conflicts and dilemmas are thus avoidable is a matter of dispute. W. D. Ross may be understood as claiming that conflicts are merely apparent and that by weighing all relevant facts and reasons, it can be decided which of the conflicting *prima facie* duties are really our duties (cf. Ross 1930, Searle 1980, p. 242). G. H. von Wright stated that an axiological order can "provide a safeguard against any genuine predicament" (cf. von Wright 1968, p. 68, 80). And Hare's description of "critical moral thinking," that lets principles override other, less important principles, suggests that this process can overcome all moral conflicts (Hare 1981, p. 43, 50). On the other hand, Barcan Marcus (1980) and Horty (2003, p. 564) have argued that if it is the presence of certain facts that determines the ordering, i.e. this is not an arbitrary hacking through the Gordian knot, then situations might be incomparable (if all such facts are missing), or be completely symmetrical (e.g. identical obligations towards identical twins), so conflicts remain possible.

Rather than take sides in this controversy, I will examine what is required if the method of Brewka and Nebel is to resolve all conflicts and dilemmas. For all possible[17] situations C, $I \Downarrow \neg C$ must then be a singleton: otherwise there are $\Gamma, \Delta \in I \Downarrow \neg C$ such that $\Gamma^f \cup \Delta^f \vdash_{PL} \neg C$, and there is a dilemma. So I define:

DEFINITION 7 (Uniquely prioritized imperatives). *Let* $\langle I, f, < \rangle$ *be a prioritized imperative structure, Then* $\langle I, f, < \rangle$ *is called uniquely prioritized iff for all* $C \in \mathscr{L}_{PL}$ *such that,* $\nvdash_{PL} \neg C$, $card(I \Downarrow \neg C) = 1$.

The question of a resolution of all conflicts can now be rephrased to ask what an imperative structure $\langle I, f, < \rangle$ must be like to be uniquely prioritized. The most obvious way to avoid all conflicts for any situation C is to let the strict partial order $<$ be total, i.e. for any two $i, j \in I$, either $i < j$ or $j < i$. Then there is just one full prioritization $\prec$ that preserves $<$, namely $<$ itself. This result was also noted by Nebel (1992, Proposition 11), who constructs $<$ as the strict part of a complete preorder $\leq$ – then requiring $<$ to be total makes each equivalence class $[i]$ a singleton. Gärdenfors (1984, p. 146), who also constructs $<$ as the strict part of some complete preorder $\leq$, points out that it is enough if the only choice left is between equivalents, i.e. here either $i < j$ or $j < i$ for any $i, j \in I$ with $\nvdash_{PL} i^f \leftrightarrow j^f$. But this still requires too much. It suffices that the demands of elements in each $[i]$ are chained, so for all $j_1, j_2 \in [i] : \vdash_{PL} j_1^f \rightarrow j_2^f$ or $\vdash_{PL} j_2^f \rightarrow j_1^f$ – then it does not matter in which order j_1 and j_2 appear in a full prioritization $\prec$ of $<$.[18] And one can be even more lax, as demonstrated by the following cases, which also are of a sort in which no ambiguity ever arises:

Case 1: Let $\langle I, f, < \rangle$ be $[!p_1, !p_2] < !(p_1 \wedge p_2)$, so the demands of the two higher-ranking imperatives are "doubled" by a lower one. If both, $!p_1$, $!p_2$ are in $S_\prec \in I \Downarrow \neg C$, then adding $!(p_1 \wedge p_2)$ adds nothing. Otherwise $[S_{[\prec\downarrow!(p_1 \wedge p_2)]} \cup \{!(p_1 \wedge p_2)\}]^f \vdash_{PL} \neg C$, so $!\neg(p_1 \wedge p_2)$ cannot be in $S_\prec$.

Case 2: Let $\langle I, f, <\rangle$ be $[!p_1, !p_2] < !\neg(p_1 \wedge p_2)$, so the demands of the two higher-ranking imperatives run contrary to the lower one's. Either both $!p_1$,$!p_2$ are in $S_{\prec} \in I \Downarrow \neg C$, so $!\neg(p_1 \wedge p_2)$ cannot be consistently added, or $[S_{[\prec\downarrow !\neg(p_1 \wedge p_2)]}]^f \cup \{C\} \vdash_{PL} \neg(p_1 \wedge p_2)$, so adding $!\neg(p_1 \wedge p_2)$ adds nothing.

Case 3: Let $\langle I, f, <\rangle$ be $!(p_1 \wedge p_2) < !(p_1 \wedge \neg p_2) < [!p_1, !p_2]$. If $p_1 \wedge p_2$ is consistent with C, then $I \Downarrow \neg C$ is $\{\{!(p_1 \wedge p_2), !p_1, !p_2\}\}$. Otherwise, $\{C\} \vdash_{PL} \neg(p_1 \wedge p_2)$. If $p_1 \wedge \neg p_2$ is consistent with C, then $I \Downarrow \neg C$ is $\{\{!(p_1 \wedge \neg p_2), !p_1\}\}$. Otherwise, also $\{C\} \vdash_{PL}, \neg(p_1 \wedge \neg p_2)$. Hence $\{C\} \vdash_{PL} \neg p_1$. Then any set in $I \Downarrow \neg C$ contains at most $!p_2$, depending on whether C is consistent with p_2.

Unable to make out a necessary and sufficient requirement for $card(I \Downarrow \neg C) = 1$, without reference to particular C, I can only rephrase its definition as follows:

THEOREM 7 (Property of uniquely prioritized imperatives). *Any $\langle I, f, <\rangle$, where $<$ is the strict part of some complete preorder $\leq$, is uniquely prioritized iff for all consistent $C \in \mathscr{L}_{PL}, \Gamma \in I \Downarrow \neg C, i \in I$ and $j_1, j_2 \in [i]$,*

$$\{i' \in \Gamma | i' < i\}^f \cup \{C\} \vdash_{PL} j_1^f \rightarrow j_2^f \text{ or } \{i' \in \Gamma | i' < i\}^f \cup \{C\} \vdash_{PL} j_2^f \rightarrow j_1^f.$$

Let a semantics be called a *uniquely prioritized imperative semantics* iff it defines the truth of $\mathscr{L}_{DDL}$-sentences using (td-6), but only considers uniquely prioritized imperative structures. It validates the additional axiom scheme:

(RMon) $P(D/C) \rightarrow (O(A/C) \rightarrow O(A/C \wedge D))$.

Consider the system *PD* that was sound and (weakly) complete with respect to prioritized imperative semantics. The system that results when (RMon) is added is Hansson's (1969) *DSDL*3 as axiomatized by Spohn (1975):

THEOREM 8 (*PD*+RMon equals *DSDL3*). *Let PD*+(RMon) *be like PD, except that* (RMon) *is added as art axiom scheme. Then PD*+(RMon) = *DSDL*3, *which is the smallest set that contains all $\mathscr{L}_{DDL}$-instances into tautologies as well as all $\mathscr{L}_{PL}$-instances of the following schemes*:

(A0) $O(A/A)$

(A1) *If* $\nvdash_{PL} A$ *then* $\vdash_{DSDL3} \neg O(\bot /A)$

(A2) $O(B \wedge C/A) \leftrightarrow (O(B/A) \wedge O(C/A))$

(A3) *If* $\vdash_{PL} A \leftrightarrow A'$ *and* $\vdash_{PL} B \leftrightarrow B'$ *then* $\vdash_{DSDL3} O(B/A) \leftrightarrow O(B'/A')$

(A4) $P(B/A) \rightarrow (O(C/A \wedge B) \leftrightarrow O(B \rightarrow C/A))$

and is closed under modus ponens.

Hansson's *DSDL*3, which is also the core of Åqivist's (1986) system **G**, usually characterizes a preference-based dyadic deontic semantics, i.e. an interpretation of deontic formulas using a "betterness relation" between valuations or possible worlds, and not using explicitly given norms as it is here. However, this extreme interpretational change did not result in a changed logical behavior of the deontic operators, i.e. *DSDL*3 can be reconstructed quite naturally using an imperative semantics with an axiological order in the spirit of von Wright (1968):

THEOREM 9 (Soundness, completeness of *DSDL*3). *DSDL3 is sound and (only) weakly complete for uniquely prioritized imperative semantics.*

The construction used to prove the completeness theorem also exhibits the following relation between priorities and contrary-to-duty norms: suppose there is a finite, or finitely based, set of deontic truths $\Delta \subseteq \mathscr{L}_{DDL}$, and $\langle I, f, < \rangle$ is a uniquely prioritized imperative structure that makes true all of Δ. The construction used to prove *DSDL*3-completeness provides a uniquely prioritized imperative structure $\langle I', f', <' \rangle$ that also makes true all of Δ, but the demands of these imperatives are now chained and so the priority relation can remain empty or let all imperatives rank equally (cf. Theorem 7). E.g. if $\langle I, f, < \rangle$ is $!p_1 < !p_2$, then $\langle I', f', <' \rangle$ is $[!(p_1 \wedge p_2), !(\neg(p_1 \wedge p_2) \to p_1), \; !((\neg(p_1 \wedge p_2) \wedge \neg p_1) \to p_2)]$. These can be viewed as contrary-to-duty norms, where the primary obligation is: to make $p_1 \wedge p_2$ true, if that is not possible, to make p_1 true, and if that is also not possible, to make p_2 true. So instead of using ranks and priorities to avoid conflicts, contrary-to-duty formalizations can be employed to produce the same effect. While there may be pragmatic reasons to attach higher priority to the commands of the king than the wishes of his jester, from the standpoint of logic, exception clauses like "if the king did not say otherwise" suffice.

6. Conclusion

Describing how a resolution of normative conflicts using priorities works is surprisingly difficult. A method based on a proposal by Horty could not solve complex cases where more than two norms conflict or overriding norms are no longer satisfiable or are themselves overridden. A method developed for theory revision by Brewka and Nebel, which creates maximally non-conflicting sets by starting with a maximal set of what is most important and incrementally adding maximally to it, is able to resolve these difficulties. Alternative mechanisms discussed in normative theory, namely Alchourrón and Makinson's exposure criterion and variants by Prakken and Sartor, have counterintuitive results in cases that Brewka and Nebel's method adequately solves. The same holds for Alchourrón's "safe contractions", but the intuition underlying his construction, that in a conflict the least important norms should be set aside, is captured in a moderated version that is equivalent to Brewka's. I explain how a semantics that models explicitly given imperatives can be used to define deontic operators. When the "preferred remainders" from Brewka and Nebel's method are thus used for the definition of a disjunctive (skeptical) dyadic deontic operator, then such a semantics is characterized by the axiom system *PD*, which resembles Kraus, Lehmann and Magidor's system *P* with the dyadic D-axiom added. Whether all conflicts can be resolved using priorities is left to philosophical dispute, but conditions are discussed that guarantee priority orderings, which do just that. For a semantics that defines its dyadic deontic operators with respect to such "uniquely prioritized" imperatives, Hansson's axiom system *DSDL*3 is proved to be sound and complete. The proof's construction also exhibits the fact that priorities are dispensable and that contrary-to-duty constructions can take their place.

Most of the approaches discussed here include conditional entities, which pose different problems like the following (rephrased from Rintanen 1994):

(1) α says: if you drink anything, then don't drive.
(2) β says: if you go to the party, then you do the driving.
(3) γ says: if you go the party, then have a drink with me.
(4) You go the party.

Suppose that β does not mind if you have one drink with γ, and γ does not care that you may be driving, and let the three imperatives be ranked in descending order. One may be tempted to reason as follows: consider first the imperative in line (1), but it has not yet been 'triggered' as you have not yet drunk anything, so it is set aside. Regarding (2), its condition is true, so you must do the driving. Still, only (3) is triggered, so you should have a drink with γ. But satisfying (2) and (3) both triggers and violates the highest-ranking imperative. Is it not more prudent to violate one of the lower-ranking imperatives instead of the higher-ranking one? For a solution, we need an adequate definition of triggering (that can handle e.g. disjunctive inputs, like Makinson and van der Torre's "basic output" (Makinson and van der Torre 2000, 2001), and to find a maximizing strategy that is consistent with the above intuition. It is clear that the present discussion has not provided the tools to properly address such problems, so these must be left to further study.

Acknowledgments

I thank David Makinson for valuable help with an earlier version, and Lou Goble, John F. Horty and Henry Prakken for helpful discussions and clarifying correspondence. An earlier version of the paper was presented to the working group Law and Logic of the XXII World Congress of Philosophy of Law and Social Philosophy (IVR 2005) 24–29: May 2005, Granada.

Notes

[1] Though some discussed approaches cover conditional imperatives, or entities that can be interpreted as such, these cause problems that are best considered separately.

[2] *PL* is based on a language $\mathscr{L}_{PL}$, defined from a set of proposition letters *Prop* $=\{p_1,$ $p_2,...\}$, Boolean connectives " $\neg$", " $\wedge$", " $\vee$", " $\rightarrow$ ", "$\leftrightarrow$" and brackets "(","")" as usual, The truth of a $\mathscr{L}_{PL}$-sentence A is defined recursively using a valuation function v : $Prop \rightarrow \{1,0\}$ (I write $v \vDash A$), starting with $v \vDash p$ iff $v(p)=1$ and continuing as usual. If $A \in \mathscr{L}_{PL}$ is true for all valuations it is called a tautology. *PL* is the set of all tautologies, and this set is used to define provability, consistency and derivability (I write $\Gamma \vdash_{PL} A$) as usual. $\top$ is an arbitrary tautology, and $\bot$ is $\neg\top$.

[3] E.g. (td-1) most closely resembles definitions of Kanger (1957) and Alchourrón and Bulygin (1981). For authors belonging to this tradition cf. Hansen (2001), Section 1 and Hansen (2004), fn. 1, in addition to which Ziemba (1971) must be mentioned.

[4] Cf. Horty (1997). The "disjunctive" ought is more commonly referred to as "skeptical" non-monotonic inference. Horty (2003) attributes the proposal to Brink (1994), yet the idea to use such a definition for (dyadic) deontic logic already appears in Lewis (1981). For alternatives in the deontic-logical treatment of normative conflicts cf. Goble (2005).

[5] Cf. Kraus et al. (1990), where, however, the proofs are done in a more general setting. Also cf. my (Hansen 2005) for constructive proofs in the manner of Spohn (1975) as well as more comparisons and truth definitions and axioms for an alternative "credulous" O-operator.

[6] Horty's definition only employs circumstances to derive consequents from a set of conditional imperatives, but this has no effect on the solution of the examples.

[7] Horty is preparing a refined version of *Binding* that solves all of the examples (private correspondence).

[8] Brewka (1991) and for Nebel cf. Rott (1993, fn. 9). For the rationale, let the ordered I be $\langle i_0, \ldots, i_{0.125}, i_{0.25}, i_{0.5}, i_1 \rangle$, with $i_1 = !p, i_{0.5} = !\neg p, i_{0.25} = !p, i_{0.125} = !\neg p$, etc., and $i_0 = !q$. We cannot tell whether p or $\neg p$ is obligatory, but this is not a case of conflict either, all imperatives $!p$ being overridden by ones demanding $\neg p$, and *vice versa*.

[9] I use notation from both, Brewka (1989, 1991), Brewka and Eiter (1999), and Nebel (1991, 1992).

[10] Also cf. Sartor (2005, p. 734): "preference must be given to the argument such that its weakest defeasible subreasons are better than the weakest defeasible subreasons in the other".

[11] Prakken could argue that he only compares minimal conflict pairs (subarguments), which $\{!p_1, !p_2\}, \{!(p_3 \wedge \neg p_1), !(p_4 \wedge \neg p_2)\}$ is not, while $\{!p_1\} \ll_P \{!(p_3 \wedge \neg p_1)\}$ and $\{!p_2\} \ll_P \{!(p_4 \wedge \neg p_2)\}$ hold. But let $I = \{!p_1, !p_2, !(p_3 \wedge (p_4 \rightarrow \neg(p_1 \wedge p_2))), !(p_4 \wedge (p_3 \rightarrow \neg (p_1 \wedge p_2)))\}$, with $!p_1 < !(p_3 \wedge (p_4 \rightarrow \neg(p_1 \wedge p_2)))$ and $!p_2 < !(p_4 \wedge (p_3 \rightarrow \neg(p_1 \wedge p_2)))$, so with the primary targets one "bonus" secondary target is reachable. Still $\{!p_1, !p_2\} \not\ll_P \{!(p_3 \wedge (p_4 \rightarrow \neg(p_1 \wedge p_2))), !(p_4 \wedge (p_3 \rightarrow \neg(p_1 \wedge p_2)))\}$, and this is a minimal conflict pair. Yet intuitively, the argument for $p_1 \wedge p_2$ should win over any for $\neg(p_1 \wedge p_2)$.

[12] Sartor (1991) simply rejects contradictions, yet the adjustment to any A is immediate.

[13] The indicative version of the example also shows that Prakken cannot avoid counterintuitive results by replacing, in his definition of a rebuttal in (1997), his own relation $\ll_P$ by the relation $\ll_{AM}$ of Alchourón and Makinson for the comparison of minimal conflict pairs: then all arguments for q are rebutted by arguments for $\neg q$ as demonstrated, but intuitively any consistent argument including i_1 cannot be defeated and so the argument for q should win.

[14] For resolving legal arguments by summing up weights of reasons cf. Hage (1991, 1996).

[15] Also cf. Iwin (1972, p. 486): "When we are in a situation compelled to satisfy two obligations requiring contradictory actions, then the most natural way out of this difficulty consists in comparing the two obligations and not satisfying the less important one."

[16] $I/\neg C$ is the notation in Alchourrón and Makinson (1985), whereas the notation in Alchourrón (1986) would be I/C. Alchourrón's own truth definition for deontic operators employs a deontic logic as basic logic and conditional imperatives that are not treated here.

[17] If C is a contradiction, then by definition $I \Downarrow \neg C = \emptyset$.

[18] I owe this insight to Leon van der Torre (private correspondence).

Appendix: Proofs

THEOREM 1 (Existence of full prioritizations). *For every well-founded strict partial order $<$ on a set Γ there is a full prioritization $\prec$, i.e. a strict well-order that is order-preserving with respect to $<$.*

Proof. Let $<$ be a well-founded strict partial order on the set Γ. Let each $x \in \Gamma$ be assigned an ordinal α_x in the following way: $\alpha_x = 0$ for the elements in $min_< \Gamma$, and for any other $x \in \Gamma$, $\alpha_x = sup\{\alpha_y | y < x\} + 1$, where *sup* denotes the supremum of a set of ordinals. Transfinite induction available for well-founded partial ordered sets tells us that α_x is well-defined for any $x \in \Gamma$. Let the equivalence class $[x] = \{y \in \Gamma | \alpha_y = \alpha_x\}$. Finally $\prec$ is an arbitrary strict well-order on elements of the same equivalence class, and $x \prec y$ if α_x is smaller than α_y. Clearly $\prec$ is a strict well-order on Γ. To prove $x \prec y$ if $x < y$, for any $x, y \in \Gamma$, suppose $x < y$. Then α_x is in $\{\alpha_z | z < y\}$, so $\alpha_y = sup \{\alpha_z | z < y\} + 1$ is at least $\alpha_x + 1$. Hence $x \prec y$.

THEOREM 2 (Preferred remainders are remainders). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure, where $<$ is a well-founded strict partial order on I. Then $I \Downarrow A \subseteq I \curlywedge A$.*

Proof. Suppose $S_\prec \in I \Downarrow A$. Since $[S_\prec]^f \nvdash A$ there is some $\Gamma \in I \curlywedge A$ such that $S_\prec \subseteq \Gamma$. Suppose $\Gamma \nsubseteq S_\prec$. Let i be some element such that $i \in \Gamma$, but $i \notin S_\prec$. Then by the construction of $S_\prec$, $[\bigcup_{j \prec i} S_{[\prec \downarrow j]}]^f \cup \{i^f\} \vdash_{PL} A$, and since $\bigcup_{j \prec i} S_{[\prec \downarrow j]} \subseteq S_\prec$, also $[S_\prec]^f \cup \{i^f\} \vdash_{PL} A$. But then $\Gamma^f \vdash_{PL} A$, which is excluded by $\Gamma \in I \curlywedge A$. So $\Gamma = S_\prec$ and $S_\prec \in I \curlywedge A$.

THEOREM 3 (Nebel's prioritized removals). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure, where $<$ is the asymmetric part of a complete, well-founded preorder $\leq$ on I. Then $I \Downarrow A$ equals*

$$\{\Gamma \subseteq I \mid \Gamma^f \nvdash_{PL} A \text{ and } \forall \Delta \subseteq I : \text{ if } \Delta \ll_N \Gamma \text{ then } \Delta^f \vdash_{PL} A\}.$$

Proof. Left-to-right: Suppose $\Gamma \in I \Downarrow A$. Suppose for some $\Delta \subseteq I : \Delta \ll_N \Gamma$. Then there is a priority class $[i]$ such that for all $j < i, \Gamma \cap [j] = \Delta \cap [j]$ and $\Gamma \cap [i] \subset \Delta \cap [i]$. Let $\prec$ be the full prioritization used to construct $\Gamma = S_\prec$ and i be the $\prec$-first element of $[i]$ with $i \notin \Gamma$, but $i \in \Delta$. By construction of $S_\prec, [\bigcup_{j \prec i} S_{[\prec \downarrow j]}]^f \cup \{i^f\} \vdash_{PL} A$, and also $\bigcup_{j \prec i} S_{[\prec \downarrow j]} = \{j \in S_\prec \mid j \prec i\}$. Since $\prec$ is order-preserving, for all $j \in S_\prec$, if $j \prec i$, then $j < i$ or $j \in [i]$. Then $j \in \Delta$, by definition of $\ll_N$ or by choice of i. Hence $\Delta^f \vdash_{PL} A$.

Right-to-left: Suppose $\Gamma \subseteq I$, $\Gamma^f \nvdash_{PL} A$ and $\forall \Delta \subseteq I$. If $\Delta \ll_N \Gamma$, then $\Delta^f \vdash_{PL} A$. Let $\prec$ be a well ordering on I respecting $<$ which puts $i \prec j$ for any $i \in \Gamma$ and $j \in [i] \backslash \Gamma$, i.e. an ordering that positions the elements of Γ before their $\leq$-equals, and $\Delta = S_\prec$. We prove $\forall i \in I : \Delta \cap [i] = \Gamma \cap [i]$ by induction on $<$:

For the *induction basis*, let i be some $\leq$-least element. By definition of $\prec$ the elements of $\Gamma \cap [i]$ are positioned $\prec$-before any other elements of Δ, and by assumption $\Gamma^f \cap [i] \nvdash_{PL} A$, so $\Gamma \cap [i] \subseteq \Delta$ due to the construction of $S_\prec = \Delta$. Suppose there is some $j \in \Delta \cap [i]$ such that $j \notin \Gamma \cap [i]$. Then $\Gamma \cap [i] \subset \Delta \cap [i]$, and since trivially $\forall j < i : \Gamma \cap [j] = \Delta \cap [j]$ as i was $\leq$-least and so no such j exists, $\Delta \ll_N \Gamma$ holds. Then it must be that $\Delta^f \vdash_{PL} A$, but by definition of $S_\prec = \Delta$ this is excluded. So $\Delta \cap [i] \subseteq \Gamma \cap [i]$ and, hence, $\Delta \cap [i] = \Gamma \cap [i]$.

For the *induction step*, let i be some arbitrary element of I. By definition of $\prec$, the elements of $\Gamma \cap [i]$ are positioned $\prec$-before any other elements of $\Delta \cap [i]$. The induction hypothesis guarantees that for all $j \in I$ with $j < i : \Delta \cap [j] = \Gamma \cap [j]$, so $\{j \in \Delta \mid j \prec i\} = \{j \in \Gamma \mid j \prec i\}$. As $\{j \in \Gamma \mid j \prec i\}^f \cup [\Gamma \cap [i]]^f \nvdash_{PL} A$ (otherwise $\Gamma \vdash_{PL} A$), for any $j \in \Gamma \cap [i] : [\bigcup_{k \prec j} S_{[\prec \downarrow k]}]^f \cup \{j^f\} \nvdash_{PL} A$, so $j \in \Delta$ and hence $\Gamma \cap [i] \subseteq \Delta$. Suppose there is some $j \in \Delta \cap [i]$ with $j \notin \Gamma \cap [i]$. Then $\Gamma \cap [i] \subset \Delta \cap [i]$, and since also $\forall j < i : \Gamma \cap [i] = \Delta \cap [i]$ by the induction hypothesis, $\Delta \ll_N \Gamma$ holds. Then it must then be that $\Delta^f \vdash_{PL} A$, but by definition of $S_\prec = \Delta$ this is excluded. So $\Delta \cap [i] \subseteq \Gamma \cap [i]$ and hence $\Delta \cap [i] = \Gamma \cap [i]$.

Since $\forall i \in I : \Delta \cap [i] = \Gamma \cap [i]$ we have $\Gamma = \Delta = S_\prec$ and so $\Gamma \in I \Downarrow A$.

THEOREM 4 (Preferred remainders, after Ryan and Sakama & Inoue). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure, $<$ a well-founded strict partial order on I and $I \Downarrow A$ and $\mathbf{p}(I \curlywedge A)$ be as defined. Then $I \Downarrow A = \mathbf{p}(\mathrm{I} \curlywedge \mathrm{A})$.*

Proof. Left-to-right: Suppose $S_\prec \in I \Downarrow A$. Assume for r.a.a. that there is a $\Gamma \neq S_\prec$ in $I \curlywedge A$ such that for all $i \in S_\prec \backslash \Gamma$ there is a $j \in \Gamma \backslash S_\prec$ with $j \prec i$. Let i be the $\prec$-least element in $S_\prec \backslash \Gamma$, which is not empty, because otherwise $S_\prec \subseteq \Gamma$, but this is excluded by maximality of $S_\prec$, so existence of i is guaranteed by well-foundedness of $\prec$. Then $\bigcup_{k \prec i} S_{[<\downarrow k]} \subseteq \Gamma$: otherwise there is some $k \in S_\prec$ with $k \prec i$ but $k \notin \Gamma$ and so i would not be the $\prec$-least. Since $j \notin S_\prec$ it must be that $[S_{[<\downarrow j]}]^f \cup \{j^f\} \vdash_{PL} A$. But $j \prec i$, so $S_{[<\downarrow j]} \subseteq \bigcup_{k \prec i} S_{[<\downarrow k]}$, and chaining results we get $\Gamma^f \vdash_{PL} A$, which contradicts $\Gamma \in I \curlywedge A$.

Right-to-left: Suppose $\Gamma \in \mathbf{p}(I \curlywedge A)$. Let $\prec$ be a full prioritization for which the statement in the definition of $\mathbf{p}(I \curlywedge A)$ is true. By definition, $S_\prec$ is in $I \Downarrow A$ and so also in $I \curlywedge A$. Assume for r.a.a. that $S_\prec \neq \Gamma$. Then there is some $i \in \Gamma \backslash S_\prec$ such that for all $j \in S_\prec \backslash \Gamma, i \prec j$. Suppose $\bigcup_{j \prec i} S_{[\prec\downarrow j]} \nsubseteq \{j \in \Gamma \,|\, j \prec i\}$. Then there is a $j \in S_\prec \setminus \Gamma : j \prec i$ and so $i \nprec j$ by the antisymmetry of $\prec$, and $i \nless j$ since $\prec$ preserves $<$, but we assumed otherwise. So $\bigcup_{j \prec i} S_{[\prec\downarrow j]} \subseteq \{j \in \Gamma \,|\, j \prec i\}$. Since $i \in \Gamma \backslash S_\prec$ it must be that $i \notin S_\prec$, so by definition $[\bigcup_{j \prec i} S_{[\prec\downarrow j]}]^f \cup \{i^f\} \vdash_{PL} A$. So $\Gamma^f \vdash_{PL} A$, which contradicts $\Gamma \in I \curlywedge A$.

THEOREM 5 (Soundness, completeness of *PD*). *PD is sound and (only) weakly complete with respect to prioritized imperative semantics.*

Proof. (Sketch) We must prove that *PD* is (a) sound with respect to prioritized imperative semantics, (b) (weakly) complete with respect to prioritized imperative semantics and (c) only weakly complete, i.e. not compact.

