Stit, Iit, and Deontic logic for Action Types

Bentzen, Martin Mose

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Bentzen, M. M. (2010). *Stit, Iit, and Deontic logic for Action Types*. Roskilde Universitet.

Stit, Iit, and Deontic Logic for Action Types

Martin Mose Bentzen

SECTION FOR PHILOSOPHY AND SCIENCE STUDIES
ROSKILDE UNIVERSITY

This thesis is submitted in partial fulfillment of the requirements for obtaining the degree Doctor of Philosophy at the Section for Philsophy and Science Studies, Roskilde University.

This project has been conducted within the PhD program Science Studies at the Section for Philosophy and Science Studies, Roskilde University and at the Institute for Logic, Language and Computation, University of Amsterdam. The project has been supervised by Professor Stig Andur Pedersen, Section for Philosophy and Science Studies, Roskilde University and Professor Frank Veltman, The Institute for Logic, Language and Computation, University of Amsterdam.

Roskilde, December 21, 2009

Martin Mose Bentzen

Stit, Iit, and Deontic Logic for Action Types

Martin Mose Bentzen

Contents

Acknowledgements xi

1. Introduction 2
1.1 Stit theory for strategic situations 2
1.1.1 Outcomes 3
1.1.2 Actions 3
1.1.3 Agents 4
1.1.4 Values 7
1.1.5 Modalities in situations 8
1.1.6 Recent work on stit theory 10
1.2 Deontic logic 11
1.3 Logic of action 13
1.4 Philosophical logic 13
1.5 The problems confronted in this thesis 14

2. Deontic logic and iterated removal of dominated actions 17
2.1 Introduction 17
2.2 Luhmann on trust 19
2.3 Strategic situations 20
2.4 Utilitarian strategic models 21
2.5 Iterated removal of dominated actions 24
2.6 New deontic operators 28
2.7 Formalizing the examples 31
2.8 The Meinong-Chisholm thesis 33

3. Responsibility formalized 35
3.1 Introduction 35
3.2 Allowing 35
3.3 Being able to, refraining and preventing 37
3.4 Intentions in ethics and legal theory 38
3.5 Intentions in situations 39
3.6 Responsibility 42
3.7 Guilt 42
3.8 Moral blameworthiness and praiseworthiness 43
3.9 Knowledge in situations 46

3.9.1 Kanger on responsibility 48

4. Ability modalities and the metaphysics of agency 51
4.1 Introduction 51
4.2 The Brown-Horty double modality analysis of ability 51
4.3 Ability *must, can, may* and *might* 53
4.4 Logical relations between ability modalities 54
4.5 The Metaphysics of agency 57
4.5.1 Four reductionist positions about being and ability . . 58
4.6 Objections to the theory 61

5. Group responsibility 62
5.1 Joint agency 62
5.2 Joint strict agency 63
5.3 Joint refraining 65
5.3.1 Arendt on collective responsibility 68
5.4 Positive responsibility of groups 68
5.4.1 Holding members of groups personally responsible . . . 69

6. Frankfurt examples 72
6.1 Introduction 72
6.2 Factors 72
6.2.1 Formalizing the assassin example 73
6.3 Causal responsibility, agentive responsibility, overdetermination 75
6.4 The philosophical context of the Frankfurt examples 77
6.4.1 Reasoning about outcomes 79
6.4.2 Informal approaches to the Frankfurt examples 80
6.4.3 Lewis' causal responsibility approach 80
6.4.4 Approaches within stit theory 82
6.4.5 Analysis of a Frankfurt example 84
6.5 Frankfurt examples and negative responsibility 86
6.5.1 Inwagen on event particulars 86
6.5.2 Inwagen on event universals 87
6.5.3 In defence of negative responsibility 89

7. Deontic logic for action types 93
7.1 Overview of the chapter 94
7.2 Ross' paradox and free choice inferences 96
7.2.1 Ross' paradox - a problem for stit theory 98
7.3 Strong permission 99
7.3.1 Conjunction exploitation 100
7.4 Dynamic deontic logic 101
7.5 Other related approaches 103
7.6 Action types and action tokens 104
7.7 Logic 108
7.7.1 Syntax 108
7.7.2 Semantics 109
7.7.3 Validities 111
7.7.4 Non-validities 112
7.7.5 Equivalences 113
7.8 How intuitively adequate is the logic? 113
7.9 Natural language and expressivity 114

8. Action types in strategic situations 117
8.1 Syntax 117
8.2 Semantics 117
8.2.1 Evaluation of formulas 118
8.3 Action types and corresponding propositions 119
8.4 Intentions and action types 120
8.4.1 The right way of eating a pear - is not killing somebody while doing it! 120
8.5 Intended and unintended consequences of actions 122
8.5.1 Davidson's prowler example revisited 122
Summary 123
Resumé 125
Bibliography 127
Index 134

List of Figures

1.1 Four choices facing an agent a_1 6
1.2 Modalities in situations 10

2.1 Removing dominated actions 25
2.2 Hostage situation 32
2.3 The doctor's journey 32
2.4 Friends meeting in town 33

3.1 Skywalker's choice 45

4.1 Ability modalities 53
4.2 Square of opposition for action modalities 55
4.3 Cube of opposition for ability modalities 56
4.4 Square of opposition for *can* and *may* 56

5.1 A counter model to **5** 66
5.2 The choices of father and son 67
5.3 The choices of 3 agents 69

6.1 The choice of the assassin 74
6.2 Red button or green button 76
6.3 Man in a locked room 78
6.4 Frankfurt example 84
6.5 Random number generator 1 90
6.6 Random number generator 2 91

7.1 Consequences 94
7.2 Non-consequences 95
7.3 Equivalences 95

8.1 Killing and eating an pear 121
8.2 Turning on the light 122

Acknowledgements

Throughout the last three years Stig Andur Pedersen has provided me with excellent advice and unwavering support in matters logical as well as practical. I have learned a lot from Andur both professionally and personally. Andur has also had a great impact on the thesis itself, especially through his thorough comments on the technical parts of Chapter 2.

The meetings with Frank Veltman have been of tremendous importance for the development of this thesis. Especially the theory put forth in Chapter 7 has benefited a lot from his vigilant criticism. Frank has an impressive ability to always get to the core of a subject matter.

Thank you both very much.

Thanks a lot to Olivier Roy for his thorough comments on an earlier draft of Chapter 2, to Jens Ulrik Hansen for his equally thorough comments on Chapter 2, Chapter 6 and Chapter 7, and to Julia Bentzen for proof reading several chapters. Thanks to Claus Festersen and Klaus Frovin Jørgensen for advice regarding layout.

I presented Chapter 2 at the workshop *PhDs in Logic*, Ghent, February 2009 and central parts of Chapter 6 at the annual meeting of the Danish Philosophical Association, Aarhus, February 2009. Parts of Chapter 3 and Chapter 5 were presented at the *Second Workshop in Decisions, Games and Logic*, Amsterdam, July 2008, as well as at logic seminars in Roskilde, Amsterdam, and Utrecht in 2008. I thank the organizers of these events, not least the Ghent people for a fantastic workshop and for introducing me to many different flavors of Genever. I would also like to thank all my great colleagues and friends at the ILLC, University of Amsterdam and at the philosophy department at Roskilde University as well as the participants of the Logic Seminars at Roskilde University. I apologize for these generic categories, but the lists were getting way too long. A special thanks goes to Jelle Zuidema, though, for providing me with a cozy place to live in Amsterdam.

For their love and moral support I thank Julia Bentzen, Klaus Jupiter Bentzen, Aviaja Solsikke Bentzen, Marianne Mose Bentzen and Henning Bentzen.

Chapter 1

Introduction

The main purpose of this thesis is to develop tools for reasoning about the actions of free agents and for reasoning about the moral evaluation of these actions and their outcomes. Broadly speaking, the work is carried out within the tradition of philosophical logic. More specifically, contributions are made to deontic logic and to the logic of action. Even more specifically, the foundations of most of the work in this thesis are provided by stit theory.[1] In this introduction, I will say a bit about these topics in reverse order, starting out with some basic stit theory. The idea of the introduction is both to situate the thesis in a broader context and to introduce some philosophical assumptions about agency, values, situations and so on, which have guided the theoretical work to be presented in the rest of the thesis.

1.1 Stit theory for strategic situations

Stit theory (stit is an acronym for 'sees to it that') is a formal theory in the tradition of modal logic, which has been used to clarify questions arising in the philosophy of action and in the philosophy of norms, see Belnap et al. (2001), Horty (2001). The stit theory of Belnap, Perloff, Xu and Horty is set in a branching time indeterministic framework. Time is represented by a tree, the maximal branches of which are called *histories*. Any choice of an agent restricts the future to a subset of these histories. In this thesis, I will generally abstract away from time. This I do in order to focus on other aspects of situations than temporal ones, as witnessed by the rest of this thesis. The primary object of this reduced stit theory I will call a *strategic situation* or simply a *situation*. I take a strategic situation to consist of at least a set of agents, a set of actions for each agent and a set of outcomes for each action. Each outcome has an associated value.[2] These basic strategic situations will be extended with intentions, knowledge and action types in

[1] The exception is Chapter 7.

[2] The name *strategic situation* is derived from that of a *strategic game* in game theory. There is one main technical difference between a strategic game and a strategic situation and one main conceptual difference. The technical difference is that an action profile does not have to determine a single outcome in stit theory. The conceptual difference is that values do not represent instrumental rationality. Also, the objectives of stit theory differ from those of most game theory in being more affiliated with philosophy than economics. These points are elaborated on in the rest of this introduction.

the course of this thesis. Situations are represented formally by *utilitarian strategic models*. When no confusion is likely to arise, I use the term situation for both the real or imagined situation to be represented and for the formal model representing it. I will now say a bit about the individual components which constitute a strategic situation.

1.1.1 Outcomes

On several occasions in Horty (2001), the histories of stit models are referred to as 'possible outcomes of actions' e.g.:

> ...we speak of the histories belonging to an action K as the *possible outcomes* that might result from performing this action.
> (Horty; 2001, p.13)

Since the temporal aspects of situations will not be represented directly in this thesis, there will be no further mention of histories. Instead, situations are simply said to have various possible *outcomes*. These outcomes are formally the same as the possible worlds of standard relational semantics, see e.g. Chellas (1980). Indeed, it is also common to talk about possible worlds as outcomes in other areas of philosophical logic, especially when connections to probability theory are considered.

> Most representations of uncertainty ... start with a set of possible worlds, sometimes called *states* or *elementary outcomes*.[3]
> Halpern (2003)

An outcome represents one possible way the world may turn out as a result of the various choices of the agents. I usually write that a formula is true *with an outcome*. I also adopt the custom from probability theory of speaking of a set of outcomes as an *event* (in philosophical logic it is more common to call this a proposition).

1.1.2 Actions

The *actions* or *choices* of agents are represented by sets of outcomes. Thus each action or choice of each agent delimits the set of possible outcomes of

[3] That the outcomes are *elementary* means (in standard probability theory) that they are equally likely, an assumption not adopted in this thesis.

a situation. In stit theory two fundamental assumptions are made about choices. *The choices of each agent partition the outcomes* and the *independence of agents* condition.

The choices partition the outcomes

That the choices of an agent partition the outcomes of a situation implies that there is no such thing as an empty choice, a choice with no possible outcome. It also implies that every possible outcome of a situation is the outcome of some choice of each agent.