(a) Soundness. The validity of (CExt), (ExtC), (DM) and (DC) is immediate. (DN) is unrestrictedly valid since $\forall \Gamma \in I \Downarrow \neg C : \Gamma^f \vdash_{PL} \top$ is trivial. Theorem 1 guarantees the existence of a generating $\prec$, so if C is not a contradiction (then $I \Downarrow \neg C = \emptyset$ by definition) at least $\emptyset$ is in $I \Downarrow \neg C$, so $I \Downarrow \neg C \neq \emptyset$ and (DD-R) must hold. Leaving details to the reader, for (Cond), suppose $O(A/C \wedge D)$, so $\forall \Gamma \in I \Downarrow \neg(C \wedge D) : \Gamma^f \cup \{C \wedge D\} \vdash_{PL} A$ and assume for r.a.a. that $O(D \rightarrow \quad A/C)$ is false, which yields $\exists \Delta \in I \Downarrow \neg C : \Delta^f \cup \{C\} \nvdash_{PL} D \rightarrow A$ and also includes $\Delta^f \nvdash_{PL} \neg(C \wedge D)$. Likewise, for (CCMon) suppose $O(A \wedge D/C)$, which yields $\forall \Delta \in I \Downarrow \neg C : \Delta^f \cup \{C\} \vdash_{PL} A \wedge D$ and also includes $\Delta^f \nvdash_{PL} \neg(C \wedge D)$. Δ is obtained from some full prioritization $\prec$ via Definition 2, i.e. $\Delta = S_\prec^{\neg C}$. Appealing to the $\prec$-first element on which they might differ, prove that $S_\prec^{\neg C} = S_\prec^{\neg(C \wedge D)}$.

(b) Weak completeness. We must prove that for any $A \in \mathscr{L}_{DDL}$ such that $\neg A \notin PD$ there is a prioritized imperative structure $\langle \mathscr{I}, f, < \rangle$ that models A. I proved *PD* to be weakly complete for basic imperative semantics and (td-4) in Hansen (2005), so take the basic imperative structure $\langle \mathscr{I}, f \rangle$ that models A and define $< = \emptyset$. Then $I \Downarrow \neg C = I \curlywedge \neg C$ for any C. Hence $\langle \mathscr{I}, f, < \rangle$ also models A using (td-6).

(c) Non-compactness. Leaving details to the reader, I gave a counterexample to compactness of basic imperative semantics in Hansen (2005, Theorem 3), i.e. a non-satisfiable set $\Gamma \subseteq \mathscr{L}_{DDL}$ of which all finite subsets are satisfiable, which applies here as well. The tricky part is to show that the sets $I_{p_1} \cup I_{(p_1 \leftrightarrow \neg p_2)}$ and $I_{\neg p_1} \cup I_{(p_1 \leftrightarrow p_2)}$ are in $I \Downarrow \bot$, which works using the complete preorder induced by ordinal labels from the proof of Theorem 1 (for each union choose a full prioritization $\prec$ that puts its elements $\prec$-first in any equivalence class and prove that $S_\prec$ in $I \Downarrow \bot$ constructed from $\prec$ equals the respective union).

THEOREM 6 (Moderated safe contractions and preferred remainders). *Let $\langle I, f, < \rangle$ be a prioritized imperative structure, where $<$ is a well-founded strict partial order on I. Then $I \Downarrow A = \{I /\!/_\prec A \mid \prec$ is some full prioritization of $<\}$.*

Proof. I first give a construction *next*(i), containing the – compared to i – "next-larger" element in $M_\prec$, and a helpful lemma:

LEMMA 1. *Let $\prec$ be a full prioritization of $<$, $A \in \mathscr{L}_{PL}$ and $M_\prec$ be accordingly defined (cf. Definition 8). For any $i \in I$,* let

$$\text{next}(i) =_{df} min_\prec\{j \in M_\prec \mid i = j \, or \, i \prec j\}$$

contain i if $i \in M_\prec$ or else the element in $M_\prec$ that comes $\prec$-next to i. Let

$$M_i = \begin{cases} min_\subseteq\{M_{[\prec,\alpha]} \mid \text{next}(i) \subseteq M_{[\prec,\alpha]}\} & if \ \text{next}(i) \neq \emptyset, \\ M_\prec & otherwise \end{cases}$$

be the first set in the construction of $M_\prec$ that includes next(i) *if that is non-empty and otherwise the whole of $M_\prec$. Then*

$$M_i \setminus \text{next}(i) = M_\prec \cap \{j \in I | j \prec i\}.$$

Proof. For *next*(i), note that due to well-orderliness, the $\prec$-minimum of a non-empty set is a singleton, and so is *next*(i), unless there is no j in $M_\prec$ with $i = j$ or $i \prec j$, in which case it is $\emptyset$. For M_i, note that since $M_{[\prec,\alpha]}$ increases with each step α, $0 \leq \alpha < card(I)$, the inclusion-minimum is well defined. To prove the lemma, if neither $i \in M_\prec$ nor a $\prec$-next member in $M_\prec$ exists and so $M_i = M_\prec$, then the equivalence is trivial. Otherwise $M_i = M_{[\prec,\alpha]}$ for some α, and we must prove that (i) if $j \in I$ is in $M_{[\prec,\alpha]} \setminus next(i)$ then $j \prec i$, and (ii) that if $j \in M_\prec$ is not in $M_{[\prec,\alpha]} \setminus next(i)$ then $j \not\prec i$.

ad i): Let $j \in M_{[\prec,\alpha]}$. We assumed *next*(i) to be non-empty, so let *next*(i) $= \{i'\}$. By definition, i' is the element of $M_{[\prec,\alpha]}$ that was added at step α and so if $i' = j$, then j is not in $M_{[\prec,\alpha]}A \setminus next(i)$. Otherwise, by the construction of $M_{[\prec,\alpha]}$, j is in some $M_{[\prec,\beta]}$ with $\beta < \alpha$. Let β be the smallest of these. To be in the increase of

$M_{[\prec,\alpha]}, i'$ must be $\prec$-maximal in some A-kernel X in $I \setminus \bigcup_{\gamma<\alpha} M_{[\prec,\gamma]}$. Since the process is incremental, we have $I \setminus \bigcup_{\gamma<\alpha} M_{[\prec,\gamma]} \subseteq I \setminus \bigcup_{\gamma<\beta} M_{[\prec,\gamma]}$. So $X \subseteq I \setminus \bigcup_{\gamma<\beta} M_{[\prec,\gamma]}$. Hence, if $i' \prec j$, then j is not $\prec$-minimal among the $\prec$-maximal members of A-kernels in $I \setminus \bigcup_{\gamma<\beta} M_{[\prec,\gamma]}$, but this contradicts that j is the element by which $M_{[\prec,\beta]}$ was increased at step β. So $i' \not\prec j$, we supposed $i' \neq j$, and so by $\prec$-connectedness $j \prec i'$. Also, if $i=j$ or $i \prec j$, then $i'=j$ or $i' \prec j$ by the definition of *next*(i) and $j \in M_{\prec}$. By the previous result and $\prec$-transitivity, we obtain $j \prec j$, contradicting $\prec$-irreflexivity. So $i \not\prec j, i \neq j$ and hence by connectedness, $j \prec i$.

ad ii): Assume $j \in M_{\prec}$ is not in $M_{[\prec,\alpha]} \setminus next(i)$. Again let $next(i) = \{i'\}$. So i' is the element of $M_{[\prec,\alpha]}$, which is added at step α. If $j = i'$, then $i=j$ or $i \prec j$ and in both cases $j \not\prec i$. Otherwise j must be in some $M_{[\prec,\beta]}$ with $\alpha < \beta$. Let β be the smallest of these. To be in the increase of $M_{[\prec,\beta]}$, j must be $\prec$-maximal in some A-kernel X in $I \setminus \bigcup_{\gamma<\beta} M_{[\prec,\gamma]}$. Since the process is incremental, we have $I \setminus \bigcup_{\gamma<\beta} M_{[\prec,\gamma]} \subseteq I \setminus \bigcup_{\gamma<\alpha} M_{[\prec,\gamma]}$. So $X \subseteq I \setminus \bigcup_{\gamma<\alpha} M_{[\prec,\gamma]}$. Suppose $j \prec i'$, then i' is not $\prec$-minimal among the $\prec$-maximal members of A-kernels in $I \setminus \bigcup_{\gamma<\alpha} M_{[\prec,\gamma]}$. But this contradicts that i' is the element by which $M_{[\prec,\alpha]}$ was increased at step α. So $j \not\prec i'$, and we supposed $j \neq i'$. Thus by $\prec$-connectedness, $i' \prec j$ and hence $i \prec j$ by $\prec$-transitivity and definition of *next*(i).

For the theorem, it now suffices to prove that for each full prioritization $\prec, A \in \mathscr{L}_{PL}$ and accordingly constructed sets $S_{\prec}$ (Definition 2) and $M_{\prec}$ (Definition 8), $I \setminus M_{\prec} = S_{\prec}$, which is done by induction over $\prec$:

Induction basis: Let i_0 be the $\prec$-least element in I. Suppose i_0 is in $I \setminus M_{\prec} : I \setminus M_{\prec}$ contains no A-kernels, so $\{i_0\} \nvdash_{PL} A$ and so i_0 is in $S_{\prec}$. Suppose i_0 is not in $I \setminus M_{\prec}$. Then i_0 is $\prec$-maximal in some A-kernel of I. So for all other j in this A-kernel, $j \prec i$, but also $i \prec j$ because i_0 is $\prec$-least. So by $\prec$-antisymmetry this A-kernel equals $\{i_0\}$, so $\{i_0^f\} \vdash_{PL} A$ and by definition $i_0 \notin S_{\prec}$. Hence $i_0 \in I \setminus M_{\prec}$ iff $i_0 \in S_{\prec}$.

Induction step: Right-to-left: Suppose $i \notin I \setminus M_{\prec}$, so $i \in M_{\prec}$. Let α be the step that has i in its increase, i.e. $M_{[\prec,\alpha]} = \bigcup_{\beta<\alpha} M_{[\prec,\beta]} \cup \{i\}$. Since $i \in M_{\prec}$, $next(i) = i$, so $M_{[\prec,\alpha]}$ is M_i as defined above. So $M_i \setminus \{i\} = \bigcup_{\beta<\alpha} M_{[\prec,\beta]} = M_{\prec} \cap \{j \in I | j < i\}$ by Lemma 1. Hence $[I \setminus \bigcup_{\beta<\alpha} M_{[\prec,\beta]}] \cap \{j \in I | j \prec i\} = [I \setminus [M_{\prec} \cap \{j \in I | j \prec i\}]] \cap \{j \in I | j \prec i\} = [I \setminus M_{\prec}] \cap \{j \in I | j \prec i\}$ which by the induction hypothesis equals $S_{\prec} \cap \{j \in I | j \prec i\}$, and by definition this equals $\bigcup_{j \prec i} S_{[\prec\downarrow j]}$. To be in the increase of $M_{[\prec,\alpha]}, i$ must be $\prec$-maximal in some A-kernel X in $I \setminus \bigcup_{\beta<\alpha} M_{[\prec,\beta]}$, so $X \subseteq [I \setminus \bigcup_{\beta<\alpha} M_{[\prec,\beta]}] \cap \{j \in I | j \prec i\} \cup \{i\}$. So by definition of an A-kernel and the above equation, $[\bigcup_{j \prec i} S_{[\prec\downarrow j]}]^f \cup \{i^f\} \vdash_{PL} A$. Hence $i \notin S_{[\prec\downarrow i]}$ and $i \notin S_{\prec}$.

Left-to-right: Suppose $i \notin S_{\prec}$. So by definition $[\bigcup_{j \prec i} S_{[\prec\downarrow j]}]^f \cup \{i^f\} \vdash_{PL} A$ and $[\bigcup_{j \prec i} S_{[\prec\downarrow j]}]^f \nvdash_{PL} A$. By the induction hypothesis $\bigcup_{j \prec i} S_{[\prec\downarrow j]} = S_{\prec} \cap \{j \in I | j \prec i\} = [I \setminus M_{\prec}] \cap \{j \in I | j \prec i\}$. By the above lemma, $M_{\prec} \cap \{j \in I | j \prec i\} = M_i \setminus next(i)$ and so $[I \setminus M_{\prec}] \cap \{j \in I | j \prec i\} = [I \setminus [M_{\prec} \cap \{j \in I | j \prec i\}]] \cap \{j \in I | j \prec i\} = [I \setminus [M_i \setminus next(i)]] \cap \{j \in I | j \prec i\}$, which, since $next(i) \subseteq I$, equals $[[I \setminus M_i] \cup next(i)] \cap \{j \in I | j \prec i\}$, which equals $[I \setminus M_i] \cap \{j \in I | j \prec i\}$ since $next(i) \cap \{j \in I | j \prec i\} = \varnothing$ by definition of *next*(*i*). So we obtain the results that $[[I \setminus M_i] \cap \{j \in I | j \prec i\}]^f \cup \{i^f\} \vdash_{PL} A$ and $[[I \setminus M_i] \cap \{j \in I | j \prec i\}]^f \nvdash_{PL} A$. First observe $next(i) \neq \varnothing$ – otherwise $M_i = M_\prec$ by definition, so $[I \setminus M_\prec]^f \cup \{i^f\} \vdash_{PL} A$, but by definition $[I \setminus M_\prec]^f \nvdash_{PL} A$, so $i \notin [I \setminus M_\prec]$ and, hence, i is in $M_\prec$, but then $next(i) = \{i\} \neq \varnothing$. So $M_i = M_{[\prec,\alpha]}$ for some α. Next assume $i \notin M_\prec$. Then $i \in I \setminus M_i$. By the first result obtained above, i is $\prec$-maximal in some A-kernel in $I \setminus M_i$. Suppose i is not $\prec$-minimal among the $\prec$-maximal members of A-kernels in $I \setminus M_i$: then some A-kernel must lie completely in $[I \setminus M_i] \cap \{j \in I | j \prec i\}$, but the other result excluded this. So $min_\prec \sigma([I \setminus \bigcup_{\beta < \alpha} M_{[\prec,\beta]}] \curlyvee A) = \{i\}$, so i is in the increase at step α, i.e. $i \in M_i$, so $i \in M_\prec$, which contradicts its assumed negation and so is true. So $i \notin I \setminus M_\prec$.

THEOREM 7 (Property of uniquely prioritized imperatives). *Any* $\langle I, f, < \rangle$, *where* $<$ *is the strict part of some complete preorder* $\leq$, *is uniquely prioritized iff for all consistent* $C \in \mathscr{L}_{PL}, \Gamma \in I \Downarrow \neg C, i \in I$ *and* $j_1, j_2 \in [i]$,

$$\{i' \in \Gamma \,|\, i' < i\}^f \cup \{C\} \vdash_{PL} j_1^f \to j_2^f \textit{ or } \{i' \in \Gamma | \, i' < i\}^f \cup \{C\} \vdash_{PL} j_2^f \to j_1^f.$$

Proof. For the *left-to-right* direction (in contraposition), suppose there are $j_1, j_2 \in [i]$, a consistent $D \in \mathscr{L}_{PL}$ and $\Gamma \in I \Downarrow \neg D$ with neither $\{i' \in \Gamma \,|\, i' < i\}^f \cup \{D\} \vdash_{PL} j_1^f \to j_2^f$ nor $\{i' \in \Gamma \,|\, i' < i\}^f \cup \{D\} \vdash_{PL} j_2^f \to j_1^f$. Consider $C = D \wedge \neg(j_1^f \wedge j_2^f)$. If $\{i' \in \Gamma \,|\, i' < i\}^f \vdash_{PL} \neg C$, then $\{i' \in \Gamma \,|\, i' < i\}^f \vdash_{PL} \neg D \vee (j_1^f \wedge j_2^f)$, which includes $\{i' \in \Gamma \,|\, i' < i\}^f \cup \{D\} \vdash_{PL} j_1^f \to j_2^f$, contrary to what we supposed. Let $\prec_1$ be a full prioritization of $<$ that for all $i' < i$ is like Γ, i.e. it contains $\{i' \in \Gamma \,|\, i' < i\}$, and then puts j_1 $\prec_1$-first among the members of $[i]$. Likewise let $\prec_2$ be a full prioritization of $<$ that contains $\{i' \in \Gamma \,|\, i' < i\}$ and puts j_2 $\prec_2$-first among the members of $[i]$. $\{i' \in \Gamma \,|\, i' < i\}^f \vdash_{PL} \neg C$ is excluded, so regarding the constructions $S_{\prec_1}, S_{\prec_2} \in I \Downarrow \neg C, \bigcup_{k \prec_1 j_1} S_{[\prec_1 \downarrow k]} = \{i' \in \Gamma \,|\, i' < i\} = \bigcup_{k \prec_2 j_2} S_{[\prec_2 \downarrow k]}$. If $[\bigcup_{k \prec_1 j_1} S_{[\prec_1 \downarrow k]}]^f \cup \{j_1^f\} \vdash_{PL} \neg C$, then $\{i' \in \Gamma \,|\, i' < i\}^f \cup \{j_1^f\} \vdash_{PL} \neg D \vee (j_1^f \wedge j_2^f)$ or equivalently $\{i' \in \Gamma \,|\, i' < i\}^f \cup \{D\} \vdash_{PL} j_1^f \to j_2^f$, but that was excluded. So $[\bigcup_{k \prec_1 j_1} S_{[\prec_1 \downarrow k]}]^f \cup \{j_1^f\} \nvdash_{PL} \neg C$ and so $j_1 \in S_{\prec_1}$, and likewise $j_2 \in S_{\prec_2}$ is proved. But then $S_{\prec_1} \neq S_{\prec_2}$ – otherwise $[S_{\prec_1}]^f \vdash_{PL} j_1^f \wedge j_2^f$ and so $[S_{\prec_1}]^f \vdash_{PL} \neg C$, which is excluded. So *card* $(I \Downarrow \neg C) \neq 1$ for some non-contradictory $C \in \mathscr{L}_{PL}$.

For the *right-to-left* direction, assume for r.a.a. that $card(I \Downarrow \neg C) \neq 1$ for some consistent $C \in \mathscr{L}_{PL}$, so there are two full prioritizations $\prec_1$ and $\prec_2$

such that $S_{\prec_1} \neq S_{\prec_2}$ for the according $S_{\prec_1}, S_{\prec_2} \in I \Downarrow \neg C$. Then there is a $i \in I$ such that $S_{\prec_1} \cap [i] \neq S_{\prec_2} \cap [i]$: let $[i]$ be the $<$-first such equivalence class. So there is a $j_1 \in [i]$ with either $j_1 \in S_{\prec_1}$ and $j_1 \notin S_{\prec_2}$ or *vice versa*. This being equal, assume the former, so $[\bigcup_{k \prec_2 j_1} S_{[\prec_2 \downarrow k]}]^f \cup \{j_1^f\} \vdash_{PL} \neg C$. So there is a $(j_1^f \rightarrow \neg C)$-kernel X in $\bigcup_{k \prec_2 j_1} S_{[\prec_2 \downarrow k]}$, and $X \cap [i] \neq \emptyset$, else $X \subseteq S_{\prec_1}$ by choice of $[i]$ and since $j_1 \in S_{\prec_1}$ we obtain $[S_{\prec_1}]^f \vdash_{PL} \neg C$, which is excluded. We get $\{j \in S_{\prec_2} | j < i\}^f \cup \{C\} \vdash_{PL} j_2^f \rightarrow j_1^f$ or $\{j \in S_{\prec_2} | j < i\}^f \cup \{C\} \vdash_{PL} j_1^f \rightarrow j_2^f$ from the right side of the iff-clause for any $j_2 \in X \cap [i]$. Suppose the first case holds for a $j_2 \in X \cap [i]$. Since $j_2 \in X \subseteq \bigcup_{k \prec_2 j_1} S_{[\prec_2 \downarrow k]}$ we obtain $[\bigcup_{k \prec_2 j_1} S_{[\prec_2 \downarrow k]}]^f \vdash_{PL} C \rightarrow j_1^f$, and since the $(j_1^f \rightarrow \neg C)$-kernel X is in $\bigcup_{k \prec_2 j_1} S_{[\prec_2 \downarrow k]}$, this yields $[\bigcup_{k \prec_2 j_1} S_{[\prec_2 \downarrow k]}]^f \vdash_{PL} \neg C$, which is excluded. So for all $j_2 \in X \cap [i]$: $\{j \in S_{\prec_2} | j < i\}^f \cup \{C\} \vdash_{PL} j_1^f \rightarrow j_2^f$. For all other $k \in X$, i.e. $k \notin [i]$, we have $k \prec_2 i$ by $X \subseteq \bigcup_{k \prec_2 j_1} S_{[\prec_2 \downarrow k]}$ and so $k \in S_{\prec_1}$ since $j_1 \in [i]$ and due to the choice of i. So since j_1 is in $S_{\prec}$, and also $\{j \in S_{\prec_2} | j < i\} \subset S_{\prec}$ by the choice of i, for any $k \in X : [S_{\prec}]^f \cup \{C\} \vdash_{PL} k^f$. Hence, $[S_{\prec}]^f \vdash_{PL} \neg C$, which the construction of the antecedent excludes. This completes the r.a.a.

THEOREM 8 (*PD* + RMon equals *DSDL*3). *Let PD*+(RMon) *be like PD, except that* (RMon) *is added as an axiom scheme. Then PD*+(RMon) = *DSDL*3.

Proof. Easy and left as exercise to the reader.

THEOREM 9 (Soundness, completeness of *DSDL*3). *DSDL*3 *is sound and (only) weakly complete for uniquely prioritized imperative semantics.*

Proof. We must prove that *DSDL*3 is (a) sound with respect to prioritized imperative semantics, (b) (weakly) complete with respect to prioritized imperative semantics, and (c) only weakly complete, i.e. not compact.

(a) Soundness. Only the validity of (RMon) needs to be proved, as Theorem 5 proved validity of the *PD*-axioms with respect to all, not necessarily uniquely, prioritized imperative structures. The proof is like the one for (CCMon) in Theorem 5, i.e. we assume $O(A/C)$ and obtain that for all $\Delta \in I \Downarrow \neg C : \Delta^f \cup \{C\} \vdash_{PL} A$ and obtain the additionally included fact that for all such $\Delta, \Delta^f \cup \{C\} \nvdash_{PL} \neg D$ by assuming $P(D/C)$ and from *card* $(I \Downarrow \neg C) = 1$ in case C is consistent (otherwise DN and CExt make the conclusion $O(A/C \wedge D)$ trivially true).

(b) Weak completeness. The proof is a condensed version of its more widely applicable variants in Spohn (1975) and my (2001, 2005), to which I refer for further features of the construction. We must prove that for any *DSDL*3-consistent. $A \in \mathscr{L}_{DDL}$, i.e. $\neg A \notin DSDL3$, there is a uniquely

prioritized imperative structure $\langle I, f, < \rangle$ that models A. We build a disjunctive normal form of A and obtain a disjunction of conjunctions, where each conjunct is $O(B/D)$ or $\neg O(B/D)$. One disjunct must then be $DSDL3$-consistent. Let δ be that disjunct. Let the δ-restricted language $\mathscr{L}^{\delta}_{PL}$ be the PL-sentences that contain only proposition letters occurring in δ. Let $r(\mathscr{L}^{\delta}_{PL})$ be 2^{2^n} mutually non-equivalent representatives of $\mathscr{L}^{\delta}_{PL}$, n being the number of proposition letters in δ. $\mathscr{L}_{PL}$-sentences now mean their representatives in $r(\mathscr{L}^{\delta}_{PL})$. Construct a set $\varDelta \subset \mathscr{L}_{DDL}$ such that:

(a) Any conjunct of δ is in $\varDelta$.
(b) For all $B, D \in r(\mathscr{L}^{\delta}_{PL})$: either $P(B/D)$ or $O(\neg B/D) \in \varDelta$.
(c) $\varDelta$ is $DSDL3$-consistent.

It then suffices to find a uniquely prioritized imperative structure that makes true all of $\varDelta$. We identify what Hansson (1969) called the *deontic basis* in an extension $\|C\|$ (Spohn 1975 writes $\tilde{C}$) by letting $\mathcal{O}_C = \bigwedge\{A \in r(\mathscr{L}^{\delta}_{PL}) \mid O(A/C) \in \varDelta\}$ be the "sum" of everything demanded in the situation C. From this definition and (b), (c) it follows immediately that $O(\mathcal{O}_C/C) \in \varDelta$. Furthermore, observe

(O1) $O(A/C) \in \varDelta$ iff $\vdash_{PL} \mathcal{O}_C \rightarrow A$.

System of spheres: Let $S = \langle C_1, \ldots, C_n \rangle$ be defined recursively by letting
(i) $C_1 = \top$,
(ii) if $C_i \in S$ and $C_i \neq \bot$ then $C_{i+1} = C_i \wedge \neg \mathcal{O}_{C_i}$ (otherwise $i = n$).

Observe that for all C_i, C_j, $1 \leq i < j \leq n$:

(O2) $\vdash_{PL} C_j \rightarrow C_i$
(O3) $C_j \wedge \mathcal{O}_{C_i} = \bot$

Proof. Immediate (due to A0 and O1, $\vdash_{PL} O_{C_j} \rightarrow C_j$ so also $\mathcal{O}_{C_j} \wedge \mathcal{O}_{C_i} = \bot$).

(O4) $C_i \in r(\mathscr{L}^{\delta}_{PL}), 1 \leq i \leq n$
(O5) The sequence is finite, i.e. $n \neq \infty$
(O6) $\mathcal{O}_{C_1} \vee \ldots \vee \mathcal{O}_{C_n} = \top$

Proofs. (O4) is immediate from the definitions. Regarding (O5), $r(\mathscr{L}^{\delta}_{PL})$ is finite, so if S is infinite, then $C_i = C_j$ for some $1 \leq i < j \leq n$. Then $C_i \wedge \mathcal{O}_{C_i} = \bot$ due to (O3), but since $\vdash_{PL} O_{C_i} \rightarrow C_i$ due to (A0) and (O1), then $\mathcal{O}_{C_i} = \bot$ and so due to (A1) $C_i = \bot$. But then the sequence ends with i. For (O6), if $\mathcal{O}_{C_1} \vee \ldots \vee \mathcal{O}_{C_n} \neq \top$ then $\top \wedge \neg \mathcal{O}_{\top} \wedge \ldots \wedge \neg \mathcal{O}_{C_{n-1}} \wedge \neg \mathcal{O}_{C_n} \neq \bot$. The left side equals $C_n \wedge \mathcal{O}_{C_n}$, so C_n is not last in the sequence, contrary to what was assumed.

Smallest A-permitting sphere: Let C_A be the first C_i with $\nvdash_{PL} (A \wedge \mathcal{O}_{C_i}) \rightarrow \bot$, i.e. for all $j < i : (A \wedge \mathcal{O}_{C_j}) = \bot$. Then furthermore:

(O7) For any $r(\mathscr{L}^{\delta}_{PL})$-sentence $A \neq \bot$ there is some $C_A \in S$.

(O8) $\vdash_{PL} A \to C_A$

(O9) If$A \neq \bot$, then $P(A/C_A) \in \varDelta$.

(O10) $\vdash_{PL} \mathcal{O}_A \leftrightarrow (\mathcal{O}_{C_A} \wedge A)$

Proofs. (O7) is immediate from (O6). Regarding (O8) let $C_A = C_i$, so for all $j < i : \mathcal{O}_{C_j} \wedge A = \bot$, so $\vdash_{PL} A \to (\top \wedge \neg\mathcal{O}_\top \wedge \ldots \wedge \neg\mathcal{O}_{C_{i-1}})$ and thus $\vdash_{PL} A \to C_i$. For (O9), if C_A exists, then by the construction of $\varDelta$ either $P(A/C_A) \in \varDelta$ or $O(\neg A/C_A) \in \varDelta$, but in the second case $\mathcal{O}_{C_A} \wedge A = \bot$ due to (O1), which is excluded by the definition of C_A. Regarding (O10): due to (O8) and derivability of (Cond) we obtain $O(A \to \mathcal{O}_A/C_A) \in \varDelta$ from $O(\mathcal{O}_A/A) \in \varDelta$, so $\vdash_{PL} (\mathcal{O}_{C_A} \wedge A) \to \mathcal{O}_A$, which is the right-to-left direction. If $A = \bot$, then $\mathcal{O}_A = \bot$ since $\vdash_{PL} \mathcal{O}_A \to A$ due to (A0) and (O1) and then the left-to-right direction is trivial. Otherwise $P(A/C_A) \in \varDelta$ due to (O9) and $O(\mathcal{O}_{C_A}/A) \in \varDelta$ is obtained from $O(\mathcal{O}_{C_A}/C_A)$, (O8) and derivability of (RMon). Thus $\vdash_{PL} \mathcal{O}_A \to \mathcal{O}_{C_A}$ and due to (A0) and (O1) also $\vdash_{PL} \mathcal{O}_A \to A$. Both include the left-to-right version.

Canonical construction of $\langle I, f, < \rangle$: Let $I = \{C \to O_C \mid C \in S\}$, f be identity and $<$ be the strict part of $\leq = I \times I$ (or $\varnothing$, *al gusto*). *Verification:* due to the constructions of S and I it is immediate that for all $i, j \in I : \vdash_{PL} i^f \to j^f$ or $\vdash_{PL} j^f \to i^f$, so the demands of the imperatives are chained and I is uniquely prioritized (Theorem 7). All imperatives are equally ranked, so $I \Downarrow A = I \curlywedge A$.

Coincidence: We must prove that for all $B, D \in r(\mathscr{L}^{\delta}_{PL})$, $\langle I, f, < \rangle$ models $O(B/D)$ iff $O(B/D) \in \varDelta$. *Right-to-left*: Assume $O(B/D) \in \varDelta$. If $D = \bot$, then by definition $I \Downarrow \neg D = I \curlywedge \neg D = \varnothing$, so (td-6) makes $O(B/D)$ trivially true. Otherwise, by (O7) there is some $C_D \in S$. By its definition, $D \wedge \mathcal{O}_{C_D}$ is consistent, so $D \wedge (C_D \to \mathcal{O}_{C_D})$ is also consistent and so $C_D \to \mathcal{O}_{C_D}$ must be in (the) $\varGamma \in I \curlywedge \neg D$. Since $\vdash_{PL} D \to C_D$ by (O8) and $\vdash_{PL} (\mathcal{O}_{C_D} \wedge D) \to \mathcal{O}_D$ by (O10) we have $\vdash_{PL} ((C_D \to \mathcal{O}_{C_D}) \wedge D) \to \mathcal{O}_D$, so $\varGamma \cup \{D\} \vdash_{PL} \mathcal{O}_D$ and due to (O1), $\varGamma \cup \{D\} \vdash_{PL} B$ and so (td-6) makes $O(B/D)$ true. *Left-to-right:* Assume that for $\varGamma \in I \Downarrow \neg D = I \curlywedge \neg D$ we have $\varGamma \cup \{D\} \vdash_{PL} B$. If $D = \bot$, then $O(B/D) \in \varDelta$ follows trivially from (A0), (A2) and (A3). Otherwise, by (O7) there is a $C_D \in S$: $D \wedge \mathcal{O}_{C_D}$ is by definition consistent, so $D \wedge (C_D \to \mathcal{O}_{C_D})$ is consistent and $C_D \to \mathcal{O}_{C_D}$ must be in (the) set $\varGamma$. $C_D = C_i$ for some i in the construction of S. For all $j < i, \vdash_{PL} D \to C_j$ due to (O8) and (O2), and $\vdash_{PL} D \to \neg\mathcal{O}_{C_j}$ by definition of C_D, so $C_j \to \mathcal{O}_{C_j} \notin \varGamma$. For all $j > i, \vdash_{PL} \mathcal{O}_{C_D} \to \neg C_j$ by (O3), so with (O8) we obtain $\{C_D \to \mathcal{O}_{C_D}\} \cup \{D\} \vdash_{PL} C_j \to \mathcal{O}_{C_j}$ for any $C_j \to \mathcal{O}_{C_j} \in \varGamma$. So if $\varGamma \cup \{D\} \vdash_{PL} B$, then $\{C_D \to \mathcal{O}_{C_D}\} \cup \{D\} \vdash_{PL} B$. Hence $\vdash_{PL} (\mathcal{O}_{C_D} \wedge D) \to B$ by (O8), $\vdash_{PL} \mathcal{O}_D \to B$ by (O10), so $O(B/D) \in \varDelta$ by (O1).

(c) Non-Compactness. Compactness is disproved by any infinite set Δ of $\mathscr{L}_{DDL}$-sentences, of which any finite subset Δ_f is satisfiable, but not Δ. Consider

$$\begin{aligned}\Delta = \{O(p_1/\top)\} & \cup \{P(\neg p_1/\neg(p_1 \wedge \ldots \wedge p_n)) \,|\, n > 1\} \\ & \cup \{P(A \wedge \neg p_n/\neg(p_1 \wedge \ldots \wedge p_n)) \,|\, n > 1 \text{ and } \{A\} \nvdash_{PL} p_n\}\end{aligned}$$

Let $\Delta_f \subset \Delta$ be a finite subset of Δ. Let n be the highest index such that p_n occurs in some formula in Δ_f. Then $\langle I, f, < \rangle$ with $I = \{!(p_1 \wedge \ldots \wedge p_n)\}$, $<$ being $\emptyset$, satisfies $\Delta_f : I \Downarrow \bot = \{I\} = \{\{!(p_1 \wedge \ldots \wedge p_n\}\}$. So for all sets Γ in $I \Downarrow \bot : \Gamma^f \vdash_{PL} p_1$, so $O(p_1/\top)$ is true. The only $\Gamma \in I \Downarrow (p_1 \wedge \ldots \wedge p_n)$, $n > 1$, is the empty set and that plus $\neg(p_1 \wedge \ldots \wedge p_n)$ is consistent with $\neg p_1$ and $A \wedge \neg p_n$ if $\{A\} \nvdash_{PL} p_n$. So $P(\neg p_1/\neg(p_1 \wedge \ldots \wedge p_n))$ and $P(A \wedge \neg p_n/\neg (p_1 \wedge \ldots \wedge p_n))$ are true for any $n > 1$. Hence Δ is finitely satisfiable.

Suppose Δ is satisfiable, i.e. there is a uniquely prioritized imperative structure $\langle I, f, < \rangle$ that makes all Δ true. The sets $I \Downarrow \bot$ and $I \Downarrow (p_1 \wedge \ldots \wedge p_n)$ must be singletons (Definition 9). To make $O(p_1/\top)$ true, there must be some full prioritization $\prec$ such that $[S^{\bot}_{\prec}]^f \vdash_{PL} p_1$, where $S^{\bot}_{\prec} \in I \Downarrow \bot$ is as described in Definition 2. Due to PL-compactness there must be some $i \in I$ such that $[S^{\bot}_{[\prec\downarrow i]}]^f \vdash_{PL} p_1$. Let i be the $\prec$-first such element. Its existence is guaranteed by well-foundedness of $\prec$ since $\nvdash_{PL} p_1$. So $[\bigcup_{j \prec i} S^{\bot}_{[\prec\downarrow j]}]^f \nvdash_{PL} p_1$ and $i \in S^{\bot}_{\prec}$. Let n be the smallest index such that $\{i^f\} \nvdash_{PL} p_n$. i^f must be consistent since $i \in S^{\bot}_{\prec}$, so the finite length of i^f alone guarantees the existence of such an n. We have $[\bigcup_{j \prec i} S^{\bot}_{[\prec\downarrow j]}]^f \nvdash_{PL} p_1 \wedge \ldots \wedge p_n$ and so $\bigcup_{j \prec i} S^{\bot}_{[\prec\downarrow j]} \subseteq S^{(p_1 \wedge \ldots \wedge p_n)}_{\prec}$, where $S^{(p_1 \wedge \ldots \wedge p_n)}_{\prec} \in I \Downarrow ((p_1 \wedge \ldots \wedge p_n))$ is as described in Definition 2. $S^{\bot}_{[\prec\downarrow i]} \nsubseteq S^{(p_1 \wedge \ldots \wedge p_n)}_{\prec}$, for otherwise $P(\neg p_1/\neg(p_1 \wedge \ldots \wedge p_n))$ could not be true, so it must be that $[S^{\bot}_{[\prec\downarrow i]}]^f \vdash_{PL} p_1 \wedge \ldots \wedge p_n$. But then $[S^{(p_1 \wedge \ldots \wedge p_n)}_{\prec}]^f \vdash_{PL} i^f \rightarrow (p_1 \wedge \ldots \wedge p_n)$, which includes $[S^{(p_1 \wedge \ldots \wedge p_n)}_{\prec}]^f \vdash_{PL} i^f \rightarrow p_n$. Hence $P(i^f \wedge \neg p_n/\neg(p_1 \wedge \ldots \wedge p_n))$ cannot be true. But $P(A \wedge \neg p_n/\neg(p_1 \wedge \ldots \wedge p_n))$ must be true for all A such that $\{A\} \nvdash_{PL} p_n$ and i^f meets the condition. So Δ is not satisfiable.

References

Alchourrón, C. E. (1986). Conditionality and the Representation of Legal Norms, in Martino, A. A. and Socci Natali, F. (eds.), *Automated Analysis of Legal Texts: Edited Versions of selected papers from the Second International Conference on "Logic, Informatics, Law', Florence, Italy, September 1985*. North Holland, Amsterdam, 173–186.

Alchourrón, C. E. and Bulygin, E. The Expressive Conception of Norms, in Hilpinen, R. (1981). *New Studies in Deontic Logic*. Reidel: Dordrecht, 95–124.

Alchourrón, C. E. and Makinson, D. Hierarchies of Regulations and Their Logic, in Hilpinen, R. (1981). *New Studies in Deontic Logic*. Reidel: Dordrecht 95–124, 125–148.

Alchourrón, C. E. and Makinson, D. (1985). On the Logic of Theory Change: Safe Contraction. Studia Logica 44: 405–422.

Åqvist, L. (1986). Some Results on Dyadic Deontic Logic and the Logic of Preference. Synthese 66: 95–110.

Brewka, G. (1989). Preferred Subtheories: An Extended Logical Framework for Default Reasoninig, in Sridharan, N. S. (ed.), *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence IJCA1–89, Detroit, Michigan, USA, August 20–25, 1989*, San Mateo, Calif.: Kaufmann, 1043–1048.

Brewka, G. (1991). Belief Revision in a Framework for Default Reasoning, in Fuhrmann, A. and Morreau, M. (eds.), *The Logic of Theory Change, Workshop, Konstanz, Germany, October 13–15, 1989*. Springer, Berlin, 206–222.

Brewka, G. and Eiter, T. (1999). Preferred Answer Sets for Extended Logic Programs. Artificial Intelligence 109: 297–356.

Brink, D. O. (1994). Moral Conflict and Its Structure. Philosophical Review 103: 215–247.

Fehige, C. (1994). The Limit Assumption in Deontic (and Prohairetic) Logic, in Meggle, G. and Wessels, U. (eds.), *Analyomen 1: Proceedings of the 1st Conference "Perspectives in Analytical Philosophy", Saarbrücken 1991*. de Gruyter, Berlin, 42–56.

Gärdenfors, P. (1984). Epistemic Importance and Minimal Changes of Belief. Australasian Journal of Philosophy 62: 136–157.

Goble, L. (2005). A Logic for Deontic Dilemmas, Journal of Applied Logic: 961–983.

Hage, J. C. (1991). Monological Reason Based Reasoning, in Breuker, J. A., De Mulder, R. V., and Hage, J. C., *Legal Knowledge Based Systems: Model-based Legal Reasoning (Proceedings of the Fourth International Conference on Legal Knowledge Based Systems JURIX 1991, Lelystad, December 1991)*, 77–91.

Hage, J. C. (1996). A Theory of Legal Reasoning and a Logic to Match. Artificial Intelligence and Law 4: 199–273.

Hansen, J. (2001). Sets, Sentences, and Some Logics about Imperatives. Fundamenta Informaticae 48: 205–226.

Hansen, J. (2004) Problems and Results for Logics about Imperatives, Journal of Applied logic, 36–61.

Hansen, J. (2005) Conflicting Imperatives and Dyadic Deontic Logic, Journal of Applied Logic, 484–511.

Hansson, B. (1969). An Analysis of Some Deontic Logics. Nôus 3: 373–398, reprinted in Hilpinen, R. (1971). *Deontic Logic: Introductory and Systematic Readings*. Reidel: Dordrecht, 121–147.

Hare, R. M. (1981). *Moral Thinking*. Clarendon Press: Oxford.

Horty, J. F. (1997). Nonmonotonic Foundations for Deontic Logic, in Nute, D. (eds.), *Defeasible Deontic Logic*. Kluwer, Dordrecht, 17–44.

Horty, J. F. (2003). Reasoning with Moral Conflicts. Noûs 37: 557–605.

Iwin, A. A. (1972). Grundprobleme der deontischen Logik: in Wessel, H. (eds.), *Quantoren–Modalitäten–Paradoxien*. VEB Deutscher Verlag der Wissenschaften, Berlin, 402–522.

Kanger, S. (1957). New Foundations for Ethical Theory: Part 1", duplic., 42 p., reprinted in Hilpinen, R. (1971). *Deontic Logic: Introductory and Systematic Readings*. Reidel: Dordrecht, 36–58.

Kraus, S., Lehmann, D. and Magidor, M. (1990). Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. Artificial Intelligence 44: 167–207.

Lewis, D. (1981). Ordering Semantics and Premise Semantics for Counterfactuals. Journal of Philosophical Logic 10: 217–234.

Makinson, D. and van der Torre, L. (2000). Input/Output Logics. Journal of Philosophical Logic 29: 383–408.

Makinson, D. and van der Torre, L. (2001). Constraints for Input/Output Logics. Journal of Philosophical Logic 30: 155–185.

Marcus, R. B. (1980). Moral Dilemmas and Consistency. Journal of Philosophy 77: 121–136.

McNamara, P. (1995). The Confinement Problem: How to Terminate Your Mom With Her Trust. Analysis 55: 310–313.

Nebel, B. (1991). Belief Revision and Default Reasoning: Syntax-Based Approaches, in Allen, J. A., Fikes, R. and Sandewall, E. (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference, KR '91, Cambridge, MA, April 1991*. Morgan Kaufmann, San Mateo, 417–428.

Nebel, B. (1992). Syntax-Based Approaches to Belief Revision, in Gärdenfors, P. (eds.), *Belief Revision*. University Press, Cambridge, 52–88.

Prakken, H. (1997). *Logical Tools for Modelling Legal Argument*. Kluwer: Dordrecht.

Prakken, H. and Sartor, G. (1997). Argument-based Logic Programming with Defeasible Priorities. Journal of Applied Non-classical Logics 7: 25–75.

Rescher, N. (1964). *Hypothetical Reasoning*. North-Holland: Amsterdam.

Rintanen, J. (1994). Prioritized Autoepistemic Logic, in MacNish, C., Pearce, D. and Pereira, L. M. (eds.), *Logics in Artificial Intelligence, European Workshop, JELIA '94, York, September 1994, Proceedings*. Springer, Berlin, 232–246.

Ross, W. D. (1930). *The Right and the Good*. Clarendon Press: Oxford.

Rott, H. (1993). Belief Contraction in the Context of the General Theory of Rational Choice. Journal of Symbolic Logic 58: 1426–1450.

Ryan, M. (1992). Representing Defaults as Sentences with Reduced Priority, in Nebel, B., Rich, C. and Swartout, W. (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference, KR '92, Cambridge, MA, October 1992*. Morgan Kaufmann, San Mateo, 649–660.

Sakama, C. and Inoue, K. (1996). Representing Priorities in Logic Programs, in Maher, M. (eds.), *Joint International Conference and Syposium on Logic Programming JICSLP 1996, Bonn, September 1996*. MIT Press, Cambridge, 82–96.

Sartor, G. (1991). Inconsistency and Legal Reasoning, in Breuker, J. A., De Mulder, R. V., and Hage, J. C., *Legal Knowledge Based Systems: Model-based Legal Reasoning (Proceedings of the Fourth International Conference on Legal Knowledge Based Systems JURIX 1991, Lelystad, December 1991)*, 92–112.

Sartor, G. (2005) *Legal Reasoning. A Cognitive Approach to the Law*, vol. 5 of *A Treatise of Legal Philosophy and General Jurisprudence*, Springer: Dordrecht.

Searle, J. (1980). Prima-facie Obligations, in Straaten, Z. v. (ed.), *Philosophical Subjects: Essays presented to P. F. Strawson*. Clarendon Press, Oxford, 238–259.

Spohn, W. (1975). An Analysis of Hansson's Dyadic Deontic Logic. Journal of Philosophical Logic 4: 237–252.

von Wright, G. H. (1968). *An Essay in Deontic Logic and the General Theory of Action*. North Holland: Amsterdam.

Ziemba, Z. (1971). Deontic Syllogistics. Studia Logica 28: 139–159.

Prioritized Conditional Imperatives: Problems and a New Proposal[⋆]

Jörg Hansen

University of Leipzig, Institut für Philosophie
04107, Leipzig, Beethovenstraße 15, Germany
jhansen@uni-leipzig.de

Abstract. The sentences of deontic logic may be understood as describing what an agent ought to do when faced with a given set of norms. If these norms come into conflict, the best the agent can be expected to do is to follow a maximal subset of the norms. Intuitively, a priority ordering of the norms can be helpful in determining the relevant sets and resolve conflicts, but a formal resolution mechanism has been difficult to provide. In particular, reasoning about prioritized conditional imperatives is overshadowed by problems such as the 'order puzzle' that are not satisfactorily resolved by existing approaches. The paper provides a new proposal as to how these problems may be overcome.

Keywords: deontic logic, default logic, priorities, logic of imperatives

1 Drinking and Driving

Imagine you have been invited to a party. Before the event, you become the recipient of various imperative sentences:

(1) Your mother says: if you drink anything, then don't drive.
(2) Your best friend says: if you go to the party, then you do the driving.
(3) Some acquaintance says: if you go to the party, then have a drink with me.

Suppose that as a rule you do what your mother tells you – after all, she is the most important person in your life. Also, the last time you went to a party your best friend did the driving, so it really is your turn now. You can enjoy yourself without a drink, though it would be nice to have a drink with your acquaintance – your best friend would not mind if you had one drink, and your acquaintance does not care that you may be driving – but your mother would not approve of such a behavior. Making up your mind,

(4) You go to the party.

I think it is quite clear what you must do: obey your mother and your best friend, and hence do the driving and deny your acquaintance's request. However, it is not so clear what formal algorithm could explain this reasoning.

[⋆] I am grateful to Lou Goble, John F. Horty and Leon van der Torre for helpful comments and discussions in preparing this paper.

An example of a similar form was first employed in epistemic logic,[1] and has been termed the 'order puzzle' (cf. Horty [22]). For the epistemic version, consider the following sentences:

(5) You remember from physics: if you are in a car, lightning won't strike you.
(6) The coroner tells you: he was struck by lightning.
(7) Your neighbor says: he must have been drinking and driving.

Suppose that driving includes being in a car, that you firmly believe in what you remember from physics, that you believe that information by medical officers is normally based on competent investigation, and that you usually don't question your neighbor's observations, but think that sometimes she is just speculating. It seems quite clear what happens: you keep believing what you remember from school, and don't doubt what the coroner told you, but question your neighbor's information, maybe answering: "This can't be true, as the authorities found he was struck by lightning, and you can't be struck by lightning in a car".

In both cases, the problem as to how the underlying reasoning can be formally reconstructed seems so far unsolved. Both involve a ranking, or priority ordering, of the sentences involved. Concentrating on the imperative side of things, in what follows, I will consider various proposals from the literature that have been put forward to explain the reasoning about such prioritized conditionals, discuss their strengths and weaknesses in relation to problems such as the one above, and finally propose a fresh solution that solves the problem.

2 Formal Preliminaries

To formally discuss problems such as the one presented above, I shall use a simple framework: let I be a set of objects, they are meant to be (conditional) imperatives. Two functions g and f associate with each imperative an antecedent and a consequent – these are sentences from the language of a basic logic that here will be the language $\mathscr{L}_{PL}$ of propositional logic.[2] $g(i)$ may be thought of as describing the 'grounds', or circumstances in which the consequent of i is to hold, and $f(i)$ as associating the sentence that describes what must be the case if the imperative i is satisfied, its 'deontic focus' or 'demand'.[3] In accordance with tradition (cf. Hofstadter and McKinsey [20]), I write $A \Rightarrow !B$ for an $i \in I$ with $g(i) = A$ and $f(i) = B$, and $!A$ means an unconditional imperative $\top \Rightarrow !A$. Note

[1] Cf. Rintanen [36] p. 234, who in turn credits Brewka with its invention.

[2] PL is based on a language $\mathscr{L}_{PL}$, defined from a set of proposition letters $Prop = \{p_1, p_2, ...\}$, Boolean connectives '$\neg$', '$\wedge$', '$\vee$', '$\rightarrow$', '$\leftrightarrow$' and brackets '(', ')' as usual. The truth of a $\mathscr{L}_{PL}$-sentence A is defined recursively using a valuation function $v : Prop \rightarrow \{1, 0\}$ (I write $v \models A$), starting with $v \models p$ iff $v(p) = 1$ and continuing as usual. If $A \in \mathscr{L}_{PL}$ is true for all valuations it is called a tautology. PL is the set of all tautologies, and this set is used to define provability, consistency and derivability (I write $\Gamma \vdash_{PL} A$) as usual. $\top$ is an arbitrary tautology, and $\bot$ is $\neg\top$.

[3] In analogy to Reiter's default logic one might add a third function e that describes exceptional circumstances in which the imperative is not to be applied. I will not address this additional complexity here.

that $A \Rightarrow !B$ is just the name for a conditional imperative that demands B to be made true in a situation where A is true – it is not an object that is assigned truth values. I write $m(i)$ for $\ulcorner g(i) \to f(i) \urcorner$ and call $m(i)$ the 'materialization' of i, as it represents the material implication that may be thought of as corresponding to the conditional imperative. For any $i \in I$ and $\Delta \subseteq I$, instead of $f(i)$, $g(i)$, $m(i)$, $f(\Delta)$, $g(\Delta)$ and $m(\Delta)$, I may use the superscripted i^f, i^g, i^m, Δ^f, Δ^g and Δ^m for better readability.

Let $\mathcal{I}$ be a tuple $\langle I, f, g \rangle$, let $W \subseteq \mathscr{L}_{PL}$ be a set of sentences, representing 'real world facts', and $\Delta \subseteq I$ be a subset of the imperatives: then we define

$$\mathit{Triggered}_{\mathcal{I}}(W, \Delta) =_{df} \{i \in \Delta \mid W \vdash_{PL} g(i)\}.$$

So an imperative $i \in \Delta$ is triggered if its antecedent is true given W. Tradition wants it that a conditional imperative can only be fulfilled or violated if its condition is the case.[4] So I define:

$$\mathit{Satisfied}_{\mathcal{I}}(W, \Delta) =_{df} \{i \in \Delta \mid W \vdash_{PL} i^g \wedge i^f\},$$
$$\mathit{Violated}_{\mathcal{I}}(W, \Delta) =_{df} \{i \in \Delta \mid W \vdash_{PL} i^g \wedge \neg i^f\},$$

An imperative in $\mathit{Satisfied}_{\mathcal{I}}(W, \Delta)$ [$\mathit{Violated}_{\mathcal{I}}(W, \Delta)$] is called satisfied [violated] given the facts W. It is of course possible that an imperative is neither satisfied nor violated given the facts W. If an imperative is triggered, but not violated, we call the imperative satisfiable:

$$\mathit{Satisfiable}_{\mathcal{I}}(W, \Delta) =_{df} \{i \in \mathit{Triggered}_{\mathcal{I}}(W, \Delta) \mid W \nvdash_{PL} \neg i^f\}.$$

Moreover, we define

$$\mathit{Obeyable}_{\mathcal{I}}(W, \Delta) =_{df} \{\Gamma \subseteq \Delta \mid \Gamma^m \cup W \nvdash_{PL} \bot\}.$$

So a subset Γ of Δ is obeyable given W iff it is not the case that for some $\{i_1, ..., i_n\} \subseteq \Gamma$ we have $W \vdash_{PL} (i_1^g \wedge \neg i_1^f) \vee ... \vee (i_n^g \wedge \neg i_n^f)$: otherwise we know that whatever we do, i.e. given any maxiconsistent subset V of $\mathscr{L}_{PL}$ that extends $W \subseteq V$, at least one imperative in Γ is violated.[5] We speak of a *conflict of imperatives* when the triggered imperatives cannot all be satisfied given the facts W, i.e. when $\mathit{Triggered}_{\mathcal{I}}(W, \Delta)^f \cup W \vdash_{PL} \bot$. More generally speaking I will also call imperatives conflicting if they are not obeyable in the given situation.

As prioritized conditional imperatives are our concern here, we let all imperatives in I be ordered by some priority relation $< \subseteq I \times I$. The relation $<$ is assumed to be a strict partial order on I, i.e. $<$ is irreflexive and transitive, and additionally we assume $<$ to be well-founded, i.e. infinite descending chains are excluded. For any $i_1, i_2 \in I$, $i_1 < i_2$ means that i_1 takes priority over i_2 (ranks higher than i_2, is more important than i_2, etc.). A tuple $\langle I, f, g \rangle$ will be called a *conditional imperative structure*, and $\langle I, f, g, < \rangle$ a *prioritized conditional imperative structure*. If all imperatives in I are unconditional, we may drop any reference to the relation g in the tuples and call these *basic imperative structures* and *prioritized imperative structures* respectively.

[4] Cf. Rescher [35], Sosa [40], van Fraassen [10]. Also cf. Greenspan [12]: "Oughts do not arise, it seems, until it is too late to keep their conditions from being fulfilled."

[5] Terms differ here, e.g. Downing [8] uses the term 'compliable' instead of 'obeyable'.

3 Deontic Concepts

Given a set of imperatives, one may truly or falsely state that their addressee must, or must not, perform some act or achieve some state of affairs according to what the addressee was ordered to do. For instance, in the 'drinking and driving' example from sec. 1 I think it is true that the agent ought to do the driving, as this is what the second-ranking imperative, uttered by the best friend, requires the agent to do, but that it would be false to say that the agent ought to drink and drive. Statements that something ought to be done or achieved are called 'normative' or 'deontic statements', and the ultimate goal is to find a logical semantics that models the situation and defines the deontic concepts in such a way that the formal results coincide with our natural inclinations in the matter.

3.1 Deontic operators for unconditional imperatives

For unconditional imperatives, such definitions are straightforward. Given a basic imperative structure $\mathcal{I} = \langle I, f \rangle$, a monadic deontic O-operator is defined by

$(td\text{-}m1)$ $\quad \mathcal{I} \models OA$ if and only if (iff) $I^f \vdash_{PL} A$.

So obligation is defined in terms of what the satisfaction of all imperatives logically implies. With the usual truth definitions for Boolean operators, it can easily be seen that such a definition produces a normal modal operator, i.e. one that is defined by the following axiom schemes plus *modus ponens*:

(Ext) If $\vdash_{PL} A \leftrightarrow B$, then $OA \leftrightarrow OB$ is a theorem.
(M) $O(A \wedge B) \rightarrow (OA \wedge OB)$
(C) $(OA \wedge OB) \rightarrow O(A \wedge B)$
(N) $O\top$

Furthermore, $(td\text{-}m1)$ defines standard deontic logic *SDL*, which adds

(D) $OA \rightarrow \neg O \neg A$

iff the imperatives are assumed to be non-conflicting and so I^f is PL-consistent, i.e. $I^f \nvdash_{PL} \bot$. It is immediate that in the case of conflicts, $(td\text{-}m1)$ pronounces everything as obligatory, and in particular defines $O\bot$ true, thus making the impossible obligatory. If conflicts are not excluded, a solution is to only consider (maximal) subsets of the imperatives whose demands are consistent and define the O-operator with respect to these (I write $I \curlywedge \neg C$ for the set of all '$\neg C$-remainders', i.e. maximal subsets Γ of I such that $\Gamma^f \nvdash_{PL} \neg C$):

$(td\text{-}m2)$ $\quad \mathcal{I} \models OA$ iff $\forall \Gamma \in I \curlywedge \bot : \Gamma^f \vdash_{PL} A$

Quite similarly, a dyadic deontic operator $O(A/C)$, meaning that A ought to be true given that C is true, can be defined with respect to the maximal subsets of imperatives that do not conflict in these circumstances:

$(td\text{-}d1)$ $\quad \mathcal{I} \models O(A/C)$ iff $\forall \Gamma \in I \curlywedge \neg C : \Gamma^f \vdash_{PL} A$

So A is obligatory given that C is true if A is what the imperatives in any $\neg C$-remainder demand. In the case of conflicts, this definition produces a "disjunctive solution": e.g. if there are two imperatives $!A$ and $!B$ with $\vdash_{PL} C \rightarrow (A \rightarrow \neg B)$, then neither $O(A/C)$ nor $O(B/C)$ but $O(A \vee B/C)$ is true.[6]

[6] For alternative solutions to the problem of conflicts cf. Goble [11] and my [13], [14].