Independence of agents

Informally, *independence of agents* is the condition that each possible choice of each agent is consistent with every choice of every other agent. That two choices are consistent is taken to mean that they have at least one outcome in common (thus distinct choices of the *same* agent are never consistent). This condition implies that each choice of each agent is always open to that agent in the situation, no action of any agent depends on what any other agent does. Naturally, one could introduce such dependencies but doing so lies beyond the scope of this thesis. Throughout, the independence of agents condition is assumed to hold. Formally, independence of agents as described above is equivalent to the condition that each atomic action profile (one atomic choice for each agent) is non-empty. The independence of agents condition is rather strong and absolutely central to the account of agency offered by stit theory. Indeed, if some of the results in this thesis seem puzzling at first, it might in some cases be a good idea to check if they are related to the independence of agents condition.[4]

1.1.3 Agents

It will usually be assumed that the agents to be modeled are *free* and *accountable* in informal senses of those words. In particular, the concepts of

[4] Although rather strong, these two fundamental assumptions about agency are also part of the foundations of main stream game theory for strategic games. In game theory the action profiles are usually identified with the outcomes of the game - one way of understanding this is that game theory requires that the combined choices of all agents determine a unique outcome. This assumption is not part of stit theory.

responsibility presented in Chapter 3 and used in later chapters work best under those assumptions. I refrain, however, from a lengthy discussion of the concept of *free will* (Chapter 6 is somewhat of an exception to this), but refer the reader to works in mainstream philosophy, such as van Inwagen (1983), Kane (1998), Fischer and Ravizza (1998), Fischer (2005b), Mele (2006). In this thesis, the many interesting discussions on free will be crudely reduced to the following principle.

Principle 1.1. Free will presupposes indeterminism, control and purpose.

Agents cannot be free in a deterministic universe.[5] However, ontological indeterminism alone is not sufficient to establish free will. A random agent is not the same as a free agent. We also require that agents have at least some control over the way things turn out. But even mechanical devices can be used to control how events turn out. Free agents act with a purpose. Although agents may not be able to guarantee how things turn out they certainly intend things to turn out in a specific way by acting. I operationalize Principle 1.1 by representing all three factors in the formal models of situations. Indeterminism is represented by allowing more than one outcome of a situation. Control is represented by the agents' actions or choices which partition the outcomes. Purpose is represented by singling out a subset of outcomes of each action as intended outcomes. A sceptic might ask how we know that the agents make their choices freely? Which choice they finally make might e.g. be governed by a function from the available choices to a specific choice making the agents deterministic rather than free. Or there might be an objective probability connected to each choice making the choice random rather than free. This argument misses the point of formal modeling. The aim is not to explain why agents have a free will. It is assumed that they do. Rather, the aim is to find formal ways of representing these assumptions. The back and forth between informal discussions and the logical implications

[5] In modern philosophy this position is especially connected to the name of Inwagen, see van Inwagen (1983) and Kane (1998). Examples of philosophers, who accept the opposing compatibilist view that agents actually *can* be free given determinism are Fischer and Ravizza, see Fischer and Ravizza (1998). Mele defends *both* views in order to make the disjunction of the views more plausible and save free will from scepticism, see Mele (2006). Mele's book also contains a philosophical critique of the results of Libet, who argues against free will on the basis of his experiments in neuroscience. Chapter 6 of this thesis primarily consists in a discussion with a prominent compatibilist thought experiment, see Frankfurt (1969).

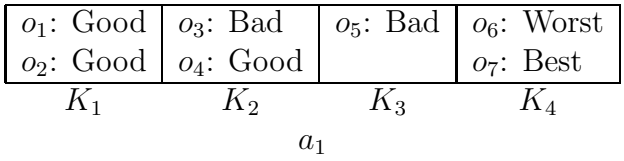

Fig. 1.1: Four choices facing an agent a_1

of the theory will help determine how natural this representation is. The fact that the models might be used equally well to represent artificial, determined agents is also besides the point. The point is that the underlying assumptions lend plausibility to the more precise concepts defined later in the thesis.

Example

Figure 1.1 provides an illustration of some of these concepts. There is just one agent in the situation, the agent a_1. There are 7 possible outcomes, $o_1, \ldots, o_7$. a_1 has four possible choices, $K_1, \ldots, K_4$. Each choice, except for K_3, restricts the possible outcomes to two. The labels 'Good', 'Bad' and so on can be regarded as propositional variables reflecting the value of the outcome. They are true with outcomes where they occur and only with these outcomes. They can be read: 'something good happens', etc. The two fundamental modalities in Horty's version of stit theory are the Chellas stit operator $[a_i\ cstit]$ and the deliberative stit operator $[a\ dstit]$.[6] With the present non-temporal version of stit theory, an agent sees to it that ϕ with an outcome according to the Chellas stit operator if and only if, every outcome of the choice made with that outcome makes ϕ true. For an agent to see to it that ϕ according to the deliberative stit operator, it is further required that it is possible that ϕ is false with some outcome of that situation. This rules out the possibility that an agent sees to necessary truths and events that happen to obtain in the entire model.[7] Nonetheless, in large parts of this thesis I stick to the simple Chellas stit operator as the primary *action*

[6] I deviate from Horty (2001) and Belnap et al. (2001), but follow recent practice in the literature on stit theory (and standard modal logic) by writing the complement ϕ of a stit operator outside of the square brackets, i.e. $[a_i\ cstit]\phi$ instead of $[a_i\ cstit : \phi]$.

[7] We might say that the latter are *presupposed* in the situation. Thus, in most situations (where agents do not die), we presuppose that air is present with every outcome, but we would not say that any agent sees to it that air is present.

modality (an exception is the joint deliberative stit operators in Chapter 4, and the concepts of responsibility in Chapter 3 are also mostly connected to the deliberative stit operator.) Now return to Figure 1.1. By making choice K_1 the agent sees to it that something good happens with either stit operator, because with every possible outcome of that choice 'Good' is true, and there is another outcome, e.g. o_3, where 'Good' is not true.

1.1.4 Values

The role values attached to outcomes play in this thesis differs from the role values play in main stream game theory or the role preferences play in main stream social choice theory. In this thesis, values or utilities are not considered to be connected to the preferences of a specific agent or a specific group of agents. In game theory, for instance, the utility functions represent the private preferences of individual decision makers. This is *not* an assumption made here. Rather, values represent *legal* or *moral* evaluations of outcomes. One purpose and effect of this is that values or preferences are severed from individual or group *choice* - the idea that choice is revealed preference is not adopted in this thesis. The ultimate choice of an agent in a given situation cannot be calculated from the legal values of the outcomes of the situation. Rather, the focus is on situations where agents might do what is wrong legally or morally. One way of conceptualizing this difference between kinds of values is by making a distinction between *public* and *private* sources of values. This distinction is an analogy to *objective* vs. *subjective* sources of information, as considered for epistemic modalities, see e.g. (Portner; 2009, p. 108). The legal values and moral values are more public than personal tastes and desires of an agent. In main stream game theory a common strategy is to consider a public source of values such as morality or the law as an *influence* on private values. These kind of considerations lie beyond the scope of this thesis. It should also be noted that although they are more stable than personal preferences, legal or moral values are not stable across all situations. Normally, in Denmark, an agent may not cross a red light in an intersection. If the agent sees a child in the street and she also sees a car approaching from far away and she still has plenty of time to save the child, the legal preference changes. If she does nothing and the child is run over she might get up to two years in prison for neglecting to help a human being in a life threatening situation when there is no particular risk to herself, see

(Greve; 2004, p. 49). A starting point for developing tools to analyze how legal preferences change over situations could be Liu (2008). However, this task is beyond the scope of this thesis.

1.1.5 Modalities in situations

The level of abstraction of situations

In this thesis, as in everyday life, situations are described via sentences. However, the languages used here are logical languages. To make the use of logic more clear it will be useful to go a bit more in depth with the anatomy of situations. Let me first indicate which level of abstraction I take sentences about a strategic situation to be on, as compared to some other theories. The following table goes from the more specific to the more abstract.

1. Utterances in a context. (Speech act theory).
2. Sentences in a context. (Kaplan's theory of Indexicals)
3. Sentences in a stit model (stit theory in Belnap et al. (2001))
4. Sentences in a situation (Game theory for strategic games, this thesis)

We can understand this as moving towards more and more abstraction in the following way. Kaplan abstracts away from concrete utterances of sentences in a context of utterance in order to preserve logical relationships between sentences and to be able to have sentences be present simultaneously in the same context. Thus, for a context c and a domain of objects $\mathfrak{U}$:

> ...the notion of ϕ *being true in* c and $\mathfrak{U}$ does not require an utterance of ϕ.
> (Kaplan; 2004, p. 781)

Belnap, Perloff, and Xu abstract speaker and place away from Kaplan's contexts, but keep the temporal aspects of contexts. Thus a formula such as $[a_i\ cstit]\phi$ is true or otherwise in a stit model, quite independently of who speaks and where that person is. In situations we also abstract away from time. Abstracting away from time is technically the same as confining stit theory to a single moment, as done in Kooi and Tamminga (2006). Informally, though, it gives us a bit more freedom in the way we think about

situations. For instance, I do not wish to require that all actions are momentary or that outcomes of the same action or different actions available to the agent occur in the same moment of time. There might be situations where an agent can choose to 'do the same thing' fast or slowly or 'today' or 'tomorrow', although we represent it as two concurrent choices in the same situation. The following discussion gives a bit more precise overview over how I consider modalities to be connected with situations.

Kinds of modalities and their relation to situations

The following modalities will be considered in this thesis: deontic, ability, action, intention, judgement and epistemic modalities. Further, it will be assumed that are three *primary temporal perspectives* on situation: before, during, and after. The modalities tend to be connected to one specific temporal perspective, excepting epistemic modalities, see Chapter 3. Figure 1.2 sum up this way of classifying the modalities. The first kind we call *situation* modalities. What an agent ought to do or is permitted to do (as expressed in sentences with deontic *must* or *may*), is able to do (as expressed with ability *can*) is usually considered before a situation takes place, and we think of these modalities as directed towards the future. Intuitively, these modalities are evaluated before an agent has chosen what to do. Since these modalities are considered before a certain outcome is determined and even before the agent is committed to a specific action, they are most naturally thought of as settled either true or false in the whole situation, i.e. for each of these modalities $\mathbf{m}$ and a situation S and a formula ϕ, we have either $M \vDash \mathbf{m}\phi$ or $M \nvDash \mathbf{m}\phi$ (where $M \vDash \phi$ is defined as $M, o \vDash \phi$ for each $o \in dom(M)$). We say that these modalities are settled in a situation or situation determinate.[8] The second kind we call *action* modalities, because they are most closely connected to a specific action that an agent is committed to.[9] In stit theory, an agent's actions partition the domain of the model, so an action is a set of outcomes. Where A is an action, let us define $M, A \vDash \phi$ as $M, o \vDash \phi$ for each $o \in A$. For action modalities M (such as *sees to it that*,

[8] This terminology comes from Horty, who calls the historical modalities *moment determinate*. The universal modalities $\mathbf{A}$ and $\mathbf{E}$ (corresponding to Horty's historical) are of course also situation determinate.

[9] The term *action* modality will thus be used in two ways: in a narrow sense about the specific action modalities *sees to it that* and *allows it that* and in a broad sense, including the intention modality.