Often, we want to use the information that we have about the circumstances also for reasoning about the obligations in these circumstances. E.g. if the set of imperatives is $\{!(p_1 \vee p_2)\}$, ordering me to either send you a card or phone you, and I cannot send you a card, i.e. $\neg p_1$ is true, I should be able to conclude that I should phone you, and so $O(p_2/\neg p_1)$ should be true. Such 'circumstantial reasoning' is achieved by the following change to the truth definition:

(td-$d2$) $\quad \mathcal{I} \models O(A/C)$ iff $\forall \Gamma \in I \curlywedge \neg C : \Gamma^f \cup \{C\} \vdash_{PL} A$

With the usual truth conditions for Boolean operators, a semantics that employs (td-$d2$) has a sound and (weakly) complete axiom system PD that equals the system P of Kraus, Lehmann, Magidor [23], defined by these axiom schemes

(DExt) If $\vdash_{PL} A \leftrightarrow B$ then $O(A/C) \leftrightarrow O(B/C)$ is a theorem.
(DM) $O(A \wedge B/C) \rightarrow (O(A/C) \wedge O(B/C))$
(DC) $O(A/C) \wedge O(B/C) \rightarrow O(A \wedge B/C)$
(DN) $O(\top/C)$
(ExtC) If $\vdash_{PL} C \leftrightarrow D$ then $O(A/C) \leftrightarrow O(A/D)$ is a theorem.
(CCMon) $O(A \wedge D/C) \rightarrow O(A/C \wedge D)$
(CExt) If $\vdash_{PL} C \rightarrow (A \leftrightarrow B)$ then $O(A/C) \leftrightarrow O(B/C)$ is a theorem.
(Or) $O(A/C) \wedge O(A/D) \rightarrow O(A/C \vee D)$

with the additional (restricted) dyadic 'deontic' axiom scheme

(DD-R) If $\nvdash_{PL} \neg C$ then $\vdash_{PD} O(A/C) \rightarrow \neg O(\neg A/C)$

added (hence the name PD).[7]

3.2 Deontic operators for conditional imperatives

Unlike their unconditional counterparts, conditional imperatives have been found hard to reason about. G. H. von Wright [47] called conditional norms the "touchstone of normative logic", and van Fraassen [10] wrote with regard to logics for conditional imperatives: "There may be systematic relations governing this moral dynamics, but I can only profess ignorance of them."

Representing a conditional imperative as an unconditional imperative that demands a material conditional to be made true yields undesired results. Most notorious is the problem of contraposition: consider a set I with the only imperative $!(p_1 \rightarrow p_2)$, meaning e.g. 'if the police stops you, show your drivers licence'. (td-$d1$) makes true $O(p_2/p_1)$, but also $O(\neg p_1/\neg p_2)$, so if you can't present your drivers licence (you don't have one) you must see to it that the police does not stop you, which is hardly what the speaker meant you to do.One may think that such problems arise from the fact that antecedents of conditional imperatives often describe states of the affairs that the agent is not supposed to, and often cannot, control. But consider the set $\{!(p_1 \rightarrow p_2), !(\neg p_1 \rightarrow p_3)\}$, it yields $O(p_2/\neg p_3)$ with (td-$d1$). Here, p_2 is what the consequent of some imperative demands, so it supposedly describes something the agent can control. Now let

[7] For proofs, and an additional "credulous ought" that defines $O(A/C)$ true if the truth of A is required to satisfy all imperatives in *some* $\neg C$-remainder, cf. my [14].

the imperatives be interpreted as ordering me to wear my best suit if it does not rain, and a rain coat if it does: it is clear nonsense that I am obliged to wear a raincoat given that I can't wear my best suit (e.g. it is in the laundry). Such problems are the reason why we cautiously use special models for conditional imperatives (i.e. conditional imperative structures), and write $p_1 \Rightarrow !p_2$ instead of $!(p_1 \rightarrow p_2)$. But this only delegates the problem from the level of representation to that of semantics, where now new truth definitions must be found.

Let $\mathcal{I} = \langle I, f, g \rangle$ be a conditional imperative structure, and let us ignore for the moment the further complication of possible conflicts between imperatives. Then the following seems a natural way to define what ought to be the case in circumstances where C is assumed to be true:

$$(td\text{-}cd1) \quad \mathcal{I} \models O(A/C) \ \text{ iff } \ [\mathit{Triggered}_{\mathcal{I}}(\{C\}, I)]^f \vdash_{PL} A$$

So dyadic obligation is defined in terms what is necessary to satisfy all imperatives that are triggered in the assumed circumstances. E.g. if $I = \{p_1 \Rightarrow !p_2\}$, with its only imperative interpreted as "if you have a cold, stay in bed", then $O(p_2/p_1)$ truly states that I must stay in bed given that I have a cold.

Like in the unconditional case, it seems important to be able to use 'circumstantial reasoning', i.e. employ the information about the situation not only to determine if an imperative is triggered, but also for reasoning with its consequent. E.g. if the set of imperatives is $\{p_1 \Rightarrow !(p_2 \vee p_3)\}$, with its imperative interpreted as expressing "if you have a cold, either stay in bed or wear a scarf", one would like to obtain $O(p_3/p_1 \wedge \neg p_2)$, expressing that given that I have a cold and don't stay in bed, I must wear a scarf. So $(td\text{-}cd1)$ may be changed into

$$(td\text{-}cd2) \quad \mathcal{I} \models O(A/C) \ \text{ iff } \ [\mathit{Triggered}_{\mathcal{I}}(\{C\}, I)]^f \cup \{C\} \vdash_{PL} A.$$

Though the step from $(td\text{-}cd1)$ to $(td\text{-}cd2)$ seems quite reasonable, such definitions have also been criticized for defining the assumed circumstances as obligatory. E.g. if the set of imperatives is $\{p_1 \Rightarrow !p_2\}$, where the imperative is interpreted as expressing "if you hit someone, apologize to him", then (td-5) makes true $O(p_1 \wedge p_2/p_1)$, and hence also $O(p_1/p_1)$, so given that I hit someone, this is something I ought to do. The criticism looses much of its edge in the present setting, where one can point to the distinction between imperatives (there is no imperative that demands p_1) and ought sentences that describe what must be true when all triggered imperatives are satisfied in the supposed circumstances: then the truth of $O(p_1/p_1)$ seems no more paradoxical than the truth of $O\top$ that is accepted in most systems of deontic logic.

3.3 Further modifications

In Makinson & van der Torre's [25] more general theory of 'input/output logic', $(td\text{-}cd1)$ is termed 'simple-minded output', and $(td\text{-}cd2)$ is its 'throughput version'.[8]As the names suggests, the authors also discuss more refined operations, which again might be considered for reasoning about conditional imperatives. One modification addresses the possibility of 'reasoning by cases' that e.g. makes

[8] If I resembles the generating set G of input/output logic, then $O(A/C)$ means that A is an output given the input C (Makinson & van der Torre write $A \in \mathit{out}(G, \{C\})$).

true $O(p_2 \vee p_4/p_1 \vee p_3)$ for a set of imperatives $I = \{p_1 \Rightarrow !p_2, p_3 \Rightarrow !p_4\}$. This may be achieved by the following definition, where $\mathscr{L}_{PL} \bot \neg C$ is the set of all maximal subsets of the language $\mathscr{L}_{PL}$ that are consistent with C:[9]

$$(td\text{-}cd3) \quad \mathcal{I} \models O(A/C) \ \text{ iff } \ \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}_{\mathcal{I}}(V, I)]^f \vdash_{PL} A$$

In the example, each set $V \subset \mathscr{L}_{PL}$ that is maximally consistent with $p_1 \vee p_3$ either contains p_1, then $p_1 \Rightarrow !p_2$ is triggered and so p_2 and also $p_2 \vee p_4$ is implied by $[\mathit{Triggered}_{\mathcal{I}}(V, I)]^f$, or it contains $\neg p_1$, but then it cannot also contain $\neg p_3$ and so must contain p_3, so $p_3 \Rightarrow !p_4$ is triggered and therefore p_4 and also $p_2 \vee p_4$ implied, so for all sets V, $p_2 \vee p_4$ is implied and so $O(p_2 \vee p_4/p_1 \vee p_3)$ made true.

In order to add 'circumstantial reasoning' to (td-cd3) – or, in Makinson & van der Torre's terms, for its 'throughput version' –, one might, in the vein of (td-d2) and (td-cd2), try this definition:

$$(td\text{-}cd4^-) \ \mathcal{I} \models O(A/C) \ \text{ iff } \ \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}_{\mathcal{I}}(V, I)]^f \cup \{C\} \vdash_{PL} A$$

But the definition seems too weak. Consider the set $\{p_1 \Rightarrow !(\neg p_2 \vee p_4), p_3 \Rightarrow !p_4\}$ and the situation $(p_1 \wedge p_2) \vee p_3$. We would expect a reasoning as follows: in this situation, either $p_1 \wedge p_2$ is true, so the first imperative is triggered but we cannot satisfy it by bringing about $\neg p_2$, and so must bring about p_4. Or p_3 is true, then the second imperative is triggered and we must again bring about p_4. So we must bring about p_4 in the given situation. But the definition fails to make true $O(p_4/(p_1 \wedge p_2) \vee p_3)$. Like Makinson and van der Torre [25], I therefore combine reasoning by cases with a stronger version of throughput:

$$(td\text{-}cd4) \quad \mathcal{I} \models O(A/C) \ \text{ iff } \ \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}_{\mathcal{I}}(V, I)]^f \cup V \vdash_{PL} A$$

As is easy to see, this resolves the difficulty: for $\{p_1 \Rightarrow !(\neg p_2 \vee p_4), p_3 \Rightarrow !p_4\}$, $O(p_4/(p_1 \wedge p_2) \vee p_3)$ is now true, as desired. However, this modification has a surprising consequence: it makes the reasoning about conditional imperatives collapse into reasoning about consequences of their materializations:

Observation 1 *By (td-cd4),* $\mathcal{I} \models O(A/C)$ *iff* $m(I) \cup \{C\} \vdash_{PL} A$.

Proof. For the right-to-left direction, for any imperative $i \in I$ and any set $V \in \mathscr{L}_{PL} \bot \neg C$, either V includes $g(i)$, so $i \in \mathit{Triggered}_{\mathcal{I}}(V, I)$ and therefore $[\mathit{Triggered}_{\mathcal{I}}(V, I)]^f \vdash_{PL} g(i) \rightarrow f(i)$, or it does not include $g(i)$, but then it includes $\neg g(i)$ by maximality, hence $V \vdash_{PL} g(i) \rightarrow f(i)$. So $[\mathit{Triggered}_{\mathcal{I}}(V, I)]^f \cup V \vdash_{PL} g(i) \rightarrow f(i)$. For the left-to-right direction, if $m(I) \cup \{C\} \nvdash_{PL} A$ then $m(I) \cup \{C\} \cup \{\neg A\}$ is consistent, so there is a $V \in \mathscr{L}_{PL} \bot \neg C$ such that $m(I) \cup \{C\} \cup \{\neg A\} \subseteq V$. It is immediate that for each $i \in \mathit{Triggered}_{\mathcal{I}}(V, I)$, $m(I) \cup V \vdash_{PL} f(i)$, so if $[\mathit{Triggered}_{\mathcal{I}}(V, I)]^f \cup V \vdash_{PL} A$ then $m(I) \cup V \vdash_{PL} A$ and since $m(I) \subseteq V$ also $V \vdash_{PL} A$. Since V was consistent and included $\neg A$, it cannot also derive A, and so by contraposition $[\mathit{Triggered}_{\mathcal{I}}(V, I)]^f \cup V \nvdash_{PL} A$.

But such an equivalence makes all the problems discussed above for identifying conditional imperatives with unconditional imperatives that demand their mate-

[9] Makinson & van der Torre's [25] call the resulting operator 'basic output', of which a syntactical version was first presented by Świrydowicz [41] p. 32.

rializations reappear, in particular the problem of contraposition.[10] So it seems we must choose between 'reasoning by cases' and 'circumstantial reasoning'.[11]

Another modification that these authors consider is that of 'reusable output': when an imperative is triggered that demands A, and A is the trigger for some imperative $A \Rightarrow !B$, then we also ought to do B. Such a modification can easily be incorporated into a truth definition and its 'throughput' version:

$$(\mathit{td\text{-}cd}5) \quad \mathcal{I} \models O(A/C) \text{ iff } [\mathit{Triggered}^*_{\mathcal{I}}(\{C\}, I)]^f \vdash_{PL} A$$
$$(\mathit{td\text{-}cd}6) \quad \mathcal{I} \models O(A/C) \text{ iff } [\mathit{Triggered}^*_{\mathcal{I}}(\{C\}, I)]^f \cup \{C\} \vdash_{PL} A$$

where $\mathit{Triggered}^*_{\mathcal{I}}(W, \Gamma)$ means the smallest subset of $\Gamma \subseteq I$ such that for all $i \in \Gamma$, if $[\mathit{Triggered}^*_{\mathcal{I}}(W, \Gamma)]^f \cup W \vdash_{PL} g(i)$ then $i \in \mathit{Triggered}^*_{\mathcal{I}}(W, \Gamma)$. Moreover, the two modifications of 'reasoning by cases' and 'reusable output' can be combined to produce the following definition and its 'throughput' variant:

$$(\mathit{td\text{-}cd}7) \quad \mathcal{I} \models O(A/C) \text{ iff } \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}^*_{\mathcal{I}}(V, I)]^f \vdash_{PL} A$$
$$(\mathit{td\text{-}cd}8) \quad \mathcal{I} \models O(A/C) \text{ iff } \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}^*_{\mathcal{I}}(V, I)]^f \cup V \vdash_{PL} A$$

The topic of 'reusable output' is discussed under the name of 'deontic detachment' in the literature on deontic logic, and there is no agreement whether such a procedure is admissible (Makinson [24] p. 43 argues in favor, whereas Sven Ove Hansson [17] p. 155 disagrees). E.g. let $I = \{!p_1, p_1 \Rightarrow !p_2, \neg p_1 \Rightarrow !\neg p_2\}$, and for its interpretation assume that it is imperative for the proper execution of your job that you develop novel methods, which make you eligible for a bonus, and that if you develop such novel methods you owe it to yourself to apply for the bonus, but that if you don't develop such methods you must not apply for the bonus. Truth definitions that accept 'deontic detachment' make true $O(p_2/\top)$, and so tell us that you ought to apply for the bonus, which seems weird since it may be that you never invent anything. However, proponents of deontic detachment may argue that in such a situation, $O(p_1 \wedge p_2/\top)$ should hold, i.e. you ought to invent new methods *and* apply for the bonus, and that the reluctance to also accept $O(p_2/\top)$ is – like the inference from "you ought to put on your parachute and jump" to "you ought to jump" – just a variant of Ross' Paradox that is usually considered harmless.

For $(\mathit{td\text{-}cd}7)$ we once again obtain $O(p_2/\neg p_3)$ for $I = \{p_1 \Rightarrow !p_2, \neg p_1 \Rightarrow !p_3\}$: for any $V \in \mathscr{L}_{PL} \bot p_3$, $\neg p_3 \in V$, furthermore either $p_1 \in V$ and so $p_1 \Rightarrow !p_2 \in \mathit{Triggered}^*_{\mathcal{I}}(V, I)$, or $\neg p_1 \in V$, then $\neg p_1 \Rightarrow !p_3 \in \mathit{Triggered}^*_{\mathcal{I}}(V, I)$, and since $\{p_3\} \cup$

[10] $(\mathit{td\text{-}cd}4^-)$ does not fare much better: though it does not include contraposition, it again makes $O(p_2/\neg p_3)$ true for $I = \{p_1 \Rightarrow !p_2, \neg p_1 \Rightarrow !p_3\}$, which is counterintuitive.

[11] Legal use of 'reasoning by cases', or *Wahlfeststellung*, is controversial. It means that if the defendant either committed crime α or crime β, the defendant would be convicted according to the milder law. A proponent would argue that since the defendant committed a crime (though it remains open which), justice demands that he should not go free, while the defense would argue that this violates the *in dubio pro reo* principle, since neither charge is sufficiently proved. After a *Reichsgericht* ruling in 1934 allowed *Wahlfeststellung* for cases in which the crimes in question were 'ethically and psychologically equivalent', the national-socialist lawmakers introduced a law prescribing its unrestricted application in 1935, considered ideological and lifted again by the Allied Control Council of Germany in 1946 (cf. [43]).

$\{\neg p_3\} \vdash_{PL} p_1$, again $p_1 \Rightarrow !p_2$ is in $\mathit{Triggered}^*_{\mathcal{I}}(V, I)$, hence $[\mathit{Triggered}^*_{\mathcal{I}}(V, I)]^f \vdash_{PL} p_2$ for all $V \in \mathscr{L}_{PL} \bot p_3$. But as we saw above, when interpreting the imperatives as ‘if it rains, wear a raincoat’ and ‘if it does not rain, wear your best jacket’, this result seems counterintuitive.[12] Note that (*td-cd*8) is again equivalent to $\mathcal{I} \models O(A/C)$ iff $m(I) \cup \{C\} \vdash_{PL} A$ and thus to (*td-cd*4) (cf. Makinson & van der Torre [25] observation 16; [26], p. 156):

Observation 2 *By* (*td-cd*8), $\mathcal{I} \models O(A/C)$ *iff* $m(I) \cup \{C\} \vdash_{PL} A$.

Proof. Similar to the proof of observation 1. For the left-to-right direction, use that for each $i \in \mathit{Triggered}^*_{\mathcal{I}}(V, I)$, $m(I) \cup V \vdash_{PL} f(i)$, which is immediate.

3.4 Operators for prioritized conditional imperatives

This paper focuses on prioritized conditional imperatives, and for these there is a further hurdle to finding the proper truth definitions for deontic concepts. Priorities are only required if the imperatives cannot all be obeyed – otherwise there is no reason not to obey all, and the priority ordering is not used. So the truth definitions must be able to deliver meaningful results for possibly conflicting imperatives. The intuitive idea is to use the information in the ordering to choose subsets of the set of imperatives under consideration that contain only the more important imperatives and leave out less important, conflicting ones, so that the resulting ‘preferred subset’ (or rather, subsets, since the choice may not always be determined by the ordering) only contains imperatives that do not conflict in the given situation. More generally, let $\mathcal{I}$ be a prioritized conditional imperative structure $\langle I, g, f, < \rangle$, and let Δ be a subset of I. Then $\mathscr{P}_{\mathcal{I}}(W, \Delta)$ contains just the subsets of Δ that are thus preferred given the world facts W. The above truth definitions can then be adapted such that they now describe something as obligatory iff it is so with respect to all the preferred subsets of the imperatives, i.e. they take on the following forms:

$$\mathcal{I} \models O(A/C) \text{ iff } \forall \Gamma \in \mathscr{P}_{\mathcal{I}}(\{C\}, I):$$

$$(\mathit{td\text{-}pcd}1) \qquad [\mathit{Triggered}_{\mathcal{I}}(\{C\}, \Gamma)]^f \vdash_{PL} A\,,$$

$$(\mathit{td\text{-}pcd}2) \qquad [\mathit{Triggered}_{\mathcal{I}}(\{C\}, \Gamma)]^f \cup \{C\} \vdash_{PL} A\,,$$

$$(\mathit{td\text{-}pcd}3) \qquad \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}_{\mathcal{I}}(V, \Gamma)]^f \vdash_{PL} A\,,$$

$$(\mathit{td\text{-}pcd}4) \qquad \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}_{\mathcal{I}}(V, \Gamma)]^f \cup V \vdash_{PL} A\,,$$

$$(\mathit{td\text{-}pcd}5) \qquad [\mathit{Triggered}^*_{\mathcal{I}}(\{C\}, \Gamma)]^f \vdash_{PL} A\,,$$

$$(\mathit{td\text{-}pcd}6) \qquad [\mathit{Triggered}^*_{\mathcal{I}}(\{C\}, \Gamma)]^f \cup \{C\} \vdash_{PL} A\,,$$

$$(\mathit{td\text{-}pcd}7) \qquad \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}^*_{\mathcal{I}}(V, \Gamma)]^f \vdash_{PL} A\,,$$

$$(\mathit{td\text{-}pcd}8) \qquad \forall V \in \mathscr{L}_{PL} \bot \neg C : [\mathit{Triggered}^*_{\mathcal{I}}(V, \Gamma)]^f \cup V \vdash_{PL} A\,.$$

So e.g. (*td-pcd*1) defines A as obligatory if the truth of A is required to satisfy the triggered imperatives in any preferred subset. Of course, the crucial and as yet missing element is the decision procedure that determines the set $\mathscr{P}_{\mathcal{I}}(\{C\}, I)$ of preferred subsets. The next section discusses several proposals to define such subsets; a new proposal is presented in the section that follows it.

[12] With respect to their out_4-operation that corresponds to (*td-cd*7), Makinson & van der Torre [25] speak of a ‘ghostly contraposition’.

4 Identifying the Preferred Subsets

4.1 Brewka's preferred subtheories

The idea that normative conflicts can be overcome by use of a priority ordering of the norms involved dates back at least to Ross [37] and is also most prominent in von Wright's work (cf. [45] p. 68, 80). However, it has turned out to be difficult to determine the exact mechanism by which such a resolution of conflicts can be achieved. This is true even when only unconditional imperatives are considered, and when special problems are left out of the picture, such that the ordering itself might be dependent on the facts (e.g. when the command of an officer in the field may override that of her superior due to unexpected circumstances), or be the subject of normative regulation (e.g. when we are commanded to obey the law of God more than the law of man). Discussing various proposals for resolution of conflicts between unconditional imperative, I have argued in [15] that an 'incremental' definition be used for determining the relevant subsets. Based on earlier methods by Rescher [34], such a definition was first introduced by Brewka [4] for reasoning with prioritized defaults. For any priority relation $<$, the idea is to consider all the 'full prioritizations' $\prec$ of $<$ (strict well orders that preserve $<$), and then work ones way from top to bottom by adding the $\prec$-next-higher imperative to the thus constructed 'preferred subtheory' if its demand is consistent with the given facts and the demands of the imperatives that were added before. For the present setting, the definition can be given as follows:

Definition 1 (Brewka's preferred subtheories).
Let $\mathcal{I} = \langle I, f, g, < \rangle$ be a prioritized conditional imperative structure, Δ be a subset of I, and $W \subseteq \mathscr{L}_{PL}$ be a set of PL-sentences. Then $\Gamma \in \mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(W, \Delta)$ iff (i) *$W \nvdash_{PL} \bot$, and* (ii) *Γ is obtained from a full prioritization $\prec$ by defining*

$$\Gamma_{[\prec\downarrow i]} = \begin{cases} \bigcup_{j \,\prec\, i} \Gamma_{[\prec\downarrow j]} \cup \{i\} & \text{if } W \cup [\bigcup_{j \,\prec\, i} \Gamma_{[\prec\downarrow j]} \cup \{i\}]^f \nvdash_{PL} \bot, \text{ and} \\ \bigcup_{j \,\prec\, i} \Gamma_{[\prec\downarrow j]} & \text{otherwise,} \end{cases}$$

for any $i \in \Delta$, and letting $\Gamma = \bigcup_{i \in \Delta} \Gamma_{[\prec\downarrow i]}$.

Clause (i) ensures that for an inconsistent set of assumed 'facts', no set is preferred. Somewhat roundabout, owed to the possibility of infinite ascending subchains, clause (ii) then recursively defines a set $\Gamma \in \mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(W, \Delta)$ for each full prioritization $\prec$: take the $\prec$-first i (the exclusion of infinite descending subchains guarantees that it exists) and if $W \cup \{i^f\} \nvdash_{PL} \bot$ then let $\Gamma_{[\prec\downarrow i]} = \{i\}$; otherwise $\Gamma_{[\prec\downarrow i]}$ is left empty.[13] Similarly, any $\prec$-later i is tested for possible addition to the set $\bigcup_{j \,\prec\, i} \Gamma_{[\prec\downarrow j]}$ of elements that were added in the step for a $j \in \Delta$ that occurs $\prec$-prior to i. Γ is then the union of all these sets.

To see how this definition works, consider the set $I = \{!(p_1 \vee p_2), !\neg p_2, !\neg p_1\}$, with the ranking $!(p_1 \vee p_2) < !\neg p_1$ and $!\neg p_2 < !\neg p_1$. For an interpretation, let $!(p_1 \vee p_2))$ be your mother's request that you buy cucumbers or spinach for dinner, $!\neg p_1$ be your father's wish that no cucumbers are bought, and $!\neg p_2$ your

[13] As usual, the union of an empty set of sets is taken to be the empty set.

sister's desire that you don't buy any spinach. We have two full prioritizations $!(p_1 \vee p_2) < !\neg p_2 < !\neg p_1$ and $!\neg p_2 < !(p_1 \vee p_2) < !\neg p_1$ – let these be termed $\prec_1$ and $\prec_2$, respectively. The construction for $\prec_1$ adds the imperative $!(p_1 \vee p_2)$ in the first step and, since no conflict with the situation arises, $!\neg p_2$ in the second step. In the third and last step, nothing is added since $!\neg p_1$ conflicts with the demands of the already added imperatives. For $\prec_2$ the only difference is that the first two imperatives are added in inverse order. Thus $\mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(W, I) = \{\{!(p_1 \vee p_2), !\neg p_2\}\}$. Using (*td-pcd*2) we obtain $O(p_1 \wedge \neg p_2/\top)$, which means that you have to buy spinach and not cucumbers, thus fulfilling your parents' requests but not your sister's, which seems reasonable.

As I showed in [15], Brewka's method is extremely successful for dealing with unconditional imperatives. It is provably equivalent for such imperatives to methods proposed by Ryan [38] and Sakama & Inoue [39], and it avoids problems of other approaches by Alchourrón & Makinson [2], Prakken [31] and Prakken & Sartor [32]. Moreover, an equally intuitive maximization method proposed by Nebel [29], [30], that adds first a maximal number of the highest-ranking imperatives, then a maximal number of the second-ranking imperatives, etc., but for its construction requires the ordering to be based on a complete preorder, can be shown to be embedded in Brewka's approach for such orderings. So my aim will be to retain Brewka's method for the unconditional case. However, when it is applied without change to conditional imperatives, the algorithm may lead to incorrect results. E.g. consider a set I with two equally ranking imperatives $\{p_1 \Rightarrow !p_2, \neg p_1 \Rightarrow !\neg p_2\}$, meaning e.g. "if you go out, wear your boots" and "if you don't go out, don't wear your boots". Since the consequents contradict each other, an unmodified application of Brewka's method produces $\mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(\{p_1\}, I) = \{\{p_1 \Rightarrow !p_2\}, \{\neg p_1 \Rightarrow !\neg p_2\}\}$, which fails to make true $O(p_2/p_1)$ by any truth definition (*td-pcd*1-8): the right set contains no imperatives that are in any way triggered by p_1. So we cannot derive that you ought to wear your boots, given that you are going out. But intuitively there is no conflict, since the conflicting obligations arise in mutually exclusive circumstances only.

4.2 A naïve approach

A straightforward way to adopt Brewka's method to the case of conditional imperatives is to use not all imperatives for the construction, but only those that are triggered by the facts W, i.e. to use $\mathit{Triggered}_{\mathcal{I}}(W, \Delta)$ instead of Δ:

Definition 2 (The naïve approach).
Let $\mathcal{I} = \langle I, f, g, < \rangle$ be a prioritized conditional imperative structure, Δ be a subset of I, and $W \subseteq \mathscr{L}_{PL}$ be a set of PL-sentences. Then $\Gamma \in \mathscr{P}^{\mathrm{n}}_{\mathcal{I}}(W, \Delta)$ iff $\Gamma \in \mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(W, \mathit{Triggered}_{\mathcal{I}}(W, \Delta))$.

The change resolves our earlier problems with Brewka's method: consider again the set of imperatives $\{p_1 \Rightarrow !p_2, \neg p_1 \Rightarrow !\neg p_2\}$, where the imperatives were interpreted as ordering me to wear my boots when I go out, and not wear my boots when I don't. The new definition produces $\mathscr{P}^{\mathrm{n}}_{\mathcal{I}}(\{p_1\}, I) = \{\{p_1 \Rightarrow !p_2\}\}$, its only 'preferred' subset containing just the one imperative that is triggered given the facts $\{p_1\}$. By any truth definition (*td-pcd*1-8), $O(p_2/p_1)$ is now defined true, so given that you go out, you ought to wear your boots, which is as it should be.

The naïve approach is clearly a conservative extension of Brewka's original method to conditional imperatives: for sets Δ of unconditional imperatives, $\mathit{Triggered}_{\mathcal{I}}(\{\top\}, \Delta) = \Delta$. It is similar to Horty's proposal in [21] in that conflicts are only removed between imperatives that are triggered (though the exact mechanism differs from Horty's). When I nevertheless call it 'naïve', this is because there are conceivable counterexamples to this method. Consider the set of imperatives $\{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$, ranked $!p_1 < \; p_1 \Rightarrow !p_2 < !\neg p_2$, and for an interpretation suppose that your job requires you to go outside p_1, that your mother, who is concerned for your health, told you to wear a scarf p_2 if you go outside, and that your friends don't want you to wear a scarf, whether you go outside or not. In the default situation $\top$ only the first imperative and the third are triggered, i.e. $\mathit{Triggered}_{\mathcal{I}}(\{\top\}, I) = \{!p_1, !\neg p_2\}$. Since their demands are consistent with each other, we obtain $\mathscr{P}^{\mathrm{n}}_{\mathcal{I}}(\{\top\}, I) = \{\{!p_1, !\neg p_2\}\}$, for which all truth definitions (*td-pcd*1-8) make $O(p_1 \wedge \neg p_2/\top)$ true. So you ought to go out and not wear a scarf, thus satisfying the first and the third imperative, but violating the second-ranking imperative. But arguably, if an imperative is to be violated, it should not be the second-ranking $p_1 \Rightarrow !p_2$, but the lowest ranking $!\neg p_2$ instead.

4.3 The stepwise approach

To avoid the difficulties of the 'naïve' approach, it seems we must not just take into account the imperatives that are triggered, but also those that become triggered when higher ranking imperatives are satisfied. To this effect, the following modification may seem reasonable:

Definition 3 (The stepwise approach).
Let $\mathcal{I} = \langle I, f, g, < \rangle$ be a prioritized conditional imperative structure, Δ be a subset of I, and $W \subseteq \mathscr{L}_{\mathrm{PL}}$ be a set of PL-sentences. Then $\Gamma \in \mathscr{P}^{\mathrm{s}}(W, \Delta)$ iff (i) *$W \nvdash_{PL} \bot$, and* (ii) *Γ is obtained from a full prioritization $\prec$ by defining*

$$\Gamma_{[\prec\downarrow i]} = \begin{cases} \bigcup_{j \,\prec\, i} \Gamma_{[\prec\downarrow j]} \cup \{i\} & \text{if } i \in \mathit{Satisfiable}_{\mathcal{I}}(W \cup [\bigcup_{j \,\prec\, i} \Gamma_{[\prec\downarrow j]}]^f, \Delta), \text{ and} \\ \bigcup_{j \,\prec\, i} \Gamma_{[\prec\downarrow j]} & \text{otherwise,} \end{cases}$$

for any $i \in \Delta$, and letting $\Gamma = \bigcup_{i \in \Delta} \Gamma_{[\prec\downarrow i]}$.

So at each step one considers what happens if the imperatives that were included so far are satisfied, and adds the current imperative only if it is satisfiable given the truth of W and the satisfaction of these previous imperatives. Note that satisfiability of an imperative, like its satisfaction and violation, presupposes that the imperative is triggered. In contrast to the naïve approach, the new definition not only includes, at each step, those imperatives that are triggered and can be satisfied given the facts and the supposed satisfaction of the previously added imperatives: it also includes those that *become* triggered when a previously added imperative is satisfied.

The modification avoids the previous difficulty: consider again the set of imperatives $\{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$, with the ranking $!p_1 < \ p_1 \Rightarrow !p_2 < !\neg p_2$. There is just one full prioritization, which for $W = \{\top\}$ yields in the first step the set $\{!p_1\}$, and in the second step $\{!p_1, p_1 \Rightarrow !p_2\}$, since $p_1 \Rightarrow !p_2$ is triggered when the previously added imperative $!p_1$ is assumed to be satisfied. In the third step, nothing is added: though the imperative $!\neg p_2$ is triggered, it cannot be satisfied together with the previously added imperatives. So we obtain $\mathscr{P}^{\mathrm{s}}_{\mathcal{I}}(\{\top\}, I) = \{\{!p_1, p_1 \Rightarrow !p_2\}\}$, and hence $O(p_1/\top)$, but not $O(p_1 \wedge \neg !p_2/\top)$, is defined true by all of (*td-pcd*1-8). Operators that accept 'deontic detachment' (as defined by *td-pcd*5-8) even make true $O(p_1 \wedge p_2/\top)$, and so in the given interpretation you must go out and wear a scarf, which now is as it should be.

However, a small change in the ordering shows that this definition does not suffice: let the imperatives now be ranked $p_1 \Rightarrow !p_2 < !p_1 < !\neg p_2$. (For the interpretation, assume that the conditional imperative to wear a scarf when leaving the house was uttered by some high-ranking authority, e.g. a doctor.) Then again $\mathscr{P}^{\mathrm{s}}_{\mathcal{I}}(\{\top\}, I) = \{\{!p_1, !\neg p_2\}\}$: in the first step, nothing is added since $p_1 \Rightarrow !p_2$ is neither triggered by the facts nor by the assumed satisfaction of previously added imperatives (there are none). In the next two steps, $!p_1$ and $!\neg p_2$ are added, as each is consistent with the facts and the satisfaction of the previously added imperatives. So again all of (*td-pcd*1-8) make true $O(p_1 \wedge \neg p_2/\top)$, i.e. you ought to go out and not wear a scarf, satisfying the second and third ranking imperatives at the expense of the highest ranking one. But surely, if one must violate an imperative, it should be one of the lower-ranking ones instead.

4.4 The reconsidering approach

The merits of the stepwise approach were that it did not only consider the imperatives that are triggered, but also those that *become* triggered when already added imperatives are satisfied. Such considerations applied to those imperatives that follow in the procedure. Yet the satisfaction of already added imperatives might also trigger higher-ranking imperatives, which by this method are not considered again. So it seems necessary, at each step, to reconsider also the higher-ranking imperatives. An algorithm that does that was first introduced for default theory by Marek & Truszczyński [28] p. 72, and later employed by Brewka in [5]; it can be reformulated for the present setting as follows:

Definition 4 (The reconsidering approach).
Let $\mathcal{I} = \langle I, f, g, < \rangle$ be a prioritized conditional imperative structure, Δ be a subset of I, and $W \subseteq \mathscr{L}_{\mathrm{PL}}$ be a set of PL-sentences. Then $\Gamma \in \mathscr{P}^{\mathrm{r}}_{\mathcal{I}}(W, \Delta)$ iff (i) *$W \nvdash_{\mathrm{PL}} \bot$, and* (ii) *$\Gamma$ is obtained from a full prioritization $\prec$ by defining*

$$\Gamma_i \ = \ \bigcup\nolimits_{j \prec i} \Gamma_\beta \ \cup \ min_\prec[Satisfiable_{\mathcal{I}}(W \cup [\bigcup\nolimits_{j \prec i} \Gamma_j]^f, \Delta) \setminus \bigcup\nolimits_{j \prec i} \Gamma_j]$$

for $i \in \Delta$, and letting $\Gamma = \bigcup_{i \in \Delta} \Gamma_i$.

The definition reconsiders at each step the whole ordering, and adds the $\prec$-first[14] imperative that has not been added previously and is satisfiable given both the

[14] For any ordering $<$ on some set Γ, $min_< \Gamma = \{i \in \Gamma \mid \forall i' \in \Gamma : \text{if } i' \neq i, \text{ then } i' \not< i\}$, and $max_< \Gamma = \{i \in \Gamma \mid \forall i' \in \Gamma : \text{if } i' \neq i, \text{ then } i \not< i'\}$, as usual.

circumstances C and the consequents of the previously added imperatives. Note that in 'Γ_i', i is used just as a index – it does not mean that i is considered for addition to the set at this step, and in fact it may be added at an earlier or later step (or not at all). To see how the definition works, consider again the example which the stepwise approach failed, i.e. the set of imperatives $\{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$, with the ranking $p_1 \Rightarrow !p_2 < !p_1 < !\neg p_2$. We are interested in the preferred sets for the default circumstances $\top$, i.e. the sets in $\mathscr{P}^{\mathrm{r}}_{\mathcal{I}}(\{\top\}, I)$. I is already fully prioritized, so there is just one such set. Applying the algorithm, we find the minimal (highest ranking) element in $\mathit{Satisfiable}_{\mathcal{I}}(\{\top\}, I)$ is $!p_1$, so this element is added in the first step. In the second step, we look for the minimal element in $\mathit{Satisfiable}_{\mathcal{I}}(\{\top\} \cup \{!p_1\}^f, I)$, other than the previously added $!p_1$. It is $p_1 \Rightarrow !p_2$, since the assumed satisfaction of all previously added imperatives triggers it, and its consequent can be true together with $\{\top\} \cup \{p_1\}$. So $p_1 \Rightarrow !p_2$ is added in this step. In the remaining third step, nothing is added: $!\neg p_2$ is not in $\mathit{Satisfiable}_{\mathcal{I}}(\{\top\} \cup \{!p_1, p_1 \Rightarrow !p_2\}^f, I)$, and all other elements in this set have been previously added. So $\mathscr{P}^{\mathrm{r}}_{\mathcal{I}}(\{\top\}, I) = \{\{!p_1, p_1 \Rightarrow !p_2\}\}$. Now all truth definitions (*td-pcd*1-8) make true $O(p_1/\top)$, but not $O(p_1 \wedge \neg !p_2/\top)$, and operators that accept 'deontic detachment' make true $O(p_1 \wedge p_2/\top)$. So, in the given interpretation, you must go out (as your job requires) and wear a scarf (as the doctor ordered you to do in case you go out), which is as it should be.

However, again problems remain. Reconsider the set $\{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$, but let the ranking now be $p_1 \Rightarrow !p_2 < !\neg p_2 < !p_1$. Let $p_1 \Rightarrow !p_2$ stand for the doctor's order to wear a scarf when going outside, let $!\neg p_2$ stand for your friends' expectation that you don't wear a scarf, and let $!p_1$ represent your sister's wish that you leave the house. Construct the set in $\mathscr{P}^{\mathrm{r}}_{\mathcal{I}}(\{\top\}, I)$ – since I remains fully prioritized, there is again just one such set. The minimal element in $\mathit{Satisfiable}_{\mathcal{I}}(\{\top\}, I)$ is $!\neg p_2$, and so is added in the first step. The minimal element in $\mathit{Satisfiable}_{\mathcal{I}}(\{\top\} \cup \{!\neg p_2\}^f, I)$, other than $!\neg p_2$, is $!p_1$ which therefore gets added in the second step. Nothing is added in the remaining step: $!\neg p_2$ and $!p_1$ have already been added, and $p_1 \Rightarrow !p_2$ is not in $\mathit{Satisfiable}_{\mathcal{I}}(\{\top\} \cup \{!\neg p_2, !p_1\}^f, I)$: though it is triggered by the assumed satisfaction of $!p_1$, its consequent is contradicted by the assumed satisfaction of $!\neg p_2$. So $\mathscr{P}^{\mathrm{r}}_{\mathcal{I}}(\{\top\}, I) = \{\{!p_1, !\neg p_2\}\}$. Hence all truth definitions (*td-pcd*1-8) again makes true $O(p_1 \wedge \neg p_2/\top)$, so you ought to go out without a scarf, again satisfying the second and third ranking imperatives at the expense of the first, which seems the wrong solution.

4.5 The fixpoint approach

To eliminate cases in which the 'reconsidering approach' still adds imperatives that can only be satisfied at the expense of violating a higher-ranking one, a 'fixpoint' approach may seem adequate. Such an approach was first proposed for default reasoning by Brewka & Eiter [6]. It tests each set that may be considered as preferred to see if it really includes all the elements that should be included: imperatives that are triggered given the facts and the assumed satisfaction of all imperatives in the set, and would be added by Brewka's [4] original method that adds the higher ranking imperatives first. The procedure translates as follows:

Definition 5 (The fixpoint approach).
Let $\mathcal{I} = \langle I, f, g, < \rangle$ be a prioritized conditional imperative structure, Δ be a subset of I, and $W \subseteq \mathscr{L}_{\mathrm{PL}}$ be a set of PL-sentences. Then

$$\Gamma \in \mathscr{P}^{\mathrm{f}}_{\mathcal{I}}(W, \Delta) \quad \textit{iff} \quad \Gamma \in \mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(W, \textit{Triggered}_{\mathcal{I}}(W \cup \Gamma^{f}, \Delta)).$$

To see how this definition works, consider the above set of imperatives $I = \{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$, with the ranking $p_1 \Rightarrow !p_2 < !\neg p_2 < !p_1$. It is immediate that the set $\{!p_1, !\neg p_2\}$ cannot be in $\mathscr{P}^{\mathrm{f}}_{\mathcal{I}}(\{\top\}, I)$: if we assume all imperatives in this set to be satisfied, then all imperatives are triggered, i.e. $\textit{Triggered}_{\mathcal{I}}(\{\top\} \cup \{!p_1, !\neg p_2\}^{f}, I) = I$. By Brewka's original method, $\mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(W, I) = \{\{p_1 \Rightarrow !p_2, !p_1\}\}$: $<$ is already fully prioritized, and for this full prioritization the method adds $p_1 \Rightarrow !p_2$ in the first step, $!\neg p_2$ cannot be added in the second step since its consequent contradicts the consequent of the previously added $p_1 \Rightarrow !p_2$, and in the third step $!p_1$ is added. So since the considered set $\{!p_1, !\neg p_2\}$ is not in $\mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(W, I)$, it is not a 'fixpoint'. Rather, as may be checked, the only 'fixpoint' in $\mathscr{P}^{\mathrm{f}}_{\mathcal{I}}(\{\top\}, I)$ is $\{p_1 \Rightarrow !p_2, !p_1\}$. For this set all truth definitions (*td-pcd*1-8) make true $O(p_1/\top)$, but no longer $O(p_1 \wedge \neg p_2/\top)$. Moreover, truth definitions like (*td-pcd*5-8) that allow 'deontic detachment' make true $O(p_1 \wedge p_2/\top)$. In the given interpretation this means that you must leave the house at your sisters request and wear a scarf, as the doctor ordered you to do in case you go out.

Though the construction now no longer makes true $O(p_1 \wedge \neg p_2/\top)$, its solution for the example, that determines the set $\{p_1 \Rightarrow !p_2, !p_1\}$ as the fixpoint of the set of imperatives $\{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$ with the ranking $p_1 \Rightarrow !p_2 < !\neg p_2 < !p_1$, seems questionable. Though this now includes the doctor's order, you now have no obligation anymore to satisfy the imperative that is second ranking, i.e. your friends' request that you don't wear a scarf; truth definitions (*td-pcd*4-8) even oblige you to violate it by wearing a scarf. Now consider the situation without the third ranking imperative $!p_1$: it can easily be verified that for a set $I = \{p_1 \Rightarrow !p_2, !\neg p_2\}$ the only fixpoint in $\mathscr{P}^{\mathrm{f}}_{\mathcal{I}}(\{\top\}, I)$ is $\{!\neg p_2\}$. So for the reduced set, (*td-pcd*2) makes true $O(\neg p_2/\top)$, i.e. you ought to obey your friends' wish. That the satisfaction of this higher ranking imperative $!\neg p_2$ should no longer be obligatory when a lower ranking imperative $!p_1$ is added, seems hard to explain. If the ranking is taken seriously, I think one should still satisfy the higher ranking imperatives, regardless of what lower ranking imperatives are added.

However, there is another, perhaps even more severe problem with the fixpoint approach.[15] Consider a new set of imperatives $\{p_1 \Rightarrow !p_2, !(p_1 \wedge \neg p_2), !p_3\}$, with the ranking $p_1 \Rightarrow !p_2 < !(p_1 \wedge \neg p_2) < !p_3$. For an interpretation, let the first imperative be again the doctor's order to wear a scarf in case you go out, the second one be your friends' request to go out and not wear a scarf, and the third ranking imperative be the wish of your aunt that you write her a letter. Try to find a fixpoint for the default circumstances, i.e. some $\Gamma \in \mathscr{P}^{\mathrm{f}}_{\mathcal{I}}(\{\top\}, I)$: either Γ contains the highest ranking imperative $p_1 \Rightarrow !p_2$ or it does not. If Γ contains it, then $p_1 \Rightarrow !p_2$ must be in $\textit{Triggered}_{\mathcal{I}}(\{\top\} \cup \Gamma^{f}, I)$. It can only be in there if also $!(p_1 \wedge \neg p_2)$ is in Γ, for otherwise $p_1 \Rightarrow !p_2$ could not be triggered. But no set that

[15] Both problems also arise for a new fixpoint approach by John F. Horty in [22].

is constructed by Brewka's method can include both of these imperatives, since their consequents contradict each other. So Γ does not contain $p_1 \Rightarrow !p_2$, contrary to our assumption. So assume Γ does not contain $p_1 \Rightarrow !p_2$. Whatever Γ is like, $\mathit{Triggered}_{\mathcal{I}}(\{\top\} \cup \Gamma^f, I)$ includes $!(p_1 \wedge \neg p_2)$. By Brewka's method, $!(p_1 \wedge \neg p_2)$ will only not be added to the set $\Gamma \in \mathscr{P}^{\mathrm{B}}_{\mathcal{I}}(\{\top\}, \mathit{Triggered}_{\mathcal{I}}(\{\top\} \cup \Gamma^f, I))$ if the consequents of previously added imperatives conflict with its consequent – but the only higher ranking imperative is $p_1 \Rightarrow !p_2$ and we already established that it is not in Γ. So $!(p_1 \wedge \neg p_2)$ is in Γ. But then $p_1 \Rightarrow !p_2$ is in $\mathit{Triggered}_{\mathcal{I}}(\{\top\} \cup \Gamma^f, I)$, and so is added to Γ in the first step of the construction, contrary to the assumption that it is not in Γ. So there is a *reductio ad absurdum* for both possible cases, hence there can be no $\Gamma \in \mathscr{P}^{\mathrm{f}}_{\mathcal{I}}(\{\top\}, I)$, i.e. there is no fixpoint. So there is also no fixpoint that contains $!p_3$, and hence none of the truth definitions make $O(p_3/\top)$ true, and so you do not even have to write to your aunt. But even if the presence of both a higher ranking conditional imperative and a lower ranking, unconditional imperative to violate it poses a problem (why should it? after all, the lower ranking imperative is outranked), it is hard to see why the subject should be left off the hook for all other, completely unrelated obligations.[16]

4.6 Discussion

For a discussion of our results so far, let us return to the 'drinking and driving' example from the introduction. Let the three imperatives:

(1) Your mother says: if you drink anything, then don't drive.
(2) Your best friend says: if you go to the party, then you do the driving.
(3) Some acquaintance says: if you go to the party, then have a drink with me.

be represented by a prioritized conditional imperative structure $\mathscr{I} = \langle I, f, g, < \rangle$, where $I = \{(p_1 \Rightarrow !\neg p_2, p_3 \Rightarrow !p_2, p_3 \Rightarrow !p_1\}$ and $p_1 \Rightarrow !\neg p_2 < p_3 \Rightarrow !p_2 < p_3 \Rightarrow !p_1$. Let the set of facts be $\{p_3\}$, i.e. you go to the party. As we noted, Brewka's original method is not tailored to be directly employed to conditional imperatives. The next three approaches, the naïve, the stepwise and the reconsidering ones, produce $\mathscr{P}^{\mathrm{n}}_{\mathcal{I}}(\{p_3\}, I) = \mathscr{P}^{\mathrm{s}}_{\mathcal{I}}(\{p_3\}, I) = \mathscr{P}^{\mathrm{r}}_{\mathcal{I}}(\{p_3\}, I) = \{\{p_3 \Rightarrow !p_2, p_3 \Rightarrow !p_1\}\}$, which by all truth definitions (*td-pcd*1-8) makes true $O(p_1 \wedge p_2/p_3)$, so you ought to drink and drive. The fixpoint approach produces $\mathscr{P}^{\mathrm{f}}_{\mathcal{I}}(\{p_3\}, I) = \{\{p_1 \Rightarrow !\neg p_2, p_3 \Rightarrow !p_1\}\}$, so all truth definitions make true $O(p_1/p_3)$, which means you ought to drink. Truth definition with 'deontic detachment' like (*td-pcd*5-8) additionally make true $O(p_1 \wedge \neg p_2/p_3)$, so you ought to drink and not drive. But the natural reaction is to ignore the third ranking imperative and drive, as your best friend asked you to do. So it seems we have to look for a different solution.

[16] An independent approach by Makinson in [24], which, however, only considers non-prioritized conditionals, also fails in this case: for the default circumstances $\top$ it produces the set $\{!(p_1 \wedge \neg p_2), !p_3\}$. $p_1 \Rightarrow !p_2$ is not considered, since its only 'label' (roughly: a conjunction of the circumstances, the imperatives' antecedents that would trigger* it, and its consequent) is inconsistent (it is $\top \wedge (p_1 \wedge \neg p_2) \wedge p_2$). But it is requires explanation why the agent should not be allowed to obey $p_1 \Rightarrow !p_2$, rather than having to violate it by satisfying $!(p_1 \wedge \neg p_2)$.

Before we do that, I will, however, question again our intuition in this matter. John F. Horty [22] has recently used a structurally similar example to argue for just the opposite, that the solution by the fixpoint approach is correct, i.e. that (in our example) you should drink and not drive. His example is that of three commands, uttered by a colonel, a major and a captain to a soldier, Corporal O'Reilly. The Colonel, who does not like it too warm in the cabin, orders O'Reilly to open the window whenever the heat is turned on. The Major, who is a conservationist, wants O'Reilly to keep the window closed during the winter. And the Captain, who does not like it to be cold, orders O'Reilly to turn the heat on during the winter. The intended representation is again the prioritized conditional imperative structure employed above for the 'drinking and driving' example, where p_1 now means that the heat is turned on, p_2 means that the window is closed, and p_3 means that it is winter. As we have seen, the fixpoint approach yields the preferred subset $\{p_1 \Rightarrow !\neg p_2, p_3 \Rightarrow !p_1\}$, making true $O(p_1/p_3)$ with (*td-pcd*1-3), and $O(p_1 \wedge \neg p_2/p_3)$ with (*td-pcd*4-8), so O'Reilly must turn on the heat and then open the window, and thus violate the Major's order. Horty argues as follows in support of the choice of this set:

> "O'Reilly's job is to obey the orders he has been given exactly as they have been issued. If he fails to obey an order issued by an officer without an acceptable excuse, he will be court-martialed. And, let us suppose, there is only one acceptable excuse for failing to obey such an order: that obeying the order would, in the situation, involve disobeying an order issued by an officer of equal or higher rank. (...) So given the set of commands that O'Reilly has been issued, can he, in fact, avoid court-martial? Yes he can, by (...) obeying the orders issued by the Captain and the Colonel (...). O'Reilly fails to obey the Major's order, but he has an excuse: obeying the Major's order would involve disobeying an order issued by the Colonel."

Horty's principle seems quite acceptable: for each order issued to the agent, the agent may ask herself if obeying the order would involve disobeying an order of a higher ranking officer (then he is excused), and otherwise follow it. The result is a set of imperatives where each imperative is either obeyed, or disobeyed but the disobedience excused. When I nevertheless think the argument is not correct, it is because I think it confuses the *status quo* and the *status quo posterior*. Obeying the Major's order does not, in the initial situation, involve disobeying the Colonel's order. Only once O'Reilly follows the Captain's order and turns on the heat, it is true that he must obey the Colonel, open the window and thus violate the Major's order. But this does not mean that he should follow the Captain's order in the first place – as by doing so he brings about a situation in which he is forced, by a higher ranking order, to violate a command from another higher ranking officer. Quite to the contrary, I think that being forced to violate a higher ranking order when obeying a lower ranking one is a case where following the lower one 'involves' such a violation, and so the only order the agent is excused from obeying is the lowest ranking command.

Another notion seems of importance in such examples: that of coherence, or coherent interpretation, of the imperatives that are accepted by an agent. Suppose I am a trainee at a factory, and over my new workplace there is a large sign: “If the light flashes, press the red button. By order of the Director.” On the first day, the foreman tells me “Don’t you ever press the red button.” A bit later a colleague comes round and tells me “Let’s have some fun. Make the light flash. Just short-circuiting it does the job”. Obviously I have not been told not to make the light flash. By doing so, I will have to do what the sign tells me and press the red button, and thus violate the foreman’s explicit order on my first day. But I can reason as follows: ‘Surely, the foreman did not want to contradict the Director’s order. But it would amount to a contradiction if the light flashes and I do as he told me and not press the button, though the sign says otherwise. So what the foreman meant was probably this: don’t press the button if the light does *not* flash. So I can safely make the light flash as my colleague told me, and then press the button, thus making everybody happy.’ (The consequence of such reasoning would probably be that I lose my job, which might be what my colleague meant by ‘fun’.) Such coherent reinterpretation plays an important role in judicial reasoning. But our concern are sets of imperatives that may stem from various sources and contain explicit conflicts. It is the preference ordering that is supposed to take care of arising conflicts. And by closer examination of the situation, if the light flashes, the apparent conflict is resolved since the foreman’s order is overridden. Yet that does not mean that I have to accept an obligation to bring about such a situation. If some order is to get me to make the light flash, I think it would have to rank at least as high as the foreman’s command, e.g. if my colleague had uttered the imperative in some state of emergency.

Consider finally this variant: suppose that if I am attacked by a man, I must fight him (to defend my life, my family etc.). Furthermore, suppose I have pacifist ideals which include that I must not fight the man. Now you tell me to provoke him, which in the given situation means that he will attack me. Let self-defense rank higher than my ideals, which in turn rank higher than your request. Should I do as you request? By the reasoning advocated by Horty, there is nothing wrong with it: I satisfy your request, defend myself as I must, and though I violate my ideals, I can point out to myself that the requirement to fight back took priority. But I think if I really do follow your advice, I would feel bad. I think this would not just be some irrational regret for having to violate, as I must, my ideals, but true guilt for having been tempted into doing something I should not have done, namely provoking the man: it caused the situation that made me violate my ideals. So I think our intuitions in the ‘drinking and driving’ example and the other cases have been correct.

5 New Strategies and a New Proposal

In the face of the difficulties encountered so far, it seems necessary to address the issue of finding an appropriate mechanism for a resolution of conflicts between prioritized conditional imperatives in a more systematic manner.