Modalities	Deontic, ability	Action, intention	Judgement
Time	Before	During	After
Settled in	Situation	Action	Outcome

Fig. 1.2: Modalities in situations

allows it that, *intends it that*), given a situation M and an action A, we have $M, A \vDash \mathbf{m}\phi$ or $M, A \nvDash \mathbf{m}\phi$. We say that these modalities are settled with an action or action determinate. Intuitively, sentences containing these modalities refer to a time after the agent has chosen what to do, but before the actual outcome has been determined. I think of this time as *during* the situation. The reason I group the intention modalities here, is that I see intentions as relative to actions, see Chapter 3 for more details. Finally, the most unstable modalities are the *outcome* modalities to which belong the various concepts of responsibility considered in Chapter 3. These are only settled with specific outcomes. These modalities have to do with judgement and achievement. What an agent can be held responsible for might vary from outcome to outcome. For instance, if an agent does not succeed in what he intended to do, we do not hold him responsible for *doing it*, only for *attempting it*. Here, the primary temporal perspective is looking back at a situation after the situation has terminated in a specific outcome. It is only for this latter type of modalities, I talk about an *actual* outcome as opposed to other possible outcomes. The fact that I embrace indeterminism prohibits me from saying about any specific outcome that it is (or will be) the actual outcome during or before a situation. When I want to single out an outcome for evaluation in these cases, I prefer to call the outcome singled out the *considered* outcome (as opposed to other possible outcomes).

1.1.6 Recent work on stit theory

Stit theory is not old for a philosophical theory, but it makes sense to divide its brief history into two phases. The early work in stit theory was primarily philosophically motivated. This phase culminated in the two major works, Belnap et al. (2001) and Horty (2001). These books contain most of the material published in philosophical journals in the 1990's and extends it in new ways. In this decade, however, most of the work in stit theory has

been done by people in computer science, where the focus is more technical and aimed at feasible implementations. Much of this work has focussed on proof theory and connections to other multi agent formalisms. Xu was the first to axiomatize stit theory, Xu (1998). Alternative axiomatizations are presented in Balbiani et al. (2008). Wansing provided a tableaux system, Wansing (2006). These axiomatizations are in effect restricted to a single moment. In Herzig and Schwarzentruber (2008) it is shown that stit theory with the joint Chellas stit operator is not axiomatizable. Formal translations between stit logic and coalition logic and ATL are presented in Broersen et al. (2006a), Broersen et al. (2006b). These papers are mainly of interest because of complexity, axiomatizability and decidability results. It would also be interesting to undertake a more conceptual comparison between the formalisms on the basis of these results, but I leave this for somebody else. More conceptually oriented work has also been done within this community. In Broersen (2008), the task of integrating epistemic modalities with the stit framework is begun, see also Chapter 3 for an alternative. Some work on stit has been done in philosophy, in particular by Thomas Müller. His work on stit theory has been centered around the connection between action and time. In Müller (2006), he considers trust over time. In Müller (2005), he considers the duration of actions. In Kooi and Tamminga (2006), stit theory is reduced to a single moment, a reduction that technically corresponds to what is done in this thesis, and utility functions for each agents are used to formalize game theoretical reasoning. A very interesting paper written by people in the computer science community is Troquard et al. (2006). Here the authors combine the modal or intensional view of *agency* provided by stit theory with a first-order view on *actions* considered as a kind of objects. In philosophy this roughly translates into integrating Davidson's view with that of stit theory. I will return to that issue in Chapter 7 and Chapter 8, but the more technical work referred to above is not the main concern of this thesis.

1.2 Deontic logic

My personal starting point for this work is an interest in deontic logic, see McNamara (2006) for a survey. In this thesis specific deontic logics will be considered and developed. For now, I will confine myself to a few remarks on the fundamental question, whether it makes sense to assign truth values to

deontic sentences in order to reason with them logically, a problem known as *Jørgensen's dilemma*, see Jørgensen (1937), Hansen et al. (2007), McNamara (2006). As mentioned above, this thesis mainly concerns public values such as legal and moral values. These are 'social facts' to such a degree that it makes sense to assign truth values to sentences on the basis of them relative to situations. Thus in a specific situation, it can be true (or false) that a person is allowed to cross the street. This truth value is relative to a situation and further it must be independent of a particular outcome, i.e. situation determinate to account for the future directed nature of deontic sentences.

Are deontic sentences performative or descriptive?

Some linguists think that deontic 'must' and 'may' are always performative in the sense that their utterances create an obligation or a permission for an addressee, see (Portner; 2009, p. 190). I do not agree with this position. I have no interest in arguing that there is such a thing as 'objective' or 'eternal' values. On the other hand, the public nature of some values make these speaker independent (e.g. independent of the particular authority of the speaker or a certain 'chain of command'). They are generally accepted by some community.[10] Further, it makes sense to abstract away from specific utterances of deontic sentences. The meaning of these sentences can be derived from the values given to outcomes in particular situations. Thus, relative to a situation and a set of values, deontic sentences have a truth value. On the other hand, I agree that deontic 'must' and 'may' *can* have interesting performative functions and that these functions have not been studied enough in the literature. Presumably, if we want to stick to the present theoretical framework, the performative aspect of such deontic sentences could be considered as the situation changing potential of these sentences. Naturally, such performative sentences will not have determinate truth values in any of the senses considered above, as they dynamically change the 'ontology' of the situation, e.g. by changing the values of outcomes or choices of agents. Rather they could be considered functions that gives a new situation from an old one, see also Chapter 7. Extending the present theory with tools of this kind lays beyond the scope of this thesis. Here deontic *ought*, *must* and *may*

[10] Whether and how the public values are or should be generated from the private values of some group of individuals is not considered in this thesis, see e.g. Kooi and Tamminga (2006).

are always used as descriptive terms relative to given valuations of outcomes of specific situations.

1.3 Logic of action

Above I have described how agency is conceived in this thesis. Historically, the modern origin of agency logic is with von Wright, but formally stit theory owes most of its debts to Chellas. Chellas' account of agency, which has been adopted by stit theory, has been criticized by Segerberg. In the following quotation the 'cones' are the actions of an agent considered as subsets of histories:

> While this is not implausible, it would have been interesting to have been told something about the connexion between the agent and those cones. What is it that makes an initial history continue in one fashion rather than another? Does the agent "do" anything at t - 1 to define a certain cone – does action consist in choosing or somehow committing oneself to a cone? Otherwise, where does action come from? And when does it take place – at t - 1, at t, at the interval [t - 1, t], or what?(Segerberg; 1992, p. 373)

The main competitors to stit theory are variants of dynamic logic, see also Chapter 7. From computer science (but with many philosophical implications) we have dynamic logic which is related to concurrent dynamic logic, see e.g. Goldblatt (1992), which is related to game logic which is related to coalition logic, see Pauly (2001). For a survey of some agency logics up until 1992, see Segerberg (1992). For a bibliography up until then, see *Selective Bibliography in the Logic of Action* (1992). For surveys of the later development, see Lindström and Segerberg (2007), Segerberg et al. (2009).

1.4 Philosophical logic

Portner distinguishes the goals of logic and semantics in linguistics as follows.

> ... the primary goal of the semanticist is to provide a precise theory of the meaning of modal expressions across languages... The goal of the logician is to systematize and understand important features of reasoning with the concepts of necessity, obligation,

> and so forth.
> (Portner; 2009, p. 29)

With this definition, the aims of the thesis are clearly logical rather than linguistic. I am not primarily engaged in finding the meaning of natural language words but in a creative exploration of concepts in order to clarify and even develop those concepts themselves. Further, I do find the idea of semantic facts that we need to discover or predict a bit unsettling. The definitions of concepts given in this thesis are meant as suggestions which are open to discussion and revision. To a philosopher it should not be a problem that the object of exploration changes as a consequence of that very exploration. Rather, this must be considered philosophical progress. Perhaps the empirically minded linguist would have a problem with this (in the sense that changing the meaning of a word as a result of her very semantic analysis of it would require some explanation on her part). However, I am perfectly ready to learn otherwise, and nothing in this thesis really hinges on these remarks. Finally, very austere logicians might not even agree with me that this a work of *logic*, since, all through the thesis, I only present (syntax and) semantics and not any proof theory of the logics. The focus is conceptual.

1.5 The problems confronted in this thesis

The following is a brief description of the central problems to be confronted and attempted solved in this thesis. The first two problems can be viewed as limitations of stit theory as it has been developed up until now. The third problem comes from deontic logic. The fourth problem comes from the philosophy of action and linguistics and the fifth problem comes from ethics.

1. Intentions in stit theory.
2. Action types in stit theory.
3. Deontic paradoxes.
4. Ability modalities.
5. Frankfurt Examples.

1. *Intentions* do not form part of the stit theory presented in Horty (2001), Belnap et al. (2001). In fact, Horty explicitly mentions some difficulties with implementing intentions into the models. On the other hand, intentions are getting studied in other closely related areas, see Rao and Georgeff (1997), Roy (2008), and they play a crucial role when assigning legal and moral responsibility for events to agents. Therefore intentions are introduced explicitly in Chapter 3. 2. Stit theory has been criticized by Segerberg for not being able to talk directly about actions. In Chapter 7 and 8 I interpret this as a lack of being able to apply modal operators to *action types*. In particular, the deontic and ability modalities, i.e. the situation modalities, seem to be applied to action types in natural language, as in 'you must swim', etc. 3. There are many deontic paradoxes, but in this thesis I treat only on the ones connected to disjunction and conjunction, *Ross' paradox* and *free choice inferences*. This is also done in Chapter 7 and Chapter 8 . The other big cycle of paradoxes connected to the conditional, see Prior (1954), Chisholm (1964), Castañeda (1981), Forrester (1984), Prakken and Sergot (1997), I leave alone, partly because I think they can be treated quite accurately with the tools provided in Horty (2001). Some of them can be solved by the temporal framework (temporal versions of the good samaritan), and some of them by conditionalizing over relevant subsets and then restricting the values to those sets (for instance, the Gentle Murder paradox, see Forrester (1984), might be solved by restricting the attention to the outcomes where there is a murder and then taking the optimal action relative to this, see (Horty; 2001, chp. 5)). I am sure much more could be said, but I do not have anything else to say and I will not return to the conditional paradoxes again in this thesis. 4. In Horty (2001) a concept of *ability* is formalized within stit theory as (with present notation) $\mathbf{E}[a_i\ cstit]\phi$ inspired by Brown (1988). I formalize some related ability modalities in Chapter 4. 5. Finally, I investigate some philosophical thought experiments known as *Frankfurt examples* within the theoretical framework provided by stit theory.

The main results, I achieved by working with these problems are the following.

- A formalization of various concepts of responsibility.
- A deontic logic for action types.

Minor results are the transposing of iterated removal of dominated actions into the framework of stit theory, the cube of opposition for ability modalities

and a proof that if God exists he is solely responsible for everything.

A chapter by chapter overview is provided as part of the summary.

Chapter 2

Deontic logic and iterated removal of dominated actions

Trust is a solution for specific problems of risk.

Luhmann (1990)

2.1 Introduction

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

Example 2.1. The Victim is held up by the evil guy. He is wondering whether to attempt to resist the evil guy or not. The Hero is wondering whether to help or not. The best outcome is when the Victim tries to resist the evil guy and the Hero helps. Nobody gets hurt, the evil guy goes to jail. With the second best outcome, the Hero helps but the Victim remains inactive. Here the evil guy gets killed and the Hero and the Victim will both suffer some bad wounds. The third best outcome is when the Hero does not help and the Victim does not resist. The Victim will get killed but without too much suffering. The worst outcome is when the Victim tries to resist and is not helped. In this case the evil guy tortures him to death.