5.1 Deontically Tailored Preferred Subsets

In the unconditional case, the reason to move from definition (td-$m1$) to (td-$m2$) was that when there are conflicts between imperatives, the former makes true the monadic deontic formula $O\bot$, i.e. the agent ought to do the logically impossible. This result was avoided by considering only maximal sets of imperatives with demands that are collectively consistent, i.e. sets that do not make $O\bot$ true. When faced with the question what dyadic deontic formula should not be true when conflicts are resolved for arbitrary situations C, the formula $O(\neg C/C)$ appears to be the dyadic equivalent: it would be weird if a mechanism for conflict resolution results in telling the agent to do something that contradicts the assumed facts.[17] So to define the set $\mathscr{P}_{\mathcal{I}}(\{C\}, I)$ for a truth definition (td-pcd1-8), we can modify Brewka's original method in such a way that it tests, at each step, for each of the constructed subsets, if the corresponding truth-definition (td-cd1-8) does not make $O(\neg C/C)$ true for this set.[18] Formally:

Definition 6 (Deontically Tailored Preferred Subsets). *Let $\mathcal{I} = \langle I, f, g, <\rangle$ be a prioritized conditional imperative structure, and $C \in \mathscr{L}_{PL}$ describe the given situation. Let (td-pcd∗) be any of the truth definitions (td-pcd1-8). Then Γ is in the set $\mathscr{P}^*_{\mathcal{I}}(\{C\}, I)$ employed by this truth definition iff* (i) $\{C\} \nvdash_{PL} \bot$, *and* (ii) *Γ is obtained from a full prioritization $\prec$ by defining*

$$\Gamma_{[\prec\downarrow i]} = \begin{cases} \bigcup_{j\,\prec\, i} \Gamma_{[\prec\downarrow j]} \cup \{i\} & \textit{if } \langle \bigcup_{j\,\prec\, i} \Gamma_{[\prec\downarrow j]} \cup \{i\}, f, g\rangle \nvDash O(\neg C/C) \textit{ by } (td\text{-}cd*), \\ \bigcup_{j\,\prec\, i} \Gamma_{[\prec\downarrow j]} & \textit{otherwise,} \end{cases}$$

for any $i \in I$, and letting $\Gamma = \bigcup_{i\in I} \Gamma_{[\prec\downarrow i]}$.

By this construction, each of the preferred subsets contains a maximal number of the imperatives such that they do not make true $O(\neg C/C)$ for the situation C and the truth definition that is employed, and so the resulting truth definition likewise avoids this truth. Such a construction of the preferred subsets might be considered 'tailored' to the truth definition in question, and any remaining deficiencies might be seen as stemming from the employed truth definition. But this being so, the method reveals a strong bias towards truth definitions that accept 'deontic detachment', and in particular truth definitions (td-pcd4-8):

Consider the set of imperatives $I = \{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$ with the ranking $!p_1 < \; p_1 \Rightarrow !p_2 < !\neg p_2$, that was used to refute the 'naïve approach'. As can be easily checked, $\mathscr{P}^*_{\mathcal{I}}(\{\top\}, I) = \{I\}$ for all truth definitions (td-pcd1-3). So by all these truth definitions, $O(p_1 \wedge \neg p_2/\top)$ is true. So they commit us to violating the second-ranking imperative, whereas intuitively, the third-ranking imperative should be violated instead. By contrast, all truth definitions (td-pcd5-8), that employ reusable output, and of course likewise (td-pcd4) that is equivalent to (td-pcd8), handle all given examples exactly as they should be. For the set $I =$

[17] For arguments why $O(\neg C/C)$ should be chosen, i.e. for their setting, the 'output' should be consistent with the 'input', rather than the formula $O(\bot/C)$ and thus consistency of output *simpliciter*, cf. Makinson & van der Torre [26] p. 158/159.

[18] The preferred subsets are thus a choice from the 'maxfamilies' defined in [26].

$\{!p_1, p_1{\Rightarrow}!p_2, !\neg p_2\}$ they produce for both, the ranking $!p_1 < \ p_1{\Rightarrow}!p_2 < !\neg p_2$ that was used to refute the 'naïve approach', and the ranking $p_1{\Rightarrow}!p_2 < !p_1 < !\neg p_2$ that was used to refute the stepwise approach, the set $\mathscr{P}^*_{\mathcal{I}}(\{\top\}, I) = \{\{!p_1, p_1{\Rightarrow}!p_2\}\}$, $* = 4, 5, 6, 7, 8$. Thus they all make true $O(p_1 \wedge p_2/\top)$, committing us to violate the lowest-ranking imperative only, as it should be for these examples. If the imperatives' ranking is instead $p_1{\Rightarrow}!p_2 < !\neg p_2 < !p_1$, that was used to refute both the 'reconsidering' and the 'fixpoint' approaches, then $\mathscr{P}^*_{\mathcal{I}}(\{\top\}, I)$ is $\{\{p_1{\Rightarrow}!p_2, !\neg p_2\}\}$, making $O(\neg p_2/\top)$ true by all these truth definitions, which thus commit us to satisfying the second ranking imperative, and not to violating it in favor of satisfying the third ranking imperative as these approaches did. Finally the set $I = \{p_1{\Rightarrow}!p_2, !(p_1 \wedge \neg p_2), !p_3\}$ with the ranking $p_1{\Rightarrow}!p_2 < !(p_1 \wedge \neg p_2) < !p_3$, that was also mishandled by the 'fixpoint approach', produces the set $\mathscr{P}^*_{\mathcal{I}}(\{\top\}, I) = \{\{p_1{\Rightarrow}!p_2, !p_3\}\}$. So it rejects the second ranking imperative, that commits to violating the higher ranking one, and keeps both others, as it should be. The 'drinking and driving' example is also handled correctly: the set $\{p_1{\Rightarrow}!\neg p_2, p_3{\Rightarrow}!p_2, p_3{\Rightarrow}!p_1\}$, with the ranking $p_1{\Rightarrow}!\neg p_2 < \ p_3{\Rightarrow}!p_2 < \ p_3{\Rightarrow}!p_1$ produces, for the situation p_3, the set $\mathscr{P}_{\mathcal{I}}(\{p_3\}, I) = \{\{p_1{\Rightarrow}!\neg p_2, p_3{\Rightarrow}!p_2\}\}$. So the third ranking imperative, that commits the agent to drinking and thus, by observation of the highest ranking imperative, prevents the agent from driving, is disregarded. Instead, the truth definitions make true $O(p_2/p_3)$, so the agent must do the driving if she goes to the party, as her best friend asked her to.

Is this the solution to our problems, then? Some uneasiness remains as to the quick way with which definitions (td-*pcd*1-3) were discharged as insufficient. Why should it not be possible to maintain, as these definitions do, that conditional imperatives only produce an obligation if they are factually triggered, while at the same time maintaining that the above examples should not be resolved the way they are by (td-*pcd*1-3)? The purpose of a truth definition for the deontic O-operator is to find a formal notion of 'ought' that reflects ordinary reasoning, and our intuitions on that matter may differ from our ideas as to what may constitute a good choice from a possibly conflicting set of prioritized conditional imperatives. I will now make a new proposal how to construct the 'preferable subsets', that keeps the positive results without committing us to prefer any of (td-*pcd*1-8) by virtue of their handling of prioritized imperatives alone.

5.2 Preferred Maximally Obeyable Subsets

What made Brewka's approach so successful is that it maximizes the number of higher ranking imperatives in the preferred subsets of a given set of unconditional imperatives: for each 'rank', a maximal number of imperatives are added that can be without making the set's demands inconsistent in the given situation. As was shown, Brewka's approach cannot be directly applied to conditional imperatives, since it makes no sense to test the demands of imperatives for inconsistencies if these imperatives may not be triggered in the same circumstances. Just considering triggered imperatives is also not enough, as was demonstrated for the 'naïve approach'. But if the maximization method is to include imperatives that are not (yet) triggered, then we must look for something else than inconsistency of demands to take the role of a threshold criterion for the maximization process.

To do so we should ask ourselves why, for the unconditional case, the aim was to find a maximal set of imperatives with demands that are collectively consistent with the situation. I think that by doing so we intend to give the agent directives that can be safely followed. While in the unconditional case this means that the agent can satisfy all the chosen imperatives, the situation is different for conditional imperatives: here an agent can also obey imperatives without necessarily satisfying them. If you tell me to visit you in case I go to Luxembourg next month, I can safely arrange to spend all of next month at home and still do nothing wrong. If we think not so much of imperatives, but of legal regulations, then I can obviously be a law-abiding citizen by simply failing to trigger any legal norm (even though this might imply living alone on an island): whether I do that or boldly trigger some of the regulations' antecedents and then satisfy those I have triggered seems not a question of logic, but of individual choice. So I think the threshold criterion to be used should be that of obeyability: we should maximize the set of imperatives the agent can thus obey, and only stop when the addition of an imperative means that, given the facts, it or some already added imperative, i.e. one that ranks higher or at least as high, can no longer be obeyed, and so will be violated.[19]

For a given set of conditional imperatives Δ and a set of factual truths W, the subsets of imperatives that can be obeyed are described by $\mathit{Obeyable}_{\mathcal{I}}(W, \Delta)$, i.e. they are those subsets $\Gamma \subseteq \Delta$ such that $W \cup \Gamma^m \nvdash_{PL} \bot$. To maximize not by collective consistency of demands, but by collective obeyability, Brewka's original approach can therefore be changed as follows:

Definition 7 (Preferred Maximally Obeyable Subsets). *Let $\mathcal{I} = \langle I, f, g, < \rangle$ be a prioritized conditional imperative structure, Δ be a subset of I, and $W \subseteq \mathscr{L}_{PL}$ be a set of PL-sentences. Then $\Gamma \in \mathscr{P}^{\circ}_{\mathcal{I}}(W, \Delta)$ iff* (i) *$W \nvdash_{PL} \bot$, and* (ii) *Γ is obtained from a full prioritization $\prec$ by defining*

$$\Gamma_{[\prec\downarrow i]} = \begin{cases} \bigcup_{j \prec i} \Gamma_{[\prec\downarrow j]} \cup \{i\} & \textit{if } \; W \cup [\bigcup_{j \prec i} \Gamma_{[\prec\downarrow j]} \cup \{i\}]^m \nvdash_{PL} \bot, \textit{ and} \\ \bigcup_{j \prec i} \Gamma_{[\prec\downarrow j]} & \textit{otherwise,} \end{cases}$$

for any $i \in \Delta$, and letting $\Gamma = \bigcup_{i \in \Delta} \Gamma_{[\prec\downarrow i]}$.

The change from Brewka's original definition is only minute: we test not the demands of the imperatives for consistency, but their materializations. Note that this is a conservative extension of Brewka's method, since for any unconditional imperative i we have $\vdash_{PL} f(i) \leftrightarrow m(i)$. As can easily be seen, the new construction solves all of the previously considered difficulties, regardless of the chosen truth definition for the deontic O-operator:

- To refute a direct application of Brewka's original method, we used the set $I = \{p_1 \Rightarrow !p_2, \neg p_1 \Rightarrow !\neg p_2\}$ with no ranking imposed. $m(I)$ is consistent and so $\mathscr{P}^{\circ}_{\mathcal{I}}(\{p_1\}, I) = \{I\}$, making $O(p_2/p_1)$ true for all definitions (td-*pcd*1-8). So you ought to wear your boots in case you go out, as it should be.

[19] While S. O. Hansson, in [17] p. 200, also advocates a move from 'consistency' to 'obeyability', what is meant there is rather the step from (*td-m*2) to (*td-d*1).

- To refute the 'naïve approach', we used the set $I = \{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$ with the ranking $!p_1 < \ p_1 \Rightarrow !p_2 < !\neg p_2$. Since I is already fully prioritized, the construction produces just one maximally obeyable subset, which is $\{!p_1, p_1 \Rightarrow !p_2\}$, as its two imperatives get added in the first two steps, and nothing is added in the third since $m(I)$ is inconsistent. All of (td-*pcd*1-8) make true $O(p_1/\top)$, none makes true the non-intuitive formula $O(\neg p_2/\top)$, and the definitions (td-*pcd*5-8) that accept 'deontic detachment' make true $O(p_1 \wedge p_2/\top)$. So you must go out and wear a scarf, which is as it should be.
- To refute the stepwise approach we used $I = \{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$ with the ordering $p_1 \Rightarrow !p_2 < !p_1 < !\neg p_2$. Still $\mathscr{P}^{o}_{\mathcal{I}}(\top\}, I) = \{\{!p_1, p_1 \Rightarrow !p_2\}\}$, so the sentences made true by truth definitions (td-*pcd*1-8) likewise do not change, and in particular the non-intuitive formula $O(\neg p_2/\top)$ is still false, and definitions (td-*pcd*5-8) that accept 'deontic detachment' make true $O(p_1 \wedge p_2/\top)$, so again you must go out and wear a scarf, which is as it should be.
- To refute the reconsidering and the fixpoint approaches we used again the set $\{!p_1, p_1 \Rightarrow !p_2, !\neg p_2\}$, but the ranking was changed into $p_1 \Rightarrow !p_2 < !\neg p_2 < !p_1$. Now $\mathscr{P}^{o}_{\mathcal{I}}(\top\}, I) = \{\{p_1 \Rightarrow !p_2, !\neg p_2\}\}$. Truth definitions (td-*pcd*1-8) make true $O(\neg p_2/\top)$ but not $O(p_1/\top)$ so you must satisfy the second ranking imperative, but not the third ranking imperative, which is as it should be.
- Troublesome for the fixpoint approach was the set $\{p_1 \Rightarrow !p_2, !(p_1 \wedge \neg p_2), !p_3\}$, with the ranking $p_1 \Rightarrow !p_2 < !(p_1 \wedge \neg p_2) < !p_3$: no fixpoint could be made out in the set and so the approach produced no preferred subset, making everything obligatory. The preferred maximally obeyable subset is $\{p_1 \Rightarrow !p_2, !p_3\}$, eliminating the second ranking imperative that demands a violation of the first, and making $O(p_3/\top)$ true under all truth definitions (td-*pcd*1-8), which again is as it should be.
- Finally, consider the 'drinking and driving' example: the set of imperatives $\{p_1 \Rightarrow !\neg p_2, p_3 \Rightarrow !p_2, p_3 \Rightarrow !p_1\}$ produces, for the situation p_3, the set of preferred maximally obeyable subsets $\mathscr{P}^{o}_{\mathcal{I}}(\{p_3\}, I) = \{\{p_1 \Rightarrow !\neg p_2, p_3 \Rightarrow !p_2\}\}$, making true $O(p_2/p_3)$ for all truth definitions (td-*pcd*1-8), so given that I go to the party I must do the driving, which is as it should be.

As could be seen, all truth definitions now produce the 'right' results in the examples used. Moreover, since all truth definitions refer to the same preferred subsets $\mathscr{P}^{o}_{\mathcal{I}}(\{C\}, I)$, it is possible to index the O-operators according to the truth definition employed, and e.g. state truths like $O^1(A/C) \wedge O^5(B/C) \rightarrow O^7(A \wedge B/C)$, meaning that if, for any maximal set of imperatives that I can obey in the situation C, imperatives are triggered that demand A, and that if I satisfy all such triggered imperatives, I will have to do B, then obeying a maximal number of imperatives includes having to do $A \wedge B$. It may well be that natural language 'ought-statements' are ambiguous in the face of conditional demands, the discussion in sec. 3 suggested this. If maximal obeyability is accepted as the threshold criterion that limits what norms an agent can be expected to conform to in a given situation, then definition 7 leaves the philosophical logician with maximal freedom as to what deontic operator is chosen.

6 Further Research

6.1 Benchmark examples for non-prioritized imperatives

Inevitably there remains further work to do. First of all, it seems worthwhile to consider what happens if the imperatives are not prioritized, in the sense that either there is no ranking between them or that they all have the same priority. It is immediate that for such imperatives, the set of preferred subsets $\mathscr{P}^{\circ}_{\mathcal{I}}(W, \varDelta)$ for a prioritized conditional imperative structure $\mathcal{I} = \langle I, f, g, < \rangle$ and a subset of the imperatives $\varDelta$, equals $max_{\subseteq} Obeyable_{\mathcal{I}}(W, \varDelta)$, i.e. the preferred subsets are just all the maximally obeyable subsets of $\varDelta$, given the facts W. There exist a number of benchmark examples for non-prioritized conditional imperatives, given by Makinson in [24], and I list without proof the solutions we obtain for these examples for the O-operators defined here.

Source and name	**Imperatives**	**Non-truths**	**Truths**
von Wright [**?**] window closing	$r \Rightarrow !c$, $s \Rightarrow !\neg c$	$O(c \wedge \neg c / r \wedge s)$	$O(c \vee \neg c / r \wedge s)$
Chisholm [7] help and inform	$!h$, $h \Rightarrow !i$, $\neg h \Rightarrow !\neg i$	$O(h/\neg h)$, $O(i/\neg h)$	$O(h \wedge i/\top)$, $O(\neg i/\neg h)$
Forrester [9] gentle murderer	$!\neg k$, $k \Rightarrow !g$	$O(g/\top)$, $O(\neg k/k)$	$O(\neg k/\top)$, $O(g/k)$
Belzer [3] Reykjavik scenario	1. $!(\neg r \wedge \neg g)$, $r \Rightarrow !g$, $g \Rightarrow !r$ 2. $!\neg r$, $!\neg g$, $r \Rightarrow !g$, $g \Rightarrow !r$	$O(\neg g/r)$	$O(g/r)$
Prakken& Sergot [33] cigarettes from a killer	$!\neg k$, $!\neg c$, $k \Rightarrow !c$	$O(\neg k/k)$	$O(c/k)$ fails!
Prakken& Sergot [33] white fence and dog	$!\neg f$,$f \Rightarrow !(f \wedge w)$, $d \Rightarrow !(f \wedge w)$	$O(\neg f/f)$, $O(\neg f/f \wedge d)$	$O(f \wedge w/f)$, $O(f \wedge w/d \wedge f)$ $O(f \wedge w/d)$ fails!
van der Torre [42] apples and pears	1. $!(a \vee p)$, $!\neg a$ 2. $!(a \vee p)$ 3. $\neg p \Rightarrow !a$, $\neg a \Rightarrow !p$	$O(\neg a/a)$	$O(\neg a \wedge p/\top)$, $O(\neg a \wedge p/\neg a)$ $O(p/\neg a)$[I] $O(p/\neg a)$
van der Torre [42] joining paths	$!a$, $!b$	$O(a \wedge b/\neg a \vee \neg b)$	$O(a \vee b/\neg a \vee \neg b)$
Makinson [24] Möbius strip	$a \Rightarrow !b$, $b \Rightarrow !c$, $c \Rightarrow !\neg a$	$O(\neg a/a)$	$O(c/a)$ fails!
Makinson [24] exclusive options	$c \Rightarrow !(a \wedge b)$, $\neg c \Rightarrow !(a \wedge \neg b)$		$O(a/\top)$[II]

[I] O-operators that accept 'circumstantial reasoning' only, i.e.(*td-pcd*2,4,6,8).

[II] O-operators that accept 'reasoning by cases' only, i.e. (*td-pcd*3,4,7,8).

So there are three benchmark examples for which our definitions fail:

In the first one, proposed by Prakken & Sergot [33] and termed 'cigarettes from a killer', the imperative $!\neg k$ is intended to mean that you should not kill a certain man, $!\neg c$ means that you should not offer this man a cigarette, and $k \Rightarrow !c$ means that if you kill the man, you should offer him a cigarette first. Prakken & Sergot argue that the solution should make true $O(c/k)$, as this applies the imperative that is more specific for the given circumstances, but none of the operators provides this result. A similar idea underlies the second example, also proposed by Prakken & Sergot [33] and termed above 'white fence and dog'. There is a general prohibition of fences $!\neg f$ except if there already is one – in that case it should be white, i.e. $f \Rightarrow !(f \wedge w)$ – or if the owner has a dog, in which case the owner should have a white fence, i.e. $d \Rightarrow !(f \wedge w)$. Again, Prakken & Sergot intend the more specific imperative to be applied in the situation where there is a dog, and so argue that $O(f \wedge w/d)$ should hold. It is true none of the operators defined above includes a 'specificity test', and I do not think that this is a defect. The legal principle *lex specialis derogat legi generali* is not universally applicable to all sets of norms, in particular if they may stem from various sources, and even in the realm of law it competes with other principles like *lex posterior*, *lex superiori*, or standard argument forms like teleological interpretation. But if in the given case the more specific imperative should take priority, we can use a priority ordering that includes $k \Rightarrow !c < !\neg c$ in case of the first example, and $d \Rightarrow !(f \wedge w) < !\neg f$ in the case of the second. Then all operators (*td*-*pcd*1-8) make true $O(c/k)$ and $O(f \wedge w/d)$, as intended.

The third example that the truth definitions fail is Makinson's [24] 'Möbius strip': here the set of imperatives is $\{a \Rightarrow !b, b \Rightarrow !c, c \Rightarrow !\neg a\}$. Makinson argues that intuitively, $O(b \wedge c/a)$ should hold. But as is immediate, any maximally obeyable set includes just two of the given imperatives, which does not suffice for the truth of $O(b \wedge c/a)$ for any of (*td*-*pcd*1-8). The argument why $O(b \wedge c/a)$ ought to be true seems to be that since the consequent of the third imperative $c \Rightarrow !\neg a$ is false in the supposed situation a, the agent cannot do anything about it even if its antecedent becomes true, and so this imperative should not be considered.[20] But is this argument sound? Even if the consequent is inevitably false, there will be a violation only if its antecedent is (made) true. Certainly, I do not think that the agent should, in such cases, be under an obligation to make the antecedent false – this would introduce a 'deontic contraposition' that, as we saw, is not generally desirable. But that does not mean that the agent should accept *an obligation* to make the antecedent true. Consider this example: a professor tells a student that next time he sees her, he must have some written paper to present. The student's mother, who is worried about his PhD not getting finished, wants him to see his professor. The fact is: he does not, and will not, have a written paper. Should he therefore have to go and see his professor? I think that it is entirely up to the agent which of the two imperatives he is going to obey, either attributing higher weight to the explicit order of his professor, or giving priority to alleviating his mother's worries. Similarly, in the case of the Möbius strip, it may be that the agent has reasons to think that she must rather disobey one of the first two imperatives than violate the third. Then the set $\{a \Rightarrow !b, b \Rightarrow !c\}$ is not an acceptable choice in the situation a, so $O(b \wedge c/a)$ should not be true, and so not providing this truth seems not a defect.

[20] Similarly, Greenspan [12] argues that "it seems that oughts are no longer in force when it is too late to see to it that their objects are fulfilled".

6.2 Theorems

Truth definitions (td-*pcd*1-8) define when a sentence of the form $O(A/C)$ is true or false with respect to a prioritized conditional structure $\mathcal{I}$ and a situation C. So I briefly consider what sentences are theorems, i. e. hold for all such structures, given the usual truth definitions for Boolean operators. It is immediate that for all truth definitions, (DExt), (DM), (DC), (DN) and (DD-R) are theorems. (DD-R) states that there cannot be both an obligation to bring about A and one to bring about $\neg A$ unless the situation C is logically impossible, so our truth definitions succeed in eliminating conflicts. All these theorems are 'monadic' in the sense that the situation C is kept fixed; in fact, they are the C-relative equivalents of standard deontic logic *SDL*. More interesting are theorems describing the relations between obligations in different circumstances. Obviously we have

(ExtC) If $\vdash_{PL} C \leftrightarrow D$ then $O(A/C) \leftrightarrow O(A/D)$ is a theorem

for all truth definitions, i.e. for equivalent situations C, the obligations do not change. As long as truth definitions are not sensitive to conflicts, e.g. for (*td-cd*1-8), we have 'strengthening of the antecedent', i.e. for these definitions

(SA) $O(A/C) \rightarrow O(A/C \wedge D)$

holds. When only maximally obeyable subsets are considered, i.e. for truth definitions (*td-pcd*1-8), both (SA) and the weaker 'rational monotonicity' theorem

(RM) $\neg O(\neg D/C) \wedge O(A/C) \rightarrow O(A/C \wedge D)$

are refuted e.g. by a set $I = \{!(p_1 \wedge p_2), !(p_1 \wedge \neg p_2), p_2 \Rightarrow \neg p_1\}$ of equally ranking imperatives: though $O(p_1/\top)$ is true and $O(\neg p_2/\top)$ false, $O(p_1/p_2)$ is false. However, for all definitions(*td-pcd*1-8), '(conjunctive) cautious monotonicity'

(CCMon) $O(A \wedge D/C) \rightarrow O(A/C \wedge D)$

holds, which states that if you should to two things and you do one of them, you still have the other one left.[21] Moreover, truth definitions (*td-pcd*2,4,6,8) validate the 'circumstantial extensionality' rule

(CExt) If $\vdash_{PL} C \rightarrow (A \leftrightarrow B)$ then $O(A/C) \leftrightarrow O(B/C)$ is a theorem

that corresponds to 'circumstantial reasoning'. All definitions that accept 'reasoning by cases', i.e. (*td-pcd*3,4,7,8), make

(Or) $O(A/C) \wedge O(A/D) \rightarrow O(A/C \vee D)$

a theorem. Note that (CExt) and (Or) derive

(Cond) $O(A/C \wedge D) \rightarrow O(D \rightarrow A/C)$,

which in turn derives (Or) in the presence of (DC), and that by adding (CExt) and (Or) we obtain again the system *PD* (cf. sec. 3). Finally, all definitions with 'deontic detachment', i.e. (*td-pcd*5,6,7,8), make

(Cut) $O(A/C \wedge D) \wedge O(D/C) \rightarrow O(A/C)$

a theorem. (Cut) is derivable given (Cond) (use Cond on the first conjunct $O(A/C \wedge D)$ to obtain $O(D \rightarrow A/C)$, agglomerate with $O(D/C)$, and from $O(D \wedge (D \rightarrow A)/C)$ derive $O(A/C)$), which syntactically mirrors the semantic

[21] This is B. Hansson's [16] theorem (19).

equivalence of definitions (*td-pcd*4) and (*td-pcd*8). Theoremhood of all of the above theorems for semantics that employ the respective truth definitions is easily proved and left to the reader (cf. my [14] and [15] as well as Makinson & van der Torre [25] for the general outline). Makinson & van der Torre's results also suggest that these theorems axiomatically define complete systems of deontic logic with respect to semantics that employ the respective truth definitions (*td-pcd*1-8), but this remains a conjecture that further study must corroborate.[22]

6.3 Questions of representation

One might wonder if it is always adequate to represent a natural language conditional imperative 'if ... then bring about that ___' by use of a set I containing an imperative i with a $g(i)$ that formalizes '...' and a $f(i)$ that formalizes '___'. This is because there is a second possibility: represent the natural language conditional imperative by an unconditional imperative $\ulcorner !(g(i) \rightarrow f(i))\urcorner$. We saw in sec. 3 that this is not generally adequate. But that does not mean that such a representation is not *sometimes* what is required. Consider the crucial imperatives in the previous examples: perhaps what your mother meant was simply 'don't drink and drive'; perhaps what the doctor meant was 'don't go out without a scarf'; perhaps the Colonel meant to tell O'Reilly not to do both, turn the heat on and keep the window closed; perhaps the sign wanted me to see to it that the light does not flash without the button being pressed, perhaps self-defense required me to see to it that I am not attacked without fighting back. These interpretations seem not wholly unreasonable, and if they are adequate, then the best representation would be by an imperative $\ulcorner !(g(i) \rightarrow f(i))\urcorner$ instead of $\ulcorner g(i) \Rightarrow ! f(i)\urcorner$. It is easy to see that with such a representation, all of the discussed methods would have resolved these examples.

What then are the conditions that make a representation by an unconditional imperative adequate? One test may be to ask: 'Would bringing about the absence of the antecedent condition count as satisfaction of the imperative?'. Would not drinking, not going out, not turning on the heat, making the light not flash, making the man not attack, count as properly reacting to the imperatives in question? It should be if what the imperatives demand is a material conditional, since then the conditional imperatives in question are equivalent to telling the agent 'either don't drink or don't drive, its your decision', 'either don't go out, or wear a scarf', 'either don't turn on the heat, or open the window', etc. Another test would be to examine if contraposition is acceptable. Can we say that your mother wanted you not to drink if you are going to drive, that the doctor wanted you to stay inside if you are not going to wear a scarf, that the Colonel wanted O'Reilly to turn off the heat if the window is closed, that the sign wants you to make the light not flash if the button is not pressed, that self-defense requires you to make the man not attack if you are not going to fight back? If the proper representation is by imperatives that demand a material conditional, then the answers should be affirmative. I do not think these are easy questions, however, and leave them to the reader to discuss and answer at his or her own discretion.

[22] For (*td-pcd*4,8), completeness of *PD* is immediate from the results in [14], [15].

6.4 The problem of permission

The definition of the deontic notion of permission in a context of conditional norms is troublesome.[23] For monadic deontic logic it is generally accepted to define (weak) permission through the absence of an obligation to the contrary, i.e. $PA =_{df} \neg O \neg A$. This has the additional effect of making $OA \vee P\neg A$ a tautology, and so there are not 'gaps' – any state of affairs is positively or negatively regulated. For dyadic operators, the analogue would be $P(A/C) =_{df} \neg O(\neg A/C)$. But this leads here to counterintuitive results: consider the set $I = \{p_1 \Rightarrow !p_2\}$, with the intended interpretation 'if you go out, wear your boots', and truth definitions ($td\text{-}pcd$1,2,3,5,6,7), i.e. those truth definitions that do not collapse into reasoning about the imperatives' materializations. For all these we have $\mathcal{I} \not\models O(p_1 \rightarrow p_2/\top)$, and so by the above definition we have $\mathcal{I} \models P(p_1 \wedge \neg p_2/\top)$. So you are permitted to go out without your boots. There are several proposals that overcome this difficulty. Von Wright, in his re-interpretation of deontic logic as rules for rational norm-giving from [46] onwards, has denied the interdefinability of obligation and permission altogether; his theory has the result that in the absence of explicit permissive norms we only have that OA implies PA, i.e. anything permitted is also obligatory. Quite similarly, Makinson & van der Torre [27] have proposed two definitions of conditional permission that, in the absence of explicit permissive norms, either make it coincide with obligation ('forward permission'), or come quite close to it, by demanding that by forbidding the behavior for the same condition, a conflict would be created for some situation ('backward permission'). All these approaches have the strange result that the less is obligatory, the less is allowed.[24] But surely one can, in some weak sense, say that given the presence of some (conditional) imperatives, an agent is still free to do A in a situation C, without saying that A is also obligatory in this situation. It is perhaps a better solution to define

$$\mathcal{I} \models P(A/C) \ \text{ iff } \ \exists \Gamma \in \mathscr{P}_{\mathcal{I}}(\{C\}, I) : \Gamma^m \cup \{C\} \nvdash_{PL} \neg A,$$

thus defining A as permissible in a situation C if there is a preferred maximally obeyable subset of the imperatives for which bringing about A does not cause a violation. For operators other than ($td\text{-}pcd$4,8), this definition is not 'gapless'. E.g. consider the set $I = \{!p_1, p_1 \Rightarrow !p_2\}$. For truth definitions that do not accept 'deontic detachment', i.e. ($td\text{-}pcd$1,2,3), we have neither $O(p_2/\top)$ nor $P(\neg p_2/\top)$: though we are not yet under an obligation to bring about p_2, we are also not permitted to bring about $\neg p_2$ and thus make satisfaction of the imperative impossible that ought to be triggered. Or consider conditional imperatives whose

[23] I do not consider here the problem of how permissive *norms*, or licenses, may be represented. For attempts to use a separate set of 'P-norms' alongside what is here the set of imperatives cf. Alchourrón & Bulygin [1], von Wright [46], Makinson [24] and Makinson & van der Torre [27].

[24] Consider the set $I = \{p_1 \Rightarrow !p_i \mid i > 1\}$, and for an interpretation suppose that I have no obligations in the rest of the world, but am a slave once I go to Australia. By 'forward' or 'backward' permission, $P(A/\neg p_1)$ is false for any A, i.e. I am not allowed to do anything if I do not go to Australia, and though $P(p_i/\top)$ holds for backward permission, it is only by virtue of p_i being obligatory down under.

consequent has become impossible to satisfy: for a set $I = \{p_1 \Rightarrow !p_2\}$ we do not have $O(\neg p_1/\neg p_2)$ for truth definitions other than (*td-pcd*4,8) since $p_1 \Rightarrow !p_2$ is not triggered in the situation $\neg p_2$, but it is also not permitted to trigger it, i.e. $P(p_1/\neg p_2)$ is not true. This deontic vagueness may indeed be adequate for such situations. Further study must determine if such a definition does not create counterintuitive results. But it is important to see that as far as reasoning about conditional norms is concerned, the old definitions of permission as the absence of prohibition, obligation as the absence of a permission to the contrary, and prohibition as the absence of permission, do no longer hold.

7 Conclusion

Reasoning about obligations when faced with different and possibly conditional imperatives is a part of everyday life. To avoid conflicts, these imperatives may be ordered by priority and then observed according to their respective ranks. The 'drinking and driving' case in the introduction presented an example of such natural reasoning. To provide a formal account is, however, additionally complicated by the fact that there are various and mutually exclusive intuitions about what belongs to the right definition of an 'obligation in the face of conditional imperatives', i.e. the definition of a deontic O-operator. Based on similar definitions of operators by Makinson & van der Torre [25], [26] for their 'input/output logic', but leaving the choice of the 'right' operator to the reader, I presented several proposals in sec. 3 for definitions of a dyadic O-operator, namely (*td-pcd*1-8). These were dependent on a choice of 'preferred subsets' among a given set of prioritized conditional imperatives. A particularly successful method to identify such subsets, but applying to unconditional imperatives only, was Brewka's [4] definition of 'preferred subtheories' within a theory. In sec. 4 I discussed various approaches that extend this method to conditional imperatives, but these failed to produce satisfactory results for a number of given examples. In sec. 5 I first examined an approach that 'tailors' the choice procedure to the truth definition for the deontic O-operator in question, where the only criterion is to avoid the truth of $O(\neg C/C)$ for possible circumstances C. Though this finally produced the intended results, it did so for truth definitions (*td-pcd*4-8) only, whereas counterexamples remained for any of the weaker truth definitions (*td-pcd*1-3). I then argued that the solution is to adapt Brewka's method in such a way that it constructs, instead of maximal subsets of imperatives that are collectively satisfiable by an agent, maximally *obeyable* subsets of the imperatives. I showed that this new proposal provides adequate solutions to all of the examples, and in particular the 'drinking and driving' example is resolved in a satisfactory fashion for all of the discussed deontic operators. In sec. 6 I demonstrated that the new proposal also includes satisfactory results for benchmark examples developed for non-prioritized conditional imperatives; I presented theorems of a deontic logic based on this proposal (though the question of their completeness had to be left open), and finally I showed that there are problems for the representation of conditional imperatives and difficulties for the definition of a deontic P-operator that further philosophical discussion and research must address.

References

1. Alchourrón, C. E. and Bulygin, E., "The Expressive Conception of Norms", in [19], 95–124.
2. Alchourrón, C. E. and Makinson, D., "Hierarchies of Regulations and Their Logic", in [19] 125–148.
3. Belzer, M., "Legal Reasoning in 3-D", *Proceedings of the First International Conference in Artificial Intelligence and Law*, Boston: ACM Press, 1987, 155–163.
4. Brewka, G., "Preferred Subtheories: An Extended Logical Framework for Default Reasoning", in: Sridharan, N. S. (ed.), *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence IJCAI-89, Detroit, Michigan, USA, August 20-25, 1989*, San Mateo, Calif.: Kaufmann, 1989, 1043–1048.
5. Brewka, G., "Reasoning about Priorities in Default Logic", in: Hayes-Roth, B. and Korf, R. E. (eds.), *Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, July 31st - August 4th, 1994*, vol. 2, Menlo Park: AAAI Press, 1994, 940–945.
6. Brewka, G. and Eiter, T., "Preferred Answer Sets for Extended Logic Programs", *Artificial Intelligence*, **109**, 1999, 297–356.
7. Chisholm, R. M., "Contrary-to-Duty Imperatives and Deontic Logic", *Analysis*, **24**, 1963, 33–36.
8. Downing, P., "Opposite Conditionals and Deontic Logic", *Mind*, **63**, 1959, 491–502.
9. Forrester, J. W., "Gentle Murder, or the Adverbial Samaritan", *Journal of Philosophy*, **81**, 1984, 193–197.
10. van Fraassen, B., "Values and the Heart's Command", *Journal of Philosophy*, **70**, 1973, 5–19.
11. Goble, L., "A Logic for Deontic Dilemmas", *Journal of Applied Logic*, **3**, 2005, 461–483.
12. Greenspan, P., "Conditional Oughts and Hypothetical Imperatives", *Journal of Philosophy*, **72**, 1975, 259–276.
13. Hansen, J., "Problems and Results for Logics about Imperatives", *Journal of Applied Logic*, **2**, 2004, 39–61.
14. Hansen, J., "Conflicting Imperatives and Dyadic Deontic Logic", *Journal of Applied Logic*, **3**, 2005, 484–511.
15. Hansen, J., "Deontic Logics for Prioritized Imperatives", *Artificial Intelligence and Law*, 2005, *forthcoming*.
16. Hansson, B., "An Analysis of Some Deontic Logics", *Nôus*, **3**, 1969, 373–398. Reprinted in [18], 121–147.
17. Hansson, S. O., *The Structure of Values and Norms*, Cambridge: Cambridge University Press, 2001.
18. Hilpinen, R. (ed.), *Deontic Logic: Introductory and Systematic Readings*, Dordrecht: Reidel, 1971.
19. Hilpinen, R. (ed.), *New Studies in Deontic Logic*,Dordrecht: Reidel, 1981.
20. Hofstadter, A. and McKinsey, J. C. C., "On the Logic of Imperatives", *Philosophy of Science*, **6**, 1938, 446–457.
21. Horty, J. F., "Reasoning with Moral Conflicts", *Noûs*, **37**, 2003, 557–605.
22. Horty, J. F., "Defaults with Priorities", 2006. Draft version of August 18, 2006, `http://www.umiacs.umd.edu/~horty/articles/2005-dp.pdf`.
23. Kraus, S., Lehmann, D. and Magidor, M., "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics", *Artificial Intelligence*, **44**, 1990, 167–207.

24. Makinson, D., "On a Fundamental Problem of Deontic Logic",in: McNamara, P. and Prakken, H. (eds.), *Norms, Logics and Information Systems*, Amsterdam: IOS, 1999, 29–53.
25. Makinson, D. and van der Torre, L., "Input/Output Logics", *Journal of Philosophical Logic*, **29**, 2000, 383–408.
26. Makinson, D. and van der Torre, L., "Constraints for Input/Output Logics", *Journal of Philosophical Logic*, **30**, 2001, 155–185.
27. Makinson, D. and van der Torre, L., "Permissions from an Input/Output Perspective", *Journal of Philosophical Logic*, **32**, 2003, 391–416.
28. Marek, V. W. and Truszczyński, M., *Nonmonotonic Logic. Context-Dependent Reasoning*, Berlin: Springer, 1993.
29. Nebel, B., "Belief Revision and Default Reasoning: Syntax-Based Approaches", in: Allen, J. A. and Fikes, R. and Sandewall, E. (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference, KR '91, Cambridge, MA, April 1991*, San Mateo: Morgan Kaufmann, 1991, 417–428.
30. Nebel, B., "Syntax-Based Approaches to Belief Revision", in: Gärdenfors, P. (ed.), *Belief Revision*, Cambridge: Cambridge University Press, 1992, 52-88.
31. Prakken, H., *Logical Tools for Modelling Legal Argument*, Dordrecht: Kluwer, 1997.
32. Prakken, H. and Sartor, G., "Argument-based Logic Programming with Defeasible Priorities", *Journal of Applied Non-classical Logics*, **7**, 1997, 25–75.
33. Prakken, H. and Sergot, M., "Contrary-to-duty obligations", *Studia Logica*, **52**, 1996, 91–115.
34. Rescher, N., *Hypothetical Reasoning*, Amsterdam: North-Holland, 1964.
35. Rescher, N., *The Logic of Commands*, London: Routledge & Kegan Paul, 1966.
36. Rintanen, J., "Prioritized Autoepistemic Logic", in: MacNish, C. and Pearce, D. and Pereira, L. M., *Logics in Artificial Intelligence, European Workshop, JELIA '94, York, September 1994, Proceedings*, Berlin: Springer, 1994, 232–246.
37. Ross, W. D., *The Right and the Good*, Oxford: Clarendon Press, 1930.
38. Ryan, M., "Representing Defaults as Sentences with Reduced Priority", in: Nebel, B. and Rich, C. and Swartout, W. (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference, KR '92, Cambridge, MA, October 1992*, San Mateo: Morgan Kaufmann, 1992, 649–660.
39. Sakama, C. and Inoue, K., "Representing Priorities in Logic Programs", in: Maher, M. (ed.), *Joint International Conference and Syposium on Logic Programming JICSLP 1996, Bonn, September 1996*, Cambridge: MIT Press, 1996, 82–96.
40. Şosa, E., "The Logic of Imperatives", *Theoria*, **32**, 1966, 224–235.
41. Świrydowicz, K., "Normative Consequence Relation and Consequence Operations on the Language of Dyadic Deontic Logic", *Theoria*, **60**, 1994, 27–47.
42. van der Torre, L., *Reasoning About Obligations*, Amsterdam: Thesis Publ., 1997.
43. Tröndle, H., "Die Wahlfeststellung", in: *Strafgesetzbuch. Leipziger Kommentar*, vol. 1, 10th ed., Berlin: Walter de Gruyter, 1985, §1, margin nos. 59–63.
44. von Wright, G. H., "A New System of Deontic Logic", *Danish Yearbook of Philosophy*, **1**, 1961, 173–182. Reprinted in [18], 105–115.
45. von Wright, G. H., *An Essay in Deontic Logic and the General Theory of Action*, Amsterdam: North Holland, 1968.
46. von Wright, G. H.: "Norms, Truth and Logic", in: von Wright, G. H., *Practical Reason: Philosophical Papers vol. I*, Oxford: Blackwell, 1983, 130–209.
47. von Wright, G. H.: "Bedingungsnormen, ein Prüfstein für die Normenlogik", in: Krawietz, W., Schelsky, H., Weinberger, O. and Winkler, G. (eds.), *Theorie der Normen*, Berlin: Duncker & Humblot, 1984, 447–456.

Introduction to Normative Multiagent Systems

Guido Boella
Dipartimento di Informatica
Università di Torino
Italy
guido@di.unito.it

Leendert van der Torre
Department of Computer Science
University of Luxembourg
Luxembourg
leendert@vandertorre.com

Harko Verhagen
Dept. of Computer and Systems Sciences
Stockholm University / KTH,
Forum 100, SE-16440 Kista, Sweden
verhagen@dsv.su.se

May 29, 2006

Normative multiagent systems as a research area can be defined as the intersection of normative systems and multiagent systems. Since the use of norms is a key element of human social intelligence, norms may be essential too for artificial agents that collaborate with humans, or that are to display behavior comparable to human intelligent behavior. By integrating norms and individual intelligence normative multiagent systems provide a promising model for human and artificial agent cooperation and co-ordination, group decision making, multiagent organizations, regulated societies, electronic institutions, secure multiagent systems, and so on.

With 'normative' we mean 'conforming to or based on norms', as in *normative behavior* or *normative judgments*. According to the Merriam-Webster Online (2005) Dictionary, other meanings of normative not considered here are 'of, relating to, or determining norms or standards', as in *normative tests*, or 'prescribing norms', as in *normative rules of ethics* or *normative grammar*. With 'norm' we mean 'a principle of right action binding upon the members of a group and serving to guide, control, or regulate proper and acceptable behavior'. Other meanings of 'norm' given by the Merriam-Webster Online Dictionary but not considered here are 'an authoritative standard or model', 'an average like a standard, typical pattern, widespread practice or rule in a group', and various definitions used in mathematics.

Normative multiagent systems are an example of the use of sociological theories in multiagent systems, and more generally of the relation between agent theory and the social sciences such as sociology, philosophy, economics, and legal science. The need for social science theories and concepts like norms in multiagent systems is now well established. For example, Wooldridge's weak notion of agency is based on flexible autonomous action (Wooldridge, 2002), and social ability as the interaction with other agents and co-operation is one of the three meanings of flexibility; the other two are reactivity as interaction with the environment, and pro-activeness as taking the initiative. In this definition autonomy refers to non-social aspects, such as operating without the direct intervention of humans or others, and have some kind of control over their actions and internal state. For some other arguments for the need for social theory in multiagent systems, see, for example, (Bond and Gasser, 1988; Conte and Gilbert, 1995; Verhagen and Smit, 1996). For a more complete discussion on the need of social theory in general, and norms in particular, see the AgentLink roadmap (roa, 2005).

Social concepts like norms are important for multiagent systems, because multiagent system research and sociology share the interest in the relation between micro-level agent behaviour and macro-level system effects. In sociology this is the (in)famous micro-macro link (Alexander et al., 1987) that focuses on the relation between individual agent behaviour and characteristics at the level of the social system. In multiagent system research, this boils down to the question "How to ensure efficiency at the level of the multiagent system whilst respecting individual autonomy?". According to Verhagen (2000) three possible solutions to this problem comprise of the use of central control which gravely jeopardizes the agent's autonomy, internalized control like the use of social laws (Shoham and Tennenholtz, 1992), and structural coordination (Ossowski, 1999) including learning norms.

Before we discuss normative multiagent systems, we consider some discussions on norms in the social sciences.

1 Norms and normative systems

In the 1960's, the sociologist Gibbs (1965) wrote an influential article on the problems concerning the definition and classification of norms, and observes that the various types of norms involve "a collective evaluation of behavior in terms of what it *ought* to be; a collective expectation as to what behavior *will be*; and/or particular *reactions* to behavior, including attempts to apply sanctions or otherwise induce a particular kind of conduct." (Gibbs, 1965, p. 589, original emphasis)

More recently, Therborn (2002) presented an overview of the role of norms for social theory and analysis. Normative action is based upon wanting to do the right thing rather than the thing that leads to ends or goals, which he calls teleological action, or the thing that leads to, expresses, or is caused by an emotion, called emotional action.

Therborn distinguishes among three kinds of norms. *Constitutive norms* define a system of action and an agent's membership in it, *regulative norms* describe the expected contributions to the social system, and *distributive norms* defining how rewards, costs, and risks are allocated within a social system. Furthermore, he distinguishes between non-institutionalized normative order, made up by personal and moral norms in day-to-day social traffic, and institutions, an example of a social system defined as a closed system of norms. Institutional normative action is equaled with role plays, i.e., roles find their expressions in expectations, obligations, and rights vis-a-vis the role holder's behaviour.

Therborn also addresses the dynamics and changing of norms. The dynamics of norms at the level of the individual agent is how norms are learned or propagated in a population. Socialization is based on identification, perceiving the compliance with the norms by other agents, or the entering of an institution. Norms are (re)enforced by the presence of incentives or sanctions. Changes in either of these three three socialization mechanisms lead to changes in the set of norms of the individual agent. These changes may be inhibited either by changes in the social system or changed circumstances, or by changes in the interpretation of the norms by the agents within the system.

Within philosophy normative systems have traditionally been studied by moral and legal philosophers. Alchourròn and Bulygin (1971) argue that a normative system should not be defined as a set of norms, as is commonly done, but in terms of consequences:

> "When a deductive correlation is such that the first sentence of the ordered pair is a case and the second is a solution, it will be called normative. If among the deductive correlations of the set α there is at least one normative correlation, we shall say that the set α has normative consequences. A system of sentences which has some normative consequences will be called a normative system." (Alchourròn and Bulygin, 1971, p.55).

In computer science, Meyer and Wieringa define normative systems as "systems in the behavior of which norms play a role and which need normative concepts in order to be described or specified" (Meyer and Wieringa, 1993, preface). They also explain why normative systems are intimately related with deontic logic.

> "Until recently in specifications of systems in computational environments the distinction between normative behavior (as it *should be*) and actual behavior (as it *is*) has been disregarded: mostly it is not possible to specify that some system behavior is non-normative (illegal) but nevertheless possible. Often illegal behavior is just ruled out by specification, although it is very important to be able to specify what should happen if such illegal but possible behaviors occurs! Deontic logic provides a means to do just this by using special modal operators that indicate the status of behavior: that is whether it is legal (normative) or not" (Meyer and Wieringa, 1993, preface).

2 Normative multiagent systems

The agents in the environment of a normative system interact with the normative system in various ways. First, from the perspective of the agents, agents can create new norms, update or maintain norms, and enforce norms, using roles defined in the normative system such as legislators or policemen. Secondly, from the perspective of social order, we can also look at the interaction between the normative system and its environment from the viewpoint of the normative system. In this viewpoint, the normative system uses the agents playing a role in it – the legislators, policemen and the like – to maintain an equilibrium in the normative multiagent system. In this perspective, we can distinguish at least two levels of equilibrium. First, norms are used to maintain social order in a normative multiagent system. Second, normative system contain a mechanism for updating themselves, to adapt to changing circumstances in its environment.

Jones and Carmo (2001) define a normative system as "Sets of agents whose interactions are norm-governed; the norms prescribe how the agents ideally should and should not behave. [...] Importantly, the norms allow for the possibility that actual behavior may at times deviate from the ideal, i.e., that violations of obligations, or of agents' rights, may occur." In our opinion, this is too general, as a normative system does not contain the agents themselves. It also is not a satisfactory definition of normative multiagent system, because it precludes the agents' control over the set of norms. We therefore use the following definition in this paper.

> A normative multiagent system is a multiagent system together with normative systems in which agents on the one hand can decide whether to follow the explicitly represented norms, and on the other the normative systems specify how and in which extent the agents can modify the norms.

Note that this definition makes no presumptions about the internal workings of an agent nor of the way norms find their expression in agent's behaviour.

Since norms are explicitly represented, according to our definition of a normative multiagent system, the question should be raised how norms are represented. Norms can be interpreted as a special kind of constraint, and represented depending on the domain in which they occur. However, the representation of norms by domain dependent constraints runs into the question what happens when norms are violated. Not all agents behave according to the norm, and the system has to deal with it. In other words, norms are not hard constraints, but soft constraints. For example, the system may sanction violations or reward good behavior. Thus, the normative system has to monitor the behavior of agents and enforce the sanctions. Also, when norms are represented as domain dependent constraints, the question will be raised how to represent permissive norms, and how they relate to obligations. Whereas obligations and prohibitions can be represented as constraints, this does not seem to hold for permissions. For example, how to represent the permission to access a resource under an access control system? Finally, when norms are represented as domain dependent constraints, the question can be raised how norms evolve.

We therefore believe that norms should be represented as a domain independent theory, for example in deontic logic (von Wright, 1951; van der Torre and Tan, 1999; van der Torre, 2003; Makinson and van der Torre, 2000; Makinson and van der Torre, 2001; Makinson and van der Torre, 2003). Deontic logic studies logical relations among obligations and permissions, and more in particular violations and contrary-to-duty obligations, permissions and their relation to obligations, and the dynamics of obligations over time. Therefore, insights from deontic logic can be used to represent and reason with norms. Deontic logic also offers representations of norms as rules or conditionals. However, there are several aspects of norms which are not covered by constraints nor by deontic logic, such as the relation between the cognitive abilities of agents and the global properties of norms.

Conte, Falconi and Sartor (1999) say that normative multiagent systems research focuses on two different sets of problems. On the one hand, they claim that legal theory and deontic logic supply a theory for of norm-governed interaction of autonomous agents while at the same time lacking a model that integrates the different social and normative concepts of this theory. On the other hand, they claim that three other problems are of interest in multiagents systems research on norms: how agents can acquire norms, how agents can violate norms, and how an agent can be autonomous. For artificial agents, norms can be designed as in legal human systems, forced upon, for example when joining an institution, or they can emerge from the agents making them norm autonomous (Verhagen, 2000). Agent decision making in normative systems and the relation between desires and obligations has been studied in agent architectures (Broersen et al., 2002), which thus explain how norms and obligations influence agent behavior.

An important question is where norms come from. Norms are not necessarily created by a single legislator, they can also emerge spontaneously, or be negotiated among the agents.

In electronic commerce research, for example, cognitive foundations of social norms and contracts are studied (Boella and van der Torre, 2006a). Protocols and social mechanisms are now being developed to support such creations of norms in multiagent systems. When norms are created, the question how they are enforced can be raised. For example, when a contract is violated, the violator may have to pay a penalty. But then there has to be a monitoring and sanctioning system, for example police agents in an electronic institution. Such protocols or roles in a multiagent system are part of the construction of social reality, and Searle (1995) has argued that such social realities are constructed by constitutive norms. This again raises the question how to represent such constitutive or counts-as norms, and how they are related to regulative norms like obligations and permissions (Boella and van der Torre, 2006a).

Not only the relation between norms and agents must be studied, but also the relation between norms and other social and legal concepts. How do norms structure organizations? How do norms coordinate groups and societies? How about the contract frames in which contracts live? How about the legal contexts in which contract frames live? How about the relation between legal courts? Though in some normative multiagent systems there is only a single normative system, there can also be several of them, raising the question how normative systems interact. For example, in a virtual community of resource providers each provider may have its own normative system, which raises the question how one system can authorize access in another system, or how global policies can be defined to regulate these local policies (Boella and van der Torre, 2006b).

Summarizing, normative multiagent systems study general and domain independent properties of norms. It builds on results obtained in deontic logic, the logic of obligations and permissions, for the representation of norms as rules, the application of such rules, contrary-to-duty reasoning and the relation to permissions. However, it goes beyond logical relations among obligations and permissions by explaining the relation among social norms and obligations, relating regulative norms to constitutive norms, explaining the evolution of normative systems, and much more.

The papers in this double special issue on normative multiagent systems address some of these issues, but they also address new research issues that are of central importance for the whole field of normative multiagent systems. These include how to combine theories of teleological action (e.g., the BDI model of agency) with models of normative action, how to model the dynamics of norms when institutions' norm sets are to be combined, the development and testing of logics of normative reasoning and dynamics, and the formalization descriptive social theories of normative action into implementable formal models.

3 NorMAS 2005

NorMAS05 was an international symposium on normative multiagent systems, organized in April 2005 by the authors of this article as part of the 2005 AISB convention (AISB standing for the Society for the Study of Artificial Intelligence and the Simulation of Behaviour). The symposium attracted papers from a variety of areas, such as the social sciences (and computational sociology in particular), computer science, and formal logics. A number of these papers representing these areas were selected for this double special issue on normative multiagent systems. Four general themes are addressed in these papers, namely intra-agent aspects of norms, interagent aspects of norms, normative systems and their borders, and combining normative systems.

3.1 Intra-agent aspects of norms

The paper "My Agents Love to Conform: Norms and Emotion in the Micro-Macro Link" by von Scheve et al. investigates the function of emotion in relation to norms in natural and artificial societies. It shows that unintentional behavior can be normative and socially functional at the same time, thereby highlighting the role of emotion. By defining norms as mental objects, the role of emotion in maintaining and enforcing norms is studied, relates these findings social structural dynamics in natural and societies,and outlines the possibilities of an application to a multi-agent architecture.

Sadri, Stati, and Toni's "Normative KGP Agents" extends the logical model of agency known as the KGP model to support agents with normative concepts, based on the roles an agent plays and the obligations and prohibitions that result from playing these roles. The proposed framework illustrates how the resulting normative concepts, including the roles, can evolve dynamically during the lifetime of the agent. It also illustrates how these concepts can be combined with the existing capabilities of KGP agents in order to plan for their goals, react to changes in the environment, and interact with other agents. Finally, the paper gives an executable specification of normative concepts that can be used directly for prototyping applications.

3.2 Interagent aspects of norms

Kibble's paper "Speech acts, commitment and multiagent communication" aims to reconsider the suitability of speech act theory as a basis for agent communication languages. It models dialogue states as deontic scoreboards which keep track of commitments and entitlements that speakers acknowledge and hearers attribute to other interlocutors and outlines an update semantics and protocol for selected locutions.

Sauro's paper "Qualitative Criteria of Admissibility for Enforced Agreements" focuses on the desirablility of artificial agents to help each other when they cannot achieve their goals, or when they profit from social exchanges. It studies the coalition formation processes supported by enforced agreements and defines two qualitative criteria that establish when a coalition is admissible to be formed. These two properties can be used when the space of possible coalitions is unknown.

3.3 Normative systems and their borders

Davidsson and Johansson classify artificial societies and identify four different types of stakeholders in their paper “On the Potential of Norm-Governed Behavior in Different Categories of Artificial Societies”. The potential of norm-governed behavior in different types of artificial societies is investigated based on the preferences of the stakeholders and how they influence the state of the society. The paper concludes that the more open a society is the more it has to rely on agent owners and designers to achieve norm-governed behavior, whereas in more closed societies the environment designers and owners may control the degree of norm-governed behavior.

Hahn, Fley, and Florian argue in “A Framework for the Design of Self-Regulation of Open Agent-based Electronic Marketplaces” that allowing self-interested agents to activate social institutions during run-time can improve the robustness of open multiagent systems. Based on sociological theory, institutions are seen as rules which have to be activated and adopted by the agent population. A framework for self-regulation of multiagent system for the domain of electronic marketplaces is developed, consisting of three different institutional forms that are defined by the mechanisms and instances that generate, change, or safeguard them. The paper shows that allowing autonomous agents both the reasoning about their compliance with a rule and the selection of the form of an institution helps to balance the trade-off between the autonomy of self-interested agents and the maintenance of social order in an open multiagent system and to ensure almost the same qualities as in closed environments.