What should the Hero and the Victim do in this situation? The Hero can reason with the sure thing principle as follows. Given that the Victim resists, it is better for me to help, in which case all is well, than not to help, which would yield the worst possible outcome. On the other hand, given that the Victim does not resist, it is still better for me to help, because the good guys suffering some wounds and killing the bad guy, is still better than letting the Victim die. What should the Victim do? It seems impossible to say. Of course, if the Hero helps, he is a lot better off resisting, the best

possible outcome of the situation. On the other hand, if the Hero does not help, he will be tortured to death by making this choice, which would be absolutely terrible. If he does not resist, he might get rescued anyway if the Hero decides to help, but the rescue will come at a high cost. On the other hand, if the Hero decides not to help, he will at least die a clean death and not be tortured. It is really a predicament. But assume now that the Victim trusts the Hero to be a good utilitarian. Suppose, in particular, that he trusts the Hero to not make a choice which is strictly dominated. In that case, the Victim trusts the Hero to help. The Victim can now reason with the sure thing reasoning as follows. If I resist, then we will easily overcome the evil guy together. On the other hand if I don't resist, I leave all the dirty work to the Hero who will have to kill the bad guy and we will both get hurt. It is thus better for me to resist.

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

Example 2.2. The Doctor needs to reach town fast from the jungle to get medicine. It is a difficult journey. She can walk through the mountains or travel by boat down the river. She can also decide to abandon the journey altogether. Nearby lives the Guide, who has heard about this. He has to decide whether to come and guide the Doctor on the journey. Naturally, if the Doctor stays home, he would rather stay home, too. But if the Doctor should decide to either walk or go by boat he will be able to get her there faster either way, possibly saving lives. In particular, if the Doctor goes by boat, the Guide's navigational skills makes him very useful. It would yield the best possible outcome, if he were to decide to come and the Doctor were decide to go by boat.

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

ought to come help. And because the Doctor trusts the Guide to trust her and she trusts the Guide, she ought to go by boat.

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

2.2 Luhmann on trust

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

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

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

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

And further,

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

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

2.3 Strategic situations

In the version of stit theory studied here, we do not consider time, see also Chapter 1. Formally, it corresponds to stit theory reduced to a single moment, as studied, e.g. in (Belnap et al.; 2001, Chapter 16), Kooi and Tamminga (2006). Intuitively, since we use only operators, whose satisfaction (in the full stit framework including time) would not depend on histories, throwing away these histories from the models at the outset should not matter logically. It makes the model theory simpler, since we essentially reduce the models to standard relational models known from modal logic, see e.g. Chellas (1980), Blackburn et al. (2001). For a formal mapping between the two kinds of models, see, Herzig and Schwarzentruber (2008). Conceptually, I do not think we should consider the models as representing *single moments*.

Rather, we should consider them as *strategic situations*, which is to say that agents act *independently* (meaning that each of their choices is consistent with any choice of any other agent), but not necessarily *simultaneously*. In the informal examples presented in this chapter the agents are in different locations, and they cannot communicate. Further, in these models, each agent is aware of all *possible* consequences of the different combination of choices and this awareness is common knowledge. Also, agents agree on which utility to assign to outcomes (only ordinal aspects of utilities are used), and this evaluation is also common knowledge. The conceptual difference between the account given here and those based on instrumental rationality such as in game theory (a difference which really only exists at a meta level) is that we do not require the utilitarian values to correspond to the individual utility functions of agents (these are not part of the formal framework). For more about knowledge and stit, see Broersen (2008), and for knowledge in strategic situations in game theory, see van der Hoek and Pauly (2007). The agents do not know which particular choices the other agents will make. The examples suggest that there might be many situations, where these assumptions are quite natural. The deontic operators presented here for the first time, represent reasoning about what such an agent *ought to do* in such situations given various levels of trust. What is also new is the construction of submodels using positive formulas. This construction is applied in the iterative removal of strictly dominated choices.

2.4 Utilitarian strategic models

Throughout this thesis, I presuppose rudimentary set theory, see e.g. Devlin (1992)[1], as well as classical logic, see e.g. Smullyan (1968), Ebbinghaus et al. (1996), as parts of the meta langauge. Otherwise, the following presentation of the formal framework is self contained. For information about deontic logic, see von Wright (1951), Hilpinen (1971) Chellas (1980), Hilpinen (1981), Nute (1997), McNamara (2006). Introductions to standard game theory are Osborne (2004), Myerson (1997).

Formally, we use *utilitarian strategic models* (or simply *models*) consisting of outcomes, agents, choices, a utility function on the outcomes and a valuation function.

[1] In this thesis, $\subset$ always denotes strict inclusion between sets.

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

Definition 2.1 (Utilitarian strategic model)**.** A utilitarian strategic model is a structure $M = \langle W, Agent, \{Choice_i \mid a_i \in Agent\}, u, V\rangle$, where

1. W is a non-empty set of *outcomes*.
2. We have finite, non-empty set of *agents*, $Agent = \{a_1, \ldots, a_n\}$
3. For each agent $a_i \in Agent$, we have a finite, non-empty set $Choice_i$ of *choices* or *actions*, $Choice_i = \{A_{i1}, \ldots, A_{im_i}\}$, $0 < m_i$, (agent a_i has m_i choices). $\emptyset \neq A_{ij} \subseteq W$. Furthermore,
   (a) For any agent a_i, the elements of $Choice_i$ partition W.
   (b) Let $A_{1j_1} \in Choice_1, \ldots, A_{nj_n} \in Choice_n$, where $1 \leq j_i \leq m_i$. Then $(A_{1j_1} \cap \ldots \cap A_{nj_n}) \neq \emptyset$. (Independence of agents)
4. u is a utility function assigning a real number to to each outcome, i.e. $u : W \to \mathbb{R}$. Intuitively, the real number represents the moral or legal value of the outcome.
5. V is a valuation function from propositional variables to subsets of W, i.e. $V : \Phi \to \mathcal{P}(W)$

When $M = \langle W, Agent, \{Choice_i \mid a_i \in Agent\}, u, V\rangle$ is a model, I sometimes write $dom(M)$ for W (the domain of M). An *event* is defined as a subset of the domain. It should be noted that these models obey the *finite choice condition*, each agent only has a finite number of choices. The choices of a group of agents $\Gamma \subseteq Agent$ is defined as follows, see (Horty; 2001, p. 31). Let $Choice = \bigcup_{a_i \in Agent} Choice_i$. Let $Select = \{s \mid s : Agent \to Choice$ and $s(a_i) \in Choice_i\}$ be the set of selection functions assigning to each agent an action of that agent. If $\Gamma = \emptyset$, then $Choice_\Gamma = dom(M) \times dom(M)$. If $\Gamma \neq \emptyset$, then:

$$Choice_\Gamma = \{\bigcap_{a_i \in \Gamma} s(a_i) \mid s \in Select\}$$

Note that $Choice_{\{i\}} = Choice_i$. The set of *states* confronting an agent $a_i \in Agent$ in a situation is defined as the possible choices of the rest of the

agents considered as a group, see (Horty; 2001, p. 67):

$$State_i = Choice_{Agent-a_i}$$

In order to get to sure thing reasoning the utility function on outcomes is lifted to a dominance ordering on actions in two steps. First, the utilities of outcomes are lifted to a *weak preference ordering* on arbitrary events $S, T \subseteq W$, in the following way.

Definition 2.2. Let M be a utilitarian strategic model and $S, T \subseteq dom(M)$. Then $S \leq T$ (T is *weakly preferred* to S) iff $u(o) \leq u(o')$ for each $o \in S$ and each $o' \in T$.

$\leq$ is a transitive relation on $\mathcal{P}(W)$. A *strong preference ordering* on propositions is defined as $S < T$ iff $S \leq T$ and not $T \leq S$. $<$ is a strict partial ordering (transitive and irreflexive) on $\mathcal{P}(W)$.

Next, this preference ordering is used to define a *dominance* ordering on actions based on Savage's *sure thing principle*. Intuitively, this principle has the following content. A choice dominates another choice whenever the first choice is preferred in every state confronting the agent. It is presupposed that the states (the possible joint actions of the rest of the agents) are causally independent of the choices made by the agent.

Definition 2.3. Let $a_i \in Agent$ and let A_{ij}, $A_{ik} \in Choice_i$. Then $A_{ij} \preceq A_{ik}$ (A_{ik} weakly dominates A_{ij}), iff. $A_{ij} \cap S \leq A_{ik} \cap S$ for each $S \in State_i$. $A_{ij} \prec A_{ik}$ (A_{ik} strongly dominates A_{ij}) iff $A_{ij} \preceq A_{ik}$ and not $A_{ik} \preceq A_{ij}$.

$\preceq$ is transitive relation on $Choice_i$ and $\prec$ a strict partial ordering on $Choice_i$.

The following result is almost immediate.

Fact 2.4. $A_{ij} \prec A_{ik}$ iff $A_{ij} \preceq A_{ik}$ and $A_{ij} \cap S < A_{ik} \cap S$ for some $S \in State_i$.

Horty's terminology deviates a bit from the standard terminology of game theory here. The concept referred to above as *strong dominance* game theorists often call *weak dominance*, see (Osborne; 2004, pp. 46-47) . They will then refrain from naming the concept called weak dominance by Horty. The concept referred to by game theorists as *strong dominance*, in Myerson (1997), or *strict dominance* in Osborne (2004) is stronger yet. To distinguish the three concepts we will refer to the latter as *strict dominance* with the understanding that this is to be taken as different from strong dominance. To get there we define the following *strict preference ordering* on events.

Definition 2.5. Let M be a utilitarian strategic model and $S, T \subseteq dom(M)$. Then $S <_s T$ (T is *strictly preferred* to S) iff $u(o) < u(o')$ for each $o \in S$ and each $o' \in T$.

Thus strict preference requires that every outcome in one event has a strictly higher utility than every outcome in another event. It is a strict partial ordering. This ordering is then lifted to a *strict dominance ordering* on actions in the following way.

Definition 2.6. Let $a_i \in Agent$ and let A_{ij}, $A_{ik} \in Choice_i$. Then $A_{ij} \prec_s A_{ik}$ (A_{ik} strictly dominates A_{ij}), iff. $A_{ij} \cap S <_s A_{ik} \cap S$ for each $S \in State_i$.

Strict dominance requires that one choice is strictly preferred to another in each state. It is obvious that if choice strictly dominates another choice it strongly dominates it, but not necessarily vice versa.
We define the set of optimal choices for an agent a_i in a model M, denoted $Optimal_i$ as the set of actions for that agent that are not strongly dominated.

Definition 2.7. $Optimal_i = \{A_{im} \in Choice_i \mid$there is no $A_{in} \in Choice_i$, such that $A_{im} \prec A_{in}\}$.

Fact 2.8. *(Horty 2001)* For any agent a_i, $Optimal_i \neq \emptyset$.

Proof. We repeat Horty's proof in the current (atemporal) framework for the convenience of the reader. Assume $Optimal_i = \emptyset$. Let $A_{in} \in Choice_i$. Since $A_{in} \notin Optimal_i$, and $\prec$ is irreflexive there is a different action $A_{il} \in Choice_i$ such that $A_{in} \prec A_{il}$. By assumption $A_{il} \notin Optimal_i$ either, and $\prec$ is transitive and irreflexive, so cycles cannot occur. Hence we can iterate the argument indefinitely, giving us an infinite subset of $Choice_i$ contradicting that an agent has only finitely many actions. □

2.5 Iterated removal of dominated actions

Intuitively, iteratively removing dominated actions involves removing dominated actions from the model, which may in turn make it possible to remove other dominated actions, and so on. The reasoning leading to the removal of these choices is based on trust between agents. In the examples given above the right results can be obtained by removing either strongly dominated choices or strictly dominated choices. However, when removing

	A_1	$o_1 : u(o_1) = 2$	$o_2 : u(o_2) = 2$
a	A_2	$o_3 : u(o_3) = 0$	$o_4 : u(o_4) = 1$
	A_3	$o_5 : u(o_5) = 1$	$o_6 : u(o_6) = 0$
		B_1	B_2

b

Fig. 2.1: Removing dominated actions

strongly dominated choices (what game theorists call removing weakly dominated choices), the following complication arises: the choices which remain after the process has been completed depends on the order in which choices are removed. As an example consider Figure 2.1.

If we first remove a's strongly dominated choice A_3, then we can remove the dominated B_1, leaving b with B_2 as her only optimal choice. On the other hand, if we first remove a's strongly dominated choice A_2, then we can remove B_2, leaving b with B_1. It can be proven that the order does not matter when iteratively removing *strictly* dominated choices, see (Osborne; 2004, Chapter 12). The connection between such an algorithm and dynamic epistemic logic has been investigated (as well as other connections) , see van Benthem (2007). It could be shown that a deontic operator based on the choices remaining after this algorithm had come to an end would be logically independent of Horty's deontic ought to do. For that reason, since the aim here is to generalize Horty's account of deontic logic, we will deal with the order problem mentioned above in another well-known way: each step in the algorithm will be to simultaneously remove every strongly dominated action of every agent. Clearly, for a given agent a_i, the set of strongly dominated choices $Choice_i - Optimal_i$ is fixed. Intuitively, this algorithm can be justified by requiring that agents trust that other agents will reason by the sure thing principle.

Now for a formal account. We call an action $A_{ij} \in Choice_i$, where $1 \leq j \leq m_i$, an *atomic action*. For an agent $a_i \in Agent$, we call the union of k_i ($k_i > 0$) actions from $Choice_i$, a *complex positive action* and we denote such a complex positive action α_i, i.e.

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

(One may think of α_i as successively picking or leaving out each atomic action

from $Choice_i$, possibly leaving out some, but picking at least one). Given a complex positive action α_i for each $a_i \in Agent$, we define an *action profile*, denoted P, as the intersection of the complex actions, i.e.

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

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

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

Definition 2.9. $M|P = \langle P, Agent, \{Choice'_i \mid a_i \in Agent\}, u' : P \to \mathbb{R}, V' \rangle$, where

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

We need to show that $M|P$ fulfils the conditions of Definition 2.1, in particular that the actions of agents partition $dom(M|P)$ and that $M|P$ fulfils the independence of agents condition. We show only the latter.

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

From this we get the following.

Fact 2.10. Let M be a strategic model and P an action profile. $M|P$ is a strategic model.

Although restricting a model to an action profile is a sufficient condition for getting a new strategic model it is not necessary. The main thing is that the new model needs to fulfil the independence of agents condition.

This rules out reductions based on the truth set of any formula, as in the public announcements considered in dynamic epistemic logic, see e.g. van Ditmarsch et al. (2007).

Given a model M, we order the set of models $M_P = \{M|P \mid P \text{ is an action profile}\}$ as follows. $M|P' <_c M|P$, iff $P \subset P'$.

Fact 2.11. M_P is finite. $<_c$ is a strict partial order on M_P with the atomic action profiles as maximal elements and M as minimal element.

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

Now, fix a model M. $\bigcap(\bigcup_{a_i \in Agent} Optimal_i)$, is an action profile, which we denote $Optimal^M$. So, by Fact 2.10, $M|Optimal^M$ is a strategic model. We define a model M_n with level of trust n as the model resulting from n rounds of simultaneously removing each strongly dominated action of each agent.

Definition 2.12. 1. $M_0 = M$.

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

Fact 2.13. 1. $dom(M_{n+1}) \subseteq dom(M_n)$.

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

3. There is an m, s.t. $M_n = M_m$, for all $n \geq m$. We call this model M_m, $Optimus'$.By $Optimus'_i$ we mean $Optimal_i^{M_m}$.

Proof. 1. Obvious from Definition 2.9.
2. Since $n > m$, M_n is obtained from M_m in a finite number of steps for each of which 1. holds, so we have $dom(M_n) \subseteq dom(M_m)$.
3.For any n, either $M_{n+1} = M_n$ or for some agent some action is dominated, in which case $Optimal^{M_{n+1}} \subset Optimal^{M_n}$, i.e. $M_n <_c M_{n+1}$. Now, since for any n, $M_n \in M_P$, and $<_c$ yields a finite partial order, this process must come to an end eventually, at the latest when it hits a maximal element (an atomic action profile, i.e. each agent is down to one non-dominated action). The models can only get smaller and they never become empty. □

Furthermore, we have the following.

Fact 2.14. For any n and $a_i \in Agent$, $\{A_{ij} \in Choice_i \mid A_{ij} \cap Optimal_i^{M_n} \neq \emptyset\} \neq \emptyset$.

In words there is a non-empty subset of actions from the original model consistent with the actions of the model restricted to $Optimal^{M_n}$. We denote this set $Optimal_i^n$, i.e. $Optimal_i^n = \{A_{ij} \in Choice_i \mid A_{ij} \cap Optimal^{M_n} \neq \emptyset\}$, call it *the n-optimal actions for agent* a_i. Similarly, for $\{A_{ij} \in Choice_i \mid A_{ij} \cap Optimus_i'\}$, we write $Optimus_i'$. We have the following.

Fact 2.15. 1. $Optimal_i^n \subseteq Optimal_i^m$, for $m < n$.

2. There is an m, such that $Optimal_i^m = Optimal_i^n$, for any $n > m$.

2.6 New deontic operators

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

Definition 2.16. $a_i \in Agent$, $p \in \Phi$.

$$\phi ::= p \mid \bot \mid \phi_1 \rightarrow \phi_2 \mid \bigcirc\phi \mid \mathbf{A}\phi \mid [a_i\ cstit]\phi \mid \bigodot_n[a_i\ cstit]\phi \mid \bigodot[a_i\ cstit]\phi$$

We define the rest of the propositional connectives, $\neg, \wedge, \vee, \leftrightarrow$, in the standard way, ($\neg\phi$ is defined as $\phi \rightarrow \bot$, and so on). As usual, we write $M, o \vDash \phi$, for ϕ is true with outcome o of model M. The truth conditions for atomic sentences, the propositional constant, and propositional connective are standard:

Definition 2.17. 1. $M, o \vDash p$ iff $o \in V(p)$, where $p \in \Phi$ (p is atomic).

2. $M, o \nvDash \bot$ never.

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

We define the *event expressed by a formula* ϕ, denoted $|\phi|_M$ as the set of outcomes where the formula is true, i.e. $|\phi|_M = \{o \mid M, o \vDash \phi\}$. As usual, by ϕ being true in a model, written $M \vDash \phi$, we mean that ϕ is true with

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

Definition 2.18. $M, o \vDash \bigcirc\phi$ iff. there is an outcome o', such that $M, o' \vDash \phi$ and for all o'', such that $u(o') \leq u(o'')$, $M, o'' \vDash \phi$.

This is a normal modal operator, validating e.g. **D** $(\neg(\bigcirc\phi \wedge \bigcirc\neg\phi))$ and **4** $(\bigcirc\phi \rightarrow \bigcirc\bigcirc\phi)$. Let a_i be an agent and o an outcome. By $Choice_i^M(o)$ we mean a function which picks out the unique action $A_{im} \in Choice_i$, such that $o \in A_{im}$. The Chellas stit operator[2] has the following truth condition. (The following definitions and validities apply to any $a_i \in Agent$).

Definition 2.19. $M, o \vDash [a_i\ cstit]\phi$ iff. $Choice_i^M(o) \subseteq |\phi|_M$.

The **A** operator is a universal modality.

Definition 2.20. $M, o \vDash \mathbf{A}\phi$ iff. for any $o' \in dom(M)$, $M, o' \vDash \mathbf{A}\phi$.

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

Fact 2.21. $\vDash \mathbf{A}\phi \rightarrow [a_i\ cstit]\phi$

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

Definition 2.22. $M, o \vDash \bigodot[a_i\ cstit]_n\phi$ iff. for each $K \in Optimal_i^n$, $K \subseteq |\phi|_M$.

The intuition behind this operator is that if ϕ being true is a necessary condition for a_i to perform an optimal action given a level of trust n, then ϕ is obligatory for a_i. Setting $n = 1$, we get Horty's deontic ought to do operator as defined in (Horty; 2001, p. 78). We define the individual ought to do operator without subscript on $Optimus_i'$. We give the individual ought to do operator without subscript the following truth condition.

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

Definition 2.23. $M, o \vDash \bigodot[a_i\ cstit]\phi$ iff. for each $K \in Optimus'_i$, $K \subseteq |\phi|_M$.

The following facts contain some validities for these operators.

Fact 2.24. For any $n, m > 0$,

1. $\vDash \phi$ implies $\vDash \bigodot[a_i\ cstit]_n\phi$
2. $\vDash \bigodot[a_i\ cstit]_n(\phi \to \psi) \to (\bigodot[a_i\ cstit]_n\phi \to \bigodot[a_i\ cstit]_n\psi)$
3. $\vDash \neg(\bigodot[a_i\ cstit]_n\phi \wedge \bigodot[a_i\ cstit]_n\neg\phi)$
4. $\vDash \bigodot[a_i\ cstit]_n\phi \to \bigodot[a_i\ cstit]_m\phi$, for $n < m$.
5. $\vDash \bigodot[a_i\ cstit]_n\phi \to \bigodot[a_i\ cstit]\phi$
6. There is some m, such that for all $n \geq m \vDash \bigodot[a_i\ cstit]_n\phi \leftrightarrow \bigodot[a_i\ cstit]\phi$

Proof. 1. Assume $\vDash \phi$. Let $o \in dom(M) = W$ for some M. $|\phi|_M = W$. Let $K \in Optimal_i^n$. Since $K \subseteq W$, it follows that $K \subseteq |\phi|_M$, so $M, o \vDash \bigodot[a_i\ cstit]_n\phi$. Hence $\vDash \bigodot[a_i\ cstit]_n\phi$. 2. Assume $M, o \vDash \bigodot[a_i\ cstit]_n(\phi \to \psi)$ and $M, o \vDash \bigodot[a_i\ cstit]_n\phi$. Let $K \in Optimal_i^n$. Since $K \subseteq |\phi|_M$ and $K \subseteq |\phi \to \psi|_M = -|\phi|_M \cup |\psi|_M$, $K \subseteq |\psi|_M$. Hence, $M, o \vDash \bigodot[a_i\ cstit]_n\psi$. 3. Assume for the sake of a contradiction that there is some model M and some outcome o, such that $M, o \vDash \bigodot[a_i\ cstit]_n\phi \wedge \bigodot[a_i\ cstit]_n\neg\phi$. Hence for each $K \in Optimal_i^n$, $K \subseteq |\phi|_M$ and $K \subseteq |\neg\phi|_M$, so $K \subseteq (|\phi|_M \cap |\neg\phi|_M) = \emptyset$. Hence $K = \emptyset$, so $Optimal_i^n = \emptyset$, which contradicts Fact 2.14. 4. Assume $M, o \vDash \bigodot[a_i\ cstit]_n\phi$ and $n < m$. By Fact 2.15 $Optimal_i^m \subseteq Optimal_i^n$. Hence each for $K \in Optimal_i^m$, $K \subseteq |\phi|_M$, and hence $M, o \vDash \bigodot[a_i\ cstit]_m\phi$ 5. Assume $M, o \vDash \bigodot[a_i\ cstit]_n\phi$. Let $K \in Optimus'_i$. Since $Optimus'_i \subseteq Optimal_i^n$, $K \subseteq |\phi|_M$, and hence $M, o \vDash \bigodot[a_i\ cstit]\phi$. 6. Take m to be such that $Optimal^{M_m} = Optimus'$. For any n such that $n \geq m$, $Optimus'_i = Optimal_i^n$, so $\vDash \bigodot[a_i\ cstit]_n\phi \leftrightarrow \bigodot[a_i\ cstit]\phi$. □

Since the set of valid formulas is obviously closed under Modus Ponens (we have $\{\phi, \phi \to \psi\} \vDash \psi$), 1. (**Necessitation**) and 2. (**K**) show that $\bigodot[a_i\ cstit]_n$ is a normal modal operator. It is easily shown for the operator without subscript, $\bigodot[a_i\ cstit]$, as well. 3. (**D**) is the characteristic deontic formula, saying that if we are on one level of trust, there can be no moral conflicts. Again, it holds for $\bigodot[a_i\ cstit]$, also. 4. shows that there is also

consistency across levels of trust, in the sense that no obligation is lost when going to higher levels of trust. From these two validities, it follows that no obligation can be contradicted on a higher level of trust. 5. and 6. show some rather obvious interactions between the subscripted and non-subscripted operators. That any obligation is preserved by the non-subscripted operator, and that there is a finite level of trust from which adding more levels of trust is unnecessary, since it just gives the same obligations. In stit theory *ability* is expressed by $\mathbf{E}[a_i\ cstit]\phi$. One can think of this formula as expressing 'a_i has the choice to enforce ϕ.' We have the following important principle of *ought implies can*.