In “Mapping Deontic Operators to Abductive Expectations”, Alberti et al. propose a mapping of deontic operators (obligations, prohibition, permission) to language entities (expectations) available within the an agent framework developed for agent interaction in open agent societies. The mapping is supported by showing a similarity between the abductive semantics for expectations and the Kripke semantics that can be given to deontic operators.

In “A Normative Framework for Agent-Based Systems”, López y López, Luck, and d’Inverno present a formal normative framework for agent-based systems that adresses two omissions of previous research on the use of norms in computational models of open societies to help to cope with the heterogeneity, the autonomy and the diversity of interests among their members. These are the lack of a canonical model of norms that facilitates their implementation and enables the description of the processes of reasoning about norms, and secondly the perspective of individual agents and what they might need to effectively reason about the society in which they participate.

3.4 Combining normative systems

Grossi et al. introduce the notion of contextual ontologies in their paper “Ontological Aspects of the Implementation of Norms in Agent-Based Electronic Institutions” and also provide a formal machinery to characterise this notion. This notion solves the problem of different institutions implementing the same set of norms in different ways presupposing divergent ontologies of the concepts in which that set of norms is formulated.

References

2005. *Agent Technology Roadmap: A Roadmap for Agent-Based Computing.*

Alchourròn, C.E. and E. Bulygin. 1971. *Normative Systems.* Springer.

Alexander, J.C., B. Giesen, R. Münch, and N.J. Smelser, editors. 1987. *The Micro-Macro Link.* University of California Press.

Boella, G. and L. van der Torre. 2006a. A game theoretic approach to contracts in multiagent systems. *IEEE Trans. SMC, Part C.*

Boella, G. and L. van der Torre. 2006b. Security policies for sharing knowledge in virtual communities. *IEEE Trans. SMC, Part A*.

Bond, A. H. and L. Gasser. 1988. An Analysis of Problems and Research in DAI. In A. H. Bond and L. Gasser, editors, *Readings in Distributed Artificial Intelligence*, pages 3–35. Morgan Kaufmann.

Broersen, J., M. Dastani, J. Hulstijn, and L. van der Torre. 2002. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447.

Conte, R., R. Falcone, and G. Sartor. 1999. Introduction: Agents and Norms: How to Fill the Gap? *Artificial Intelligence and Law*, pages 1 – 15.

Conte, R. and N. Gilbert, 1995. "Computer Simulation for Social Theory". In R. Conte and N. Gilbert, editors, *Artificial Societies: The Computer Simulation of Social Life*, chapter Computer Simulation for Social Theory, pages 1 – 18. UCL Press.

Gibbs, J. P. 1965. Norms: The Problem of Definition and Classification. *The American Journal of Sociology*, 70(5):586 – 594.

Jones, A. and J. Carmo. 2001. Deontic Logic and Contrary-to-Duties. In D. Gabbay, editor, *Handbook of Philosophical Logic.* Kluwer, page 203279.

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

Makinson, D. and L. van der Torre. 2001. Constraints for input-output logics. *Journal of Philosophical Logic*, 30(2):155–185.

Makinson, D. and L. van der Torre. 2003. Permissions from an input/output perspective. *Journal of Philosophical Logic*, 32(4):391–416.

Merriam-Webster OnLine. 2005. The Language Center. `www.m-w.com/`.

Meyer, J-J. and R. Wieringa, editors. 1993. *Deontic Logic in Computer Science: Normative System Specification.* Wiley.

Ossowski, S. 1999. *Co-ordination in Artificial Agent Societies*. Springer.

Searle, J. R. 1995. *The Construction of Social Reality*. The Free Press.

Shoham, Y. and M. Tennenholtz. 1992. On the Synthesis of Useful Social Laws for Artificial Agent Societies (Preliminary Report). In *Proceedings of the National Conference on Artificial Intelligence*, pages 276–281, San Jose, CA.

Therborn, G. 2002. Back to Norms! On the Scope and Dynamics of Norms and Normative Action. *Current Sociology*, 50(6):863 – 880.

van der Torre, L. 2003. Contextual deontic logic: Normative agents, violations and independence. *Annals of Mathematics and Artificial Intelligence*, 37(1-2):33–63.

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

Verhagen, H. 2000. *Norm Autonomous Agents*. Ph.D. thesis, Department of System and Computer Sciences, The Royal Institute of Technology and Stockholm University, Sweden.

Verhagen, H. and R. Smit. 1996. Modelling Social Agents in a Multiagent World. In W. Van de Velde and J. W. Perram, editors, *, Position Papers MAAMAW 1996, Technical Report 96-1*. Vrije Universiteit Brussel - Artificial Intelligence Laboratory.

von Wright, G.H. 1951. Deontic logic. 60:1–15.

Wooldridge, M. 2002. *An Introduction to MultiAgent Systems*. Wiley.

Bibliography

Guido Boella

Guido Boella received the PhD degree at the University of Torino in 2000. He is currently professor at the Department of Computer Science of the University of Torino. His research interests include multi-agent systems, in particular, normative systems, institutions and roles using qualitative decision theory. He is the co-chair of the first workshops on normative multi-agent systems (NorMas05), on coordination and organization (CoOrg05), and the AAAI Fall Symposium on roles (Roles05).

Leendert van der Torre

Leendert van der Torre received the Ph.D. degree in computer science from Erasmus University Rotterdam, The Netherlands, in 1997. He is currently a Full Professor at the University of Luxembourg. He has developed the so-called input/output logics and the BOID agent architecture. His current research interests include deontic logic, qualitative game theory, and security and coordination in normative multiagent systems.

Harko Verhagen

Harko Verhagen received his Ph.D. degree in computer and systems sciences from Stockholm University (Sweden) in 2000 and is currently an associate professor at the department. His research has focussed on simulation of organizational behaviour, simulation as a scientific method, the use of sociological theories in multiagent systems research and more in particular theories on norms and autonomy.

Ten challenges for Normative Multiagent Systems

Guido Boella
Dipartimento di Informatica
Università di Torino
Italy
guido@di.unito.it

Leendert van der Torre
Computer Science and Communication Research Unit
University of Luxembourg
Luxembourg
leendert@vandertorre.com

Harko Verhagen
Department of Computer and Systems Sciences
Stockholm University / KTH
Sweden
verhagen@dsv.su.se

May 29, 2008

Abstract

At the second international workshop on normative multiagent systems, for short NorMAS07 [4], held at Schloss Dagstuhl, Germany, in March 2007, a shift was identified in the research community from a legal to an interactionist view on normative multiagent systems. In this paper we discuss the shift, examples, and ten new challenges in this more dynamic setting.

1 Towards a more dynamic interactionist view

Traditionally normative systems have been studied in philosophy, sociology, law, and ethics, and during the past two decades they have been studied in deontic logic in computer science (ΔEON). Normative multiagent systems is a research area where the traditional normative systems and ΔEON research fields meet agent research. The proposed solutions to the ΔEON research problems are changing, and solutions based on multiagent systems are increasing. Gradually the ΔEON research focus changes from logical relations among norms, to, for example, agent decision making, and to systems in which norms are created and in which agents can play the role of legislators. The eighth conference on Deontic Logic in Computer Science in 2006 in Utrecht, the Netherlands had as special focus "artificial normative systems" [10, 9], and the seventh conference [13, 14] in 2004 in Madeira, Portugal had as special theme "deontic logic and multiagent systems." Continuing this trend, the third workshop on normative multiagent systems is co-located in Luxembourg in July 2008 with the ninth conference on Deontic Logic in Computer Science [10, 21], which has as special topic "security and trust," and the fourth workshop will again be a Dagstuhl seminar to be held in March 2009.

The Agentlink Roadmap [15, Fig. 7.1.] observes that norms must be introduced in agent technology in the medium term for infrastructure for open communities, reasoning in open environments and trust and reputation. After four days of discussion, the participants of the second workshop on normative multiagent systems agreed to the following consensus definition:

"A normative multiagent system is a multiagent system organized by means of mechanisms to represent, communicate, distribute, detect, create, modify, and enforce norms, and mechanisms to deliberate about norms and detect norm violation and fulfilment."

The shift towards a more dynamic interactionist view on normative multiagent systems is reflected in the way this definition builds on its predecessor which emerged at the first workshop on normative multiagent systems held in 2005 as a symposium of the Artificial Intelligence and Simulation of Behaviour convention (AISB) in Hatfield, United Kingdom: "A normative multiagent system is a multiagent system together with normative systems in which agents on the one hand can decide whether to follow the explicitly represented norms, and on the other the normative systems specify how and in which extent the agents can modify the norms" [3]. The emphasis has shifted from representation issues to the mechanisms used by agents to coordinate themselves, and in general to organize the multiagent system. Norms are communicated, for example, since agents in open systems can join a multiagent system whose norms are not known. Norms are distributed among agents, for example, since when new norms emerge the agent could find a new coalition to achieve its goals. Norm violations and norm compliance are detected, for example, since spontaneous emergence norms of among agents implies that norm enforcement cannot be delegated to the multiagent infrastructure.

This shift of interest marks the passage of focus from the more static legalistic view of norms (where power structures are fixed) to the more dynamic interactionist view of norms (where agent interaction is the base for norm related regulation. This ties in to what Strauss [18] called "negotiated order", Goffmans [11] view on institutions, and Giddens' [8] structuration theory). The workshop vote on next generation scenarios for normative multiagent systems clearly preferred social scenarios like virtual communities and Second Life (over 50%) to more classical e-commerce settings where centralized solutions like e-institutions are used (less than 20%).

The legalistic view of normative multiagent systems is a top-down view which considers the normative system as a regulatory instrument to regulate emerging behavior of open systems without enforcing the desired behavior. Agents are often motivated by sanctions to stick to norms, rather than by their sharing of the norms. Even if agents are allowed some freedom to create norms, this freedom is mostly restricted to the possibility for agents to create contracts to regulate the interaction among them.

The interactionist view on normative multiagent systems represents a bottom-up view. In this autonomous individually oriented view norms can be seen, e.g., as regularities of behavior which emerge without any enforcement system because agents conform to them either because their goals happen to coincide, or because they feel themselves as part of the group or because they share the same values of other agents. Sanctions are not always necessary, where by sanctions we mean formal measures towards norm violating agents carried out by agents whose task it is to sanction norm violations, because social blame and spontaneous exclusion of non-conforming agents are often enough. This interactionist view, which has been promoted in the multiagent systems community by Cristiano Castelfranchi [6], becomes essential in applications related to virtual communities. In Second Life, for example, communities emerge in which the behavior of its members show increasing homogeneity.

To put this shift from legal to interactionist view into perspective, we can identify five levels in the development of normative multiagent systems. At level 1 of off-line norm design [17], norms are imposed by the designer and automatically enforced, and agents cannot organize themselves by means of norms. At level 2 of norm representation, norms are explicitly represented, they can be used in agent communication and negotiation, and a simple kind of organizations and institutions can be created. At level 3 of norm manipulation, a legal reality is created in which agents can add and remove norms following the rules of the normative system. Whereas existing normative multiagent systems are still at one of these first three levels of norm autonomy (for an introduction to norm autonomy in multiagent systems, see [22]), multiagent system research is now moving to level 4 of social reality, and is concerned with the ten challenges discussed in Section 3 below. We believe that there is at least one more level to be dealt with in the future. At level 5, the norms create a new moral reality. This goes beyond present studies in machine ethics [1], which is more concerned with agent decision making in the context of norms, which is an issue dealt with at each level of normative multiagent systems, than with creating a new ethics.

Clearly, for each level the development of the normative multiagent system will take a much larger effort than the development of similar systems at lower levels. For example, if norm are explicitly represented (level 2) rather than built into the system (level 1), then the system has to be much more flexible to deal with the variety of normative systems that may emerge. However, it may be expected that normative multiagent systems realized at higher levels will have a huge effect on social interaction, in particular on the web. We discuss some examples and several research needs that arise in this more dynamic interactionist view on normative multiagent systems.

2 Examples of an interactionist view

We illustrate the more dynamic interactionist viewpoint on normative multiagent systems using virtual communities in virtual reality settings like Second Life. In these virtual communities, human agents interact with artificial agents in a virtual world. The new communication instruments offered by the internet have resulted in the creation of virtual communities of users sharing information, emotions, or hobbies. When the interaction possibilities are multiplied in applications like Second Life or multi-player online games, new scenarios emerge. In particular, given the higher degree of freedom of behavior with respect to the real world, and the unaccountability offered by anonymity, on the one hand, as said above, spontaneous communities emerge showing regularities of behavior. However, to preserve the autonomy of the members of these communities, interactionist mechanisms for regulating behavior are needed. Thus, members of communities should be endowed with tools to make the community norms explicit and communicable to preserve their members' autonomy.

The participants will eventually end up creating their own norms and rules, even if in virtual communities like Second Life and in multi-player games normative infrastructure is imposed by the designers. Sometimes, the rules created by the participants counter the designers' objectives and rules and players start to play in ways unforeseen by the game designers. An example is discussed by Peter Ludlow [16] from Sony's EverQuest. EverQuest is a multiplayer online game where gamers are supposed to fight each other in a world of snakes, dragons, gods, and the Sleeper. Sony intended the Sleeper to be unkillable and gave it extreme high hit points. However, a combined combat of close to 200 players nearly succeeded to kill the 'animal'. Unfortunately, Sony decided to intervene and rescue the monster. Most of the discussion on this example has highlighted the decrease in trust of the game players in Sony, despite the fact that the next day Sony let the game players beat the Sleeper. However, in this paper we would like to highlight what this story tells us about the goals of game players, and its consequences for necessary technology in games. The following quote illustrates the excitement in killing the Sleeper.

> A supposedly [player-vs.-player] server banded together 200 people. The chat channels across the server were ablaze, as no less than 5,000 of us listened in, with OMG theyre attempting the Sleeper! Good luck d00dz! Everyone clustered near their screens, sharing the thrill of the fight, the nobility of the attempt and the courage of those brave 200. Play slowed to a crawl on every server as whispers turned to shouts, as naysayers predicted, It can't be done or It will drop a rusty level 1 sword and most of us just held our breath, silently urging them forward. Rumors abounded: If they win, the whole EQ

> world stops and you get the text from the end of Wizardry 1, or If they win, the president of Sony will log on and congratulate them. With thousands watching and waiting, the Sleepers health inched ever downward.
>
> . . .
>
> [EverQuest player] Ghenwivar writes, On Monday, November 17th, in the most amazing and exciting battle ever, [EverQuest guilds] Ascending Dawn, Wudan and Magus Imperialis Magicus defeated Kerafyrm, also known as The Sleeper, for the first time ever on an EverQuest server. The fight lasted approximately three hours and about 170180 players from [EverQuest server] Rallos Zeks top three guilds were involved. Hats off to everyone who made this possible and put aside their differences in order to accomplish the impossible. Congratulations RZ!!!" [16]

Normative multiagent systems study multiagent technology to support the emergent cooperation in online multi-player games like EverQuest [2]. The example illustrates that the game had been so well wrought that a real coalition of communities of players had formed, one that was able to set aside the differences between the communities, at least for a night, in pursuit of a common goal. This was not intended or foreseen by Sony, and getting two hundred people to focus on accomplishing the same task is a challenge.

> "Why, you might ask, would anyone waste four hours of their life doing this? Because a game said it couldn't be done.
>
> This is like the Quake freaks that fire their rocket launchers at their own feet to propel themselves up so they can jump straight to the exit and skip 90% of the level and finish in 2 seconds. Someone probably told them they couldn't finish in less than a minute.
>
> Games are about challenges, about hurdles or puzzles or fights overcome. To some players, the biggest hurdle or challenge is how to do what you (the designer) said couldn't happen. If you are making a game, accept this." [16]

A typical problem in virtual communities is caused by the ease in which new participants can enter the community, known as "newbies". The virtual communities should be able to defend itself from dangerous new players, and normative systems are a way to pose virtual gates to such communities. "Griefers would also maintain numerous alts that were sent out into greater Alphaville in attempts to scam and disrupt other houses. Because alts were usually abandoned soon after they had been created, they appeared to others as new characters, and this had the effect of making many players highly suspicious of newbies, and of generating virtual gated communities in response." [16] However, a virtual space should be able to deal with honest new participants. It has been noted that existing communities establish practices which tend to exclude newly entered participant in the virtual space: "Processes of norm building were visible, resulting in patterns of established users versus outsiders; new bonds were created, and users experience an appropriation of this newly created virtual public space: parts of the Digital City were 'taken over' by active established users who behaved as a closed community and were perceived accordingly by the outsiders." [20]

As illustrated by the "newbies" example, there are some aspects in which normative systems for virtual communities are more challenging than traditional regulations. For example, the construction of autonomous virtual communities cannot ground itself on an external legal system - apart from most serious cases like frauds going beyond the virtual environment - as in e-commerce applications that ground the validity of online contracts on the relevant human regulations. Consequently, these normative systems should be developed separately, in the same way as different national systems are created independently. Another issue is related to the possibility to augment actions in virtual scenarios: in these scenarios characters can be created with their own behavior that have more abilities then humans in the real world (e.g., flying, walking through wall), objects nor existing in reality, and even places. Moreover, the abilities of characters are not only related to the ones of their players: e.g., an avatar in Second Life entering a dancing room can acquire new dancing abilities which it did not have before and will lose afterwards. Thus the autonomy of characters assumes new dimensions.

3 Ten research challenges for the interactionist view

For the ten challenges posed by the interactionist viewpoint, we take the perspective from an agent programmer, and consider which kinds of tools like programming primitives, infrastructures, protocols, and mechanisms she needs to deal with norms in the example scenario. Similar needs exist at the requirements analysis level, or the design level, but we have chosen for the programming level since it makes the discussion more concrete, and this level is often ignored when norms are discussed. The list is not exhaustive, and there is some overlap between the challenges. Our aim is to illustrate the range of topics which have to be studied, and we therefore do not attempt to be complete.

Challenge 1 *Tools for agents supporting communities in their task of recognizing, creating, and communicating norms to agents.*

Even if social norms emerge informally, e.g., when a community becomes more complex and more open, an explicit representation of norms becomes necessary. There are still numerous philosophical problems for the representation of norms, see, for example, [12]. However, the new problem is the role of the agents and humans involved in the interaction with the multiagent system.

Challenge 2 *Tools for agents to simplify normative systems, recognize when norms have become redundant, and to remove norms.*

Challenge 2 is the counterpart of Challenge 1, because the natural tendency of overregulation creates the need for a counterbalance. Since all norms come with a cost, for example to process them, to communicate them, to maintain them, or to enforce them, norms should only be introduced when they are really needed, and they should be removed as soon as they are no longer needed. For example, when the number of violations is increasing, this is typically a case where norms must be changed or removed, rather than where norm enforcement has to be increased.

Challenge 3 *Tools for agents to enforce norms.*

If we allow communities of agents to create their own normative multiagent systems, then the issue of how to enforce the norms arises. In case a centralized approach is needed, the infrastructure should support the enforcement of norms created by the communities. In a distributed approach, roles should be defined for agents in charge of monitoring and sanctioning. The virtual environment can offer new opportunities for norm enforcement not found in the usual environments. For example, evidence about agent behaviors can be collected via the logfiles of the system.

Challenge 4 *Tools for agents to preserve their autonomy.*

Challenge 4 is the counterpart of Challenge 3, because there is a natural tendency to enforce norms by regimenting them into the system. The danger highlighted by Castelfranchi [7] is related to the "formalization" of the informal. Norms have the nature of general directives which cannot cover all cases nor avoid all conflicts with other norms. Thus, normative multiagent systems need to preserve the autonomy of agents regarding the making of decisions about norm compliance and norm violation. Agents in charge of monitoring and enforcing norms should be flexible enough to preserve the autonomy of the "norm subject" agents with respect to norm violations, for instance in circumstances that differ from the circumstances which the norms have been defined to preserve and where norm compliance is not advantageous for the normative multiagent system.

Challenge 5 *Tools for agents to construct organizations.*

As the example about EverQuest example shows, cooperation among the participants of virtual reality can result in coalitions which can achieve results which go beyond the ones reachable by their members. This is of great interest for participants in virtual reality, also because Second Life is becoming a place where business takes place. Thus, participants should be given some facilities and tools which allow the construction and management of organizations to achieve their goals. Note that in the real world such mechanisms exist, first of all the laws which allow the creation of organizations and attribute the responsibilities to different entities. E-institutions as proposed in multiagent systems can be a starting point, but they are often too flat - i.e., not hierarchically organized - and they usually do not support the dynamics of the underlying normative systems by allowing the creation of new norms.

Challenge 6 *Tools for agents to create intermediate concepts and normative ontology, for example to decide about normative gaps.*

In real institutions norms have a fuzzy character in the sense that they are not able to cover all possible situations. In particular because new situations can arise, e.g., due to technological advancement (for instance: is a digital signature the same as an handwritten signature?) This problem increases exponentially in virtual worlds where all kind of new behaviors and objects can be defined. The solution in real normative system is to endow some agents with powers to decide whether a new concept is subsumed by another one. The role of agents in the logical reasoning of a normative system is something which is still missing in the state of the art of the field.

Challenge 7 *Tools for agents to decide about norm conflicts.*

This challenge is related to Challenge 6 since norms do not cover all possible cases and conflicts between norms are possible. Thus agents need a mechanism to take decisions in situations of conflicting norms. The mechanism cannot always be automated, for example because the degree of freedom in virtual world to create new behaviors and objects norms may become underspecified. Thus, the problem is to define normative systems, where, like in human normative systems, roles are defined and role keepers are empowered to take decision when automated reasoning alone is not enough. At some point, the view of the normative system as a self contained logical system is not viable anymore.

Challenge 8 *Tools for agents to voluntarily give up some norm autonomy by allowing automated norm processing in agent acting and decision making.*

In many examples, the autonomy of the agent must be adjusted to the context. In general avatars are graphical representations of users of a system and can be seen as interface agents. Avatars living in Second life are interface agents for human players but also increasingly for autonomous agents. Consider the example above, where new abilities like dancing are automatically added to the avatar. Moreover, even if now prohibited, autonomous agents should be allowed to on the player's behalf cope with events that occur when the player is not online. It is possible to envisage a scenario where avatars are partially programmed to take autonomous decisions when the player is off-line. Among these decisions is whether to comply with norms of the community the avatar is acting into.

Note that these mechanisms are useful not only when the avatar is acting autonomously on behalf of its off-line owner, but also during the activity of the player. In real life norms are often violated just by distraction or ignorance or by lack of resources and the violator does not gain anything by its deviant behavior. The same will eventually happen in virtual worlds, especially when norms to be respected will not be necessarily intuitive or similar to the ones of real world. In these cases, the decision to conform to norms can be left to the avatar and the player can be relieved from this task. E.g., consider the case of communities where nudity of avatars is prohibited. The player could simply leave to its avatar the burden to conform to the norms by automatically disabling actions which are deviant with respect to the norms.

Challenge 9 *Tools for conviviality.*

Since scenarios like Second life are aiming at people having pleasant social interactions, and norms may interfere with the goals of the players, the impact of norms on this dimension must be considered. Norms should not constrain the freedom of participants too much and allow to avoid unpleasant behavior from other agents, but there is also a more subtle effect to be considered. Social interaction is regulated by social conventions, which can be modeled as a sort of institution. Part of the fun of "living" in Second life, like when participating in a carnival or when embodying a character of a drama depends - according to Taylor [19, 5] who calls this effect "conviviality" - is the temporary displacement with respect to the usual norms of social life. In particular, in the sense that in social relations the player acquires new social powers which he does not have in his first life.

The tools for conviviality should study social dependencies among players and indicate how these dependencies can be made less unbalanced by attributing more social powers to some players. Note that, as in the example about automatic learning of dancing abilities in Section 2, adding social powers in a virtual reality can take a more extended sense, since in the real world physical abilities cannot be added. Tools for conviviality should also facilitate the introduction of new participants in a virtual community by addressing the "newbies" problems.

Challenge 10 *Tools for legal responsibility of the agents and their principals.*

Nowadays, agents become subjects of human legislation. For example, it is debated if agents have responsibilities beyond the ones attributed to their owner, or if agents can be really attributed mental states which are to be taken into account in the attribution of responsibilities. However, in scenarios like Second life, new questions arise. Participants accept the rules of the game and they should be made aware whether following the rules of some communities leads to infringement of real legislations.

References

[1] M. Anderson and S. Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–26, 2007.

[2] G. Boella, P. Caire, and L. van der Torre. Norm negotiation in online multi-player games. *Knowledge and Information Systems*, To appear.

[3] G. Boella, L. van der Torre, and H. Verhagen. Introduction to normative multiagent systems. *Computation and Mathematical Organizational Theory, special issue on normative multiagent systems*, 12(2-3):71–79, 2006.

[4] G. Boella, L. van der Torre, and H. Verhagen, editors. *Normative Multiagent Systems*, Dagstuhl Seminar Proceedings 07122, 2007.

[5] P. Caire, S. Villata, G. Boella, and L. van der Torre. Conviviality masks in multiagent systems. In *Proceedings of the Seventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS'08)*, 2008.

[6] C. Castelfranchi. Modeling social action for AI agents. *Artificial Intelligence*, 103(1-2):157–182, 1998.

[7] C. Castelfranchi. Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations. *Journal of Applied Logic*, 1(1-2):47–92, 2003.

[8] A. Giddens. *The Constitution of Society.* University of California Press, 1984.

[9] L. Goble and J.J. Ch. Meyer, editors. *Deontic Logic and Artificial Normative Systems, 8th International Workshop on Deontic Logic in Computer Science, DEON 2006, Utrecht, The Netherlands, July 12-14, 2006, Proceedings*, volume 4048 of *Lecture Notes in Computer Science.* Springer, 2006.

[10] L. Goble and J.J. Ch. Meyer. Revised versions of papers presented in the proceeding of the eighth international workshop on deontic logic in computer science (DEON06). *Journal of Applied Logic*, in press.

[11] E. Goffman. *The Presentation of Self in Everyday Life.* Doubleday, 1959.

[12] J. Hansen, G. Pigozzi, and L. van der Torre. Ten philosophical problems in deontic logic. In G. Boella, L. van der Torre, and H. Verhagen, editors, *Normative Multi-agent Systems*, volume 07122 of *Dagstuhl Seminar Proceedings.* Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2007.

[13] A. Lomuscio and D. Nute, editors. *Deontic Logic in Computer Science, 7th International Workshop on Deontic Logic in Computer Science, DEON 2004, Madeira, Portugal, May 26-28, 2004. Proceedings*, volume 3065 of *Lecture Notes in Computer Science.* Springer, 2004.

[14] A. Lomuscio and D. Nute. Revised versions of papers presented in the proceeding of the seventh international workshop on deontic logic in computer science (DEON04). *Journal of Applied Logic*, 3(3-4), 2005.

[15] M. Luck, P. McBurney, and C. Preist. *Agent Technology: Enabling Next Generation Computing (A Roadmap for Agent Based Computing).* AgentLink, 2003.

[16] P. Ludlow and M. Wallace. *The Second Life Herald.* MIT Press, Cambridge (MA), 2007.

[17] Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: Off-line design. *Artificial Intelligence*, 73(1-2):231–252, 1995.

[18] A. Strauss. *Negotiations: Varieties, Contexts, Processes and Social Order.* San Francisco, Jossey-Bass, 1978.

[19] M. Taylor. Oh no it isn't: Audience participation and community identity. *Trans, Internet journal for cultural sciences*, 1(15), 2004.

[20] P. van den Besselaar. E-community versus e-commerce: The rise and decline of the Amsterdam digital city. *AI and Society*, 15(3):280–288, 2001.

[21] R. van der Meyden and L. van der Torre, editors. *Deontic Logic in Computer Science, 9th International Conference on Deontic Logic in Computer Science, DEON 2008, Luxembourg, July 16-18, 2008, Proceedings*, LNCS, Berlin, in press. Springer.

[22] H. Verhagen. *Norm Autonomous Agents.* PhD thesis, Department of System and Computer Sciences, The Royal Institute of Technology and Stockholm University, Sweden, 2000.