Fact 2.25. For any $n \vDash \bigodot[a_i\ cstit]_n\phi \rightarrow \mathbf{E}[a_i\ cstit]\phi$

Proof. Assume $M, o \vDash \bigodot[a_i\ cstit]_n\phi$. Then for each $K \in Optimal_i^n$, $K \subseteq |\phi|_M$. Since $Optimal_i^n \neq \emptyset$, let $K \in Optimal_i^n$ and let $o' \in K$. Now $Choice_i^M(o') = K \subseteq |\phi|_M$, hence $M, o' \vDash [a_i\ cstit]\phi$. Hence $M, o \vDash \mathbf{E}[a_i\ cstit]\phi$. □

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

Fact 2.26. For any level of trust m and any $o \in W$, $M, o \vDash \bigodot[a_i\ cstit]_m\phi$ iff $M \vDash \bigodot[a_i\ cstit]_m\phi$

Proof. Right to left is trivial. For left to right, assume $M, o \vDash \bigodot[a_i\ cstit]_m\phi$ and let $o' \in dom(M)$. Since, for each $K \in Optimal_i^m$, $K \subseteq |\phi|_M$, $M, o' \vDash \bigodot[a_i\ cstit]_m\phi$. □

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

2.7 Formalizing the examples

The first example is represented by Figure 2.2. The atomic formula R is true iff the Victim resists. The atomic formula H is true iff the Hero helps.

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

Fig. 2.2: Hostage situation

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

Fig. 2.3: The doctor's journey

Considered as a formal model, M, we have $M \not\models \bigodot[a_i\ cstit]_1 R$, with agents who only trust themselves, it is not the case that the Victim ought to resist. On the other hand we have $M \models \bigodot[a_i\ cstit]_2 R$, with agents trusting themselves and each other, the Victim ought to resist. In this case, since we are down to one action per agent, we have $M_2 = Optimus'$, so $M \models \bigodot[a_i\ cstit]_2 \phi \leftrightarrow \bigodot[a_i\ cstit]\phi$. Adding further levels of trust will not give us any more obligations. This example also shows that for some model, $M \not\models \bigodot[a_i\ cstit]\phi \rightarrow \bigodot[a_i\ cstit]_1 \phi$. The account thus generalizes Horty's individual ought to do, Horty (2001), which is our $\bigodot[a_i\ cstit]_1$, because more propositions may be obligatory on this account. It is of course a matter of context, whether agents are justified in trusting each other and the indexed operator gives us flexibility to meet different modeling needs in this respect. The second example is treated in a similar way. It is represented by Figure 2.3. B is a propositional atom meaning that the Doctor goes by boat. Here we have $dom(M_1) = \{o_1, \ldots, o_4\}$ (The Doctor staying home is dominated), and $dom(M_2) = \{o_1, o_3\}$ (The Guide staying home is dominated). We have $dom(M_3) = \{o_3\}$ (The Doctor walking is dominated). Since we are down to an atomic action profile, no further levels of trust will subtract more from the model. We thus have $M \models \bigodot[Doctor\ cstit]B$, the Doctor ought to go by boat.

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

Fig. 2.4: Friends meeting in town

2.8 The Meinong-Chisholm thesis

The Meinong-Chisholm thesis[3] is the following claim:

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

The Meinong- Chisholm thesis stands refuted with the theory presented here. There are cases where it ought to be that the agent sees to it that ϕ is still not equivalent to that the agent ought to do to it that ϕ, for instance, we might have $M, o \vDash \bigcirc[a_i \ cstit]\phi$, but not $M, o \vDash \odot[a_i \ cstit]\phi$. In certain situations, genuine group reasoning (individuals acts as parts of groups) can get us closer, e.g. in *hi-low* scenarios. Here is an example of such a scenario.

Example 2.3. Two friends, Friend 1 and Friend 2, (who cannot communicate beforehand) both face the choice of going to town to meet their buddy. It would yield the best outcome if they both went. However, going to town alone is futile and a big waste of energy. So the two outcomes, where one friend goes and the other stays home, are the worst. If both Friends stay home, it is better than if one goes in vain, but not as good as both meeting up in town.

Formally, the situation looks like in Figure 2.4, where G_1 is a propositional atom, which means 'Friend 1 goes to town.'

Even though this situation appears to have a similar structure to the first example, they are in fact essentially different. The difference is simply that none of the actions of either agent are strongly dominated. Therefore $M|optimal_M = M$. It follows that e.g. $M, o_1 \nvDash \odot[Friend_1 \ cstit]G_1$, we cannot say that Friend 1 ought to see to it that he goes. On the other hand going is a necessary condition for obtaining the best outcome in the situation,

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

so we have $M, o_1 \vDash \bigcirc[Friend_1\ cstit]G_1$. It ought to be that Friend 1 sees to it that he goes. One way of getting there is to extend the theory to also cover agents trusting in *groups*. This extension, which is postponed for further research, could be based on Horty's group ought operator, see Horty (2001). Even such an account, however, would not validate the Meinong-Chisholm thesis, since there are pure coordination situations, where the agents simply cannot know what they ought to do, whether they identify with a group or not. We can transform the situation above into a pure coordination situation, by assuming that the utility of both agents staying home is exactly the same as of meeting up in town. Here, conditional accounts of ought to do (given that Friend 1 stays, friend 2 ought to stay, etc), seem appropriate, but this too, is beyond the scope of this thesis. It is clear, though that such conditional oughts cannot tell agents what to do in a pure coordination situation *as a whole*, but only when fixing certain circumstances, which we take as the antecedent of the conditional ought.

Chapter 3

Responsibility formalized

3.1 Introduction

In this chapter I formalize concepts of responsibility within stit theory. The focus is on responsibility for *events*, formally considered as subsets of outcomes of situations. We distinguish two main kinds of responsibility, *positive responsibility* connected to the idea of *seeing to it that*, and *negative responsibility* connected to the idea of *being able to prevent*. First, we introduce the basic agentive concepts we need, *allowing, being able to, refraining, preventing*. Next, we introduce intentions into the situations. Finally we introduce the concepts of responsibility.

3.2 Allowing

There is an agentive, non-normative, non-evaluative way of using the word 'allow', a sense of simply letting something happen. In particular, *allowing* in this sense does not imply giving a permission. Consider the following examples:

Example 3.1.

1. 'Peter is allowing the window to be open.'

2. 'Anne allows her hair to grow long.'

In this thesis I will permit myself to paraphrase sentences like the ones above as follows:

1.' 'Peter allows it that the window is open.'

How can we formalize this agentive concept of *allows it that* within stit theory? Consider Figure 1.1 again. By making choice K_4, even if the best is true with the outcome (o_7 is the actual outcome), the agent a_1 did not see to it that the best happened, because there is another possible outcome of the same choice, (o_6), where the very worst happens. There are many situations like this, which involve choosing between taking a great risk and going for a safe bet. Now consider the situation where the worst actually does happen. We would here not say either that the agent saw to it that the worst happened. We would say, however, that the agent *allowed it to happen*, since the agent made the choice that made it possible. Similarly, if

the best happens it was also allowed by the agent, in the sense that the agent made the choice, which made it possible, or that it was a possible outcome of a choice made by the agent. Although the outcomes are assigned moral qualities (good, bad, and so on) in this example, these moral qualities are not essential for this definition of allowing. It works just as well in morally irrelevant examples such as the following. If an agent can choose to roll a die or not roll a die, and she rolls it and it shows two spots, the agent allowed it that the die shows 2 spots.

In general, an agent allows it that ϕ if and only if ϕ is true with a possible outcome of that agent's choice. It is easily seen that this suggestion corresponds exactly to interpreting the dual of the Chellas stit operator here written $\langle a_i\ cstit\rangle$ as *allows it that*. Since the dual is defined as $\neg[a_i\ cstit]\neg$ this operator gets the following truth condition.

Fact 3.1. $M, o \vDash \langle a_i\ cstit\rangle\phi$ iff. $Choice_i^M(o) \not\subseteq (dom(M) \setminus |\phi|_M)$.

In other words, $\langle a_i\ cstit\rangle\phi$ is true with an outcome o, iff there is some $o' \in Choice_i^M(o)$, such that $M, o' \vDash \phi$. Since the Chellas stit is an **S5** modality, **T**, $\phi \rightarrow \langle a_i\ cstit\rangle\phi$, is valid. Thus anything that happens to be the case with the actual outcome of any choice the agent allows to happen. Children are dying in Africa. I am allowing that to happen. Even things outside the visible universe, I have no chance of knowing about, I am allowing. Furthermore, since **4** is valid, we have $\langle a_i\ cstit\rangle\langle a_i\ cstit\rangle\phi \rightarrow \langle a_i\ cstit\rangle\phi$. Things that might have been happening with my choice but are not, I am also allowing. Before completely dismissing this as absurd, consider the following. There are cases where we are held responsible for events that might have obtained as a consequence of our choice, events we allowed to happen, even when they did not. For instance, agents might be criticized and even be penalized in a court of law for taking a risk, even if the actual outcome was not harmful. An everyday example is getting a penalty for driving while being intoxicated.

Still, the agents seem to have too little control over events they allow in this sense to hold them responsible for these events. In view of this we will not hold agents responsible just because they allow things in the very weak sense conveyed by the dual of the Chellas stit. It will take more than that to be responsible for an event.

We also note that an analogy to a traditional problem in the philosophy

of language arises, *Ross' paradox*, see Ross (1941).[1] The following inference is an example.

Example 3.2. 1. 'Anne allows her hair to grow long.'

2. 'Anne allows her hair to grow long or to burn.' (From 1.)

Finally, in this chapter I waver a bit between using present tense as opposed to past tense (the agent allows it that vs. the agent allowed it that). In view of the discussion of the different modalities in the introduction, it is most natural to use past tense when talking about responsibility, since this can only truly be assigned when the situation is settled in a certain outcome. In particular, a specific outcome is needed in order to determine whether an event actually obtained or just might have obtained.

3.3 Being able to, refraining and preventing

In Chapter 2, the Chellas stit was introduced. It is now time to introduce the deliberative stit operator $[a_i\ dstit]$. For this to be true an added *negative condition* must be fulfilled.

Definition 3.2. $M, o \vDash [a_i\ dstit]\phi$ iff. $Choice_i^M(o) \subseteq |\phi|_M$ and for some $o' \in dom(M)$, $M, o' \nvDash \phi$.

Intuitively, an agent can only truly be said to see to it that an event obtains, if the choice made by the agent is sufficient for ensuring that the event obtains *and* it is at least possible that this event did not obtain in the given situation. The ability concepts to follow will be defined for both the Chellas stit and the deliberative stit. Stit theory offers the following account of individual ability. The concept of an agent *a being able to* ϕ is defined with the formulas $\mathbf{E}[a_i\ dstit]\phi$ or $\mathbf{E}[a_i\ cstit]\phi$.[2] In the semantics the latter

[1] Ross formulated his paradox for imperatives , which he took to include what is now called deontic sentences. The example shows that there is a similar problem in the context of agentive sentences. In Belnap et al. (2001), it is argued that imperatives can be represented with stit sentences, but I do not wish to commit to that standpoint.

[2] In general, at least two concepts of ability can be distinguished, one agent based, related to what different people under normal circumstances are able to do (as in he is able to speak), and one situation based related to what can be done in a specific situation (as in he is able to speak right now). It is the situation based concept of ability that I am considering in this thesis. Similar remarks apply to the concept of being responsible below.

formula is true with an outcome, iff a has some choice where ϕ is true with all the outcomes of that choice. Informally, if *a can do* ϕ. In Chapter 4 I analyze this composite or double modality notion of ability in greater detail. Right now, it would disturb the flow of my argument too much to do so. We need the concept, however, in order to understand the definition of *refraining*. *Refraining* is defined as having the ability to see to something but allowing it not to happen, formally an agent a refrains from ϕ iff. $\mathbf{E}[a_i\ cstit]\phi \wedge \langle a_i\ cstit\rangle\neg\phi$ or with the deliberative stit $\mathbf{E}[a_i\ dstit]\phi \wedge \neg[a_i\ dstit]\phi$. Finally, we define an agent *a preventing* ϕ, simply as $[a_i\ cstit]\neg\phi$ or $[a_i\ dstit]\neg\phi$, the agent sees to it that ϕ does not happen.

3.4 Intentions in ethics and legal theory

Intentions are important in ethics and legal theory. Culpability or blameworthiness, for instance, is often defined in terms of the intentions of the moral or legal agent. We judge agents harder for doing things on purpose than by accident or even by not paying sufficient attention. How can we ever hope to represent e.g. legal responsibility without considering what the agents intended to do? However, so far intentions are not considered in stit theory. On the other hand intentions have been considered in BDI (Belief Intention Desire) logic, see Rao and Georgeff (1997) and Meyer and Veltman (2007) for references. Here the basic idea is that intentions function as a filter on desires and desires function as a filter on beliefs. In Meyer and Veltman (2007) a very simple multi-agent 'Belief-Intention' logic is suggested, where intentions are just serial relations on a set of possible worlds. This comes very close to what we will do here, except in the stit framework actions are given explicit treatment. A suggestion which is even closer in spirit to what we will do is the theory presented in Roy (2008). Roy represents intentions by sets of outcomes in a way very similar to what will be done here. The main difference comes from the fact that he applies his theory to problems in game theory. Thus he is interested in connecting intentions to an agent's utility functions and to solution concepts in game theory. Since I mainly consider cases where an agent's values diverge from e.g. legal or moral values and we only represent the legal values, these sort of connections are not very relevant. The main intuition underlying my account is that each action token has a non-empty subset of *intended outcomes*, which is only relative to this particular action and not to the overall situation.

3.5 Intentions in situations

Intentions are represented indirectly by means of the outcomes. It will be assumed that for any action A there is a non-empty subset I_A of *intended outcomes of* A. Thus intended outcomes are relative to an action. If an agent picks up an apple, the intended outcomes are ones where she has picked up the apple. If an agent closes the door, the intended outcomes are the ones where the door is closed. Thus the focus is not on what the agent finally decides to do given a situation, but on what the agent intends to happen given some choice in a situation. The set of intended outcomes may be seen to represent the purpose of that particular action, seen from the perspective of the agent.

Definition 3.3. For each $a_i \in Agent$, and each action $A_{ij} \in Choice_i$, there is a non-empty subset $I_{A_{ij}} \subseteq A_{ij}$, called the set of intended outcomes of A_{ij}

We define a *utilitarian strategic model with intentions* as a utilitarian strategic model with a set of intended outcomes for each action of each agent as defined above. We introduce an intention operator $[a_i\ iit]$ for each agent. The informal reading of $[a_i\ iit]\phi$ is 'a_i intends it that ϕ.' The set of well formed formulas is suitably extended and the truth condition for this operator is as follows.

Definition 3.4. $M, o \vDash [a_i\ iit]\phi$ iff. $I_{Choice_i^M(o)} \subseteq |\phi|_M$.

Informally, an agent intends it that ϕ with an outcome o, if ϕ is true with all intended outcomes of the choice made at o, $Choice_i^M(o)$. The 'intends it that' operator is a normal **D45** operator, which validates **K**, necessitation and the following.

Fact 3.5. 1. $\neg([a_i\ iit]\phi \wedge [a_i\ iit]\neg\phi)$

2. $[a_i\ iit]\phi \rightarrow [a_i\ iit][a_i\ iit]\phi$

3. $\neg[a_i\ iit]\phi \rightarrow [a_i\ iit]\neg[a_i\ iit]\phi$

4. $[a_i\ cstit]\phi \rightarrow [a_i\ iit]\phi$

1. An agent cannot intend ϕ and its negation. 2. If an agent intends it that ϕ, then the agent intends it that the agent intends it that ϕ. 3. If an agent does not intend it that ϕ, then the agent intends it that the agent does

not intend it that ϕ. 4. If an agent sees to it that ϕ with the Chellas stit, then the agent intends it that ϕ. From 4. it follows that $[a_i\ dstit]\phi \rightarrow [a_i\ iit]\phi$. Although an agent cannot see to it that ϕ without intending it that ϕ, it is possible for an agent to allow it that ϕ and intend it that $\neg\phi$. Should intentions be closed under logical consequence? Horty does not seem to think so.

> ... one could imagine that an agent might see to it that A holds and that B holds as well without intentionally seeing to it that they hold jointly...
> (Horty; 2001, p. 17)

With the present proposal, this is clearly not possible. We have $([a_i\ iit]\phi \wedge [a_i\ iit]\psi) \rightarrow [a_i\ iit]\phi \wedge \psi$ and $[a_i\ dstit]\phi \rightarrow [a_i\ iit]\phi$. Therefore $[a_i\ dstit]\phi$ and $[a_i\ dstit]\psi$ clearly contradicts $\neg[a_i\ iit]\phi \wedge \psi$. In general, this question opens a number of possible objections related similar problems in deontic and epistemic contexts. Suffice it to give two examples.

Donald Davidson considers a case, where an agent has to decide to flip a light switch and turn on the light. Unbeknownst to the agent, a prowler is waiting outside. The outcomes where the light is turned on are exactly the same as the ones where the prowler is alerted. Yet, we do not wish to say that the agent intended to alert the prowler. (Davidson; 1963, p. 24). At first, it may seem that Davidson's problem is not really relevant to the situations represented in this thesis. Since we only consider situations where every agent knows every possible outcome, the agent *does know* that turning on the light is equal to alerting the prowler (with the terminology used later in the chapter the validity in the model of $L \leftrightarrow P$ implies the validity in the model of $\mathbf{K}_i^{pre}(L \leftrightarrow P)$). For this reason, we might say Davidson's example lies beyond the scope of this thesis. However, the problem is much more general and not essentially connected to the epistemic conditions of situations. It is related to what is called *double effects* in Ethics and Law, see e.g. Duff (1982), Chisholm (1970), Greve (2004). Although it is inevitably painful to get an injection, the doctor does not intend to cause me pain. Technically, double effects concern the counterfactual nature of intentions. One way of seeing this problem, is that a formula expressing a set of intended outcomes seems to be more stable under some counterfactual circumstances than the contingent logical equivalences given by the actual situation. The example

given by Roy, see (Roy; 2008, p. 27), taken from Bratman, concern two different decision makers who are going to bomb a munitions plant placed next to a school. We assume that the decision makers are completely aware that they cannot bomb the munitions plant without bombing the school and vice versa. Yet, there is a difference between *terror bomber*, who primarily wants to bomb the school and *strategic bomber*, who wants to bomb the munitions plant. In particular, strategic bomber can claim that he does not intend to bomb the school with the following argument: had the school been placed elsewhere, he would still intend to bomb the munitions plant. Roy's suggested way of getting out this problem involves redescribing the situation to bring out the counterfactual aspects, however, he does not pursue a theory of such alternative descriptions systematically, (the problem is not central to his thesis) and I shall also not pursue this strategy further here. On the one hand, this leaves us with the following immediate problem: the representation of intentions via intended outcomes makes our concept of intention diverge some from the concept as used in natural language. On the other hand, the fact that intentions are closed under logical consequence, does not seem any more or less serious than other propositional attitudes to propositions. For instance, in possible worlds semantics:

> ... we seem to predict wrongly that a person who believes a proposition p should also believe any proposition that is true in the same worlds as p. (Kratzer; 2007, p. 4-5)

Perhaps some of the various ways of dealing with this problem for other attitudes could be adapted for intentions. For instance, we might distinguish between what an agent implicitly intends and explicitly intends in a way similar to implicit vs. explicit knowledge, see Fagin et al. (1995). A specific suggestion quite similar to that will be made in Chapter 8. An alternative more in tune with the counterfactual understanding of double effects would be to extend the models to sets of possible situations with a suitable correspondence between outcomes of these. The set of intended outcomes of a situation would then be defined across a subset of these possible situations. This suggestion comes very close to the way the problem is dealt with in the seminal paper on BDI logic, see Rao and Georgeff (1997). For now, however, I leave this problem and concentrate on defining concepts of responsibility using the simple concept of intentions presented above.

3.6 Responsibility

By *positive responsibility* we mean the kind of responsibility an agent has for whatever she deliberatively sees to. So we define it as follows.

Definition 3.6. An agent a_i is positively responsible for ϕ at o, iff $M, o \vDash [a_i\ dstit]\phi$

Thus, formally there is no difference between being positively responsible for ϕ and seeing to it that ϕ. By *negative responsibility* we mean the kind of responsibility an agent has for whatever she could have prevented. There are two subcases of negative responsibility, *strict liability* and *liability for risking. Strict liability* is defined as follows.

Definition 3.7. An agent a_i is strictly liable for ϕ at o, iff. $M, o \vDash \phi \wedge \mathbf{E}[a_i\ dstit]\neg\phi$

An agent is *strictly liable* for ϕ iff. ϕ happened and the agent could have prevented ϕ by making another choice. In determining strict liability, the intentions or *mens rea* of the agent does not matter. This is typically the case in situations of omissions or neglect. An agent did not intend for ϕ, but she did not do what was required to prevent it. A related concept is that of *being liable for risking* ϕ. This is defined as follows.

Definition 3.8. An agent a_i is liable for risking ϕ at o, iff. $M, o \vDash \neg\phi \wedge \langle a_i\ cstit\rangle\phi \wedge \mathbf{E}[a_i\ dstit]\neg\phi$

Thus an agent is liable for risking ϕ, if ϕ did not actually happen, but it could have happened with the choice made by the agent, and the agent could have prevented ϕ by making another choice.

3.7 Guilt

In determining the *guilt* of an agent, the intentions of the agents are important. We say that an agent is *guilty* of ϕ iff she is responsible for ϕ and intended it that ϕ. If an agent is positively responsible for ϕ, she is automatically guilty of ϕ, as established by the validity $[a_i\ dstit]\phi \rightarrow [a_i\ iit]\phi$. However, we also say that an agent a_i is guilty of ϕ, if a_i is strictly liable for ϕ and a_i intended it that ϕ.

Definition 3.9. An agent a_i is *guilty of* ϕ at o, iff either

1. a_i is positively responsible for ϕ at o, or

2. $M, o \vDash \phi \wedge \mathbf{E}[a_i\ dstit]\neg\phi \wedge [a_i\ iit]\phi$

Thus the second sufficient condition of *guilt* is defined by adding the intention on top of *strict liability*. When we want to distinguish the two kinds of guilt we call the latter *negative guilt*. We define an agent as *guilty of attempt at* ϕ if the agent is *liable for risking* ϕ and the agent intended it that ϕ.

Definition 3.10. An agent a_i is *guilty of attempting* ϕ at o, iff. $M, o \vDash \neg\phi \wedge \langle a_i\ cstit\rangle\phi \wedge \mathbf{E}[a_i\ dstit]\neg\phi \wedge [a_i\ iit]\phi$

The concept of negative guilt and the concept of guilt of attempt are especially relevant in cases where the agent cannot see to it or enforce an event, but where the final outcome of the situation depends on luck or favorable circumstances. For instance, an assassin might not be able to see to it that she hits and kills a politician, but we will still consider her guilty of murder, if she kills the politician, intended to kill him and could have refrained from killing him. We will consider her guilty of attempted murder, if the same obtains, except that she misses her intended target.

We define the abilitive concept *regulative control* over ϕ as the simultaneous ability to see to it that ϕ and to prevent ϕ in a situation. Thus an agent is both negatively responsible and positive responsible for ϕ, if she has regulative control. A lot of the everyday events we are responsible for are of this kind, we can ensure them or prevent them at will, especially events closely connected to the control we have over own bodies. The concept of guilt as defined above is not evaluative, being guilty of ϕ does not imply that ϕ is good or bad, right or wrong. To get to the concepts of blameworthiness and praiseworthiness, we use the deontic operators defined in Chapter 2.

3.8 Moral blameworthiness and praiseworthiness

In order to be morally or legally blameworthy for ϕ is it not enough to be guilty in the sense defined above. It must also be wrong to do ϕ. Likewise there is a deontic condition for praiseworthiness. We now define concepts of moral blameworthiness, in terms of ought to do operators from the previous chapter. We say that an agent is blameworthy of ϕ, if the agent is guilty of ϕ, and ϕ ought to be prevented.

Definition 3.11. An agent a_i is *blameworthy of* ϕ at an outcome o, iff a_i is guilty of ϕ at o and $M, o \vDash \bigodot[a_i\ cstit]\neg\phi$.

This concept of blameworthiness does not distinguish between guilt because of positive responsibility and guilt because of negative responsibility. However, in the following definitions of blameworthiness of attempting, blameworthiness by neglect and blameworthiness of risking only negative responsibility comes into play. Intuitively, an agent deserves blame for attempting ϕ, if the agent ought to have prevented ϕ , ϕ did not actually happen, but the agent intended it that ϕ.

Definition 3.12. An agent a_i is *blameworthy of attempting* ϕ at o, iff a_i is guilty of attempting ϕ at o and $M, o \vDash \bigodot[a_i\ cstit]\neg\phi$.

Intuitively, an agent is deserves blame of ϕ because of neglect, if ϕ happened, the agent did not intend ϕ to happen, the agent could have prevented ϕ and the agent ought to have prevented ϕ.

Definition 3.13. An agent a_i is *blameworthy of* ϕ *because of neglect* at o, iff a_i is strictly liable for ϕ at o and $M, o \vDash \neg[a_i\ iit]\phi \wedge \bigodot[a_i\ cstit]\neg\phi$.

An agent is blameworthy of risking ϕ, if ϕ did not happen but might have happened with the choice of the agent, the agent did not intend it that ϕ, the agent could have prevented ϕ and the the agent ought to have prevented ϕ.

Definition 3.14. An agent a_i is *blameworthy of risking* ϕ at o, iff a_i is liable for risking ϕ at o and $M, o \vDash \neg[a_i\ iit]\phi \wedge \bigodot[a_i\ cstit]\neg\phi$.

If we interpret $\bigodot[a_i\ cstit]\neg\phi$ as 'ϕ is illegal' the defined concepts correspond to legal concepts. Blameworthiness corresponds to legal culpability of an event and blameworthiness of attempting an event corresponds to legal culpability of attempt. Blameworthiness by neglect and blameworthiness of risking corresponds to legal culpability of an event and of risking an event due to an omission.

In the simplest case assigning praise is symmetrical to assigning blame. An agent deserves praise for ϕ if the agent sees to it that ϕ (the first sufficient condition for being guilty) and the agent ought to do ϕ.

Definition 3.15. An agent a_i is positively praiseworthy of ϕ at an outcome o, iff $M, o \vDash [a_i\ cstit]\phi \wedge \bigodot[a_i\ cstit]\phi$.

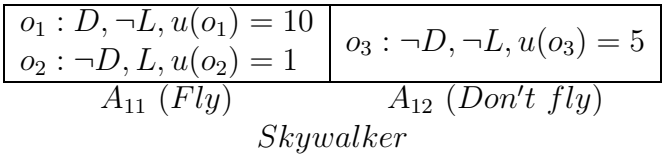

Fig. 3.1: Skywalker's choice

However, things can become more interesting, when we start thinking about negative responsibility and praise. There are situations, where no action is dominated, and yet an agent can deserve praise for doing or attempting something. Consider the following example.

Example 3.3. In *Star Wars Episode IV: A New Hope* (1977) Luke Skywalker has to make a very unlikely shot from his space ship to destroy the Death Star. If he had decided not to fly this dangerous mission, he would not have been to blame (we may assume). However, if he decides to fly it and succeeds he deserves praise. If he gets killed, he even deserves praise for trying.

This example is represented by Figure 3.1. Here D means that the Death Star is destroyed, and L means that Luke dies. We assume that the best possible outcome is when the Death Star is destroyed and Luke survives, the worst is when the Death Star is not destroyed and Luke dies and that the outcome where Luke survives but the Death Star is not destroyed has an intermediate value. In this case none of Luke's actions dominates the other. In particular, it is not the case that Skywalker ought to see to it that the Death Star is destroyed, since ought implies can and Skywalker cannot guarantee a hit. On the other hand, it ought to be that the Death Star is destroyed, since D is true with the best possible outcome of the situation. Since Skywalker's intention by flying is to destroy the Death Star, we have $I_{A_{11}} = \{o_1\}$. We have $I_{A_{12}} = \{o_3\}$. We have that Skywalker is guilty of destroying the Death Star at o_1, and guilty of attempting to destroy the Death Star at o_2. We say that an agent a_i is negatively praiseworthy for doing ϕ, if a_i is negatively guilty of ϕ and it ought to be that ϕ.

Definition 3.16. An agent a_i is negatively praiseworthy for doing ϕ at an outcome o, iff. a_i is negatively guilty of ϕ at o, and $M, o \vDash \bigcirc\phi$.

We say that an agent a_i is praiseworthy of trying to do ϕ, if a_i is guilty of attempting ϕ and it ought to be that ϕ.

Definition 3.17. An agent a_i is praiseworthy of attempting ϕ at an outcome o, iff. a_i is guilty of attempting ϕ at o, and $M, o \vDash \bigcirc\phi$.

In some cases, it is considered an omission if an agent does not make an attempt. In cases where somebody might risk ruining his own or somebody else's property in order to save another human being (e.g. throwing a valuable floating object into a lake to try to save a person from drowning) without putting their own life at risk, they might be legally responsible for neglect if they do not do anything.[3] The structure of such cases are the same as in the Skywalker example. Here the worst outcome is when the object is destroyed and the person is not saved from drowning, the intermediate outcome is when the person is not saved and the object is not destroyed and the best outcome is when the person is saved from drowning and the object is destroyed.

The most important precursor of this work is Stig Kanger's definition of individual responsibility. I would like to conclude this chapter by comparing the definitions given above with Kanger's. However, Kanger's definition explicitly refers to the knowledge of agents. Thus, in order to make a full comparison, it is necessary to spell out the epistemic conditions of agents.

3.9 Knowledge in situations

The primary aim of this thesis is to analyze concepts related to actions, ability and norms. The following discussion of epistemic notions will be kept brief.

In accordance with the three temporal views on situations discussed in the introduction, three kinds of knowledge may be distinguished. It is possible to distinguish between an agent's knowledge *before, during*, and *after* a given situation. Rather than explicitly introducing a temporal structure to situations (as would be the common way to go in stit theory), the strategy here is a little different. Each agent a_i will be given 3 partitions representing her knowledge before, during, and after a situation. These will be denoted $\sim_i^{pre}$, $\sim_i^{dur}$, $\sim_i^{post}$. So far (and also henceforth), the different outcomes of situations are meant to represent ontological uncertainty. Now, we can *also* talk about the epistemic uncertainty of agents. There are several options when deciding how to define these epistemic accessibility relations, but the conditions stipulated as follows are probably the simplest. They utilize already

[3] This is for instance the case according to Danish law, see (Greve; 2004, p. 131).

used partitions. Before the situation the epistemic uncertainty is total i.e. it is exactly the same as ontological uncertainty. For a situation M and for each agent a_i, $\sim_i^{pre} = dom(M) \times dom(M)$. During the situation the agent knows what choice she has made but not what choices the other agents have made, i.e. $\sim_i^{dur} = Choice_i$ (the partition given by the actions of the agent). After the situation there is no epistemic uncertainty, i.e. the epistemic relation is the identity relation on the domain. For any $o, o' \in dom(M)$, $o \sim_i^{post} o'$ iff. $o = o'$. To the symbols are added three epistemic operators for each agent a_i, written $\mathbf{K}_i^{pre}$, $\mathbf{K}_i^{dur}$, $\mathbf{K}_i^{post}$, and the set of well-formed formulas is extended as expected. We define a *utilitarian strategic models with intentions and knowledge* as a utilitarian strategic model with three epistemic relations for each agent as defined above. The truth conditions of the epistemic operators are defined as usual.

Definition 3.18. 1. $M, o \vDash \mathbf{K}_i^{pre}\phi$ iff for each o', such that $o \sim_i^{pre} o'$, $M, o' \vDash \phi$.

2. $M, o \vDash \mathbf{K}_i^{dur}\phi$ iff for each o', such that $o \sim_i^{dur} o'$, $M, o' \vDash \phi$.

3. $M, o \vDash \mathbf{K}_i^{post}\phi$ iff for each o', such that $o \sim_i^{post} o'$, $M, o' \vDash \phi$.

The following validities give the flavor of these epistemic conditions. In general, before the situation, agents have knowledge of the modalities which are settled in the situation, i.e. the deontic modalities and the ability modalities. During the situation, agents also have knowledge about the action modalities, i.e. the stit and iit modalities, and after the situation the agents also have knowledge about everything, including the various concepts of responsibility.

Fact 3.19. 1. All operators are normal **S5** modalities.

2. $\vDash \mathbf{K}_i^{pre}\phi \rightarrow \mathbf{K}_i^{dur}\phi \rightarrow \mathbf{K}_i^{post}\phi$ (Knowledge is not forgotten).

3. $\vDash \bigodot[a_i\ cstit]\phi \rightarrow \mathbf{K}_i^{pre} \bigodot[a_i\ cstit]\phi$.

4. $\vDash \mathbf{E}[a_i\ dstit]\phi \rightarrow \mathbf{K}_i^{pre}\mathbf{E}[a_i\ dstit]\phi$

5. $\vDash [a_i\ dstit]\phi \rightarrow \mathbf{K}_i^{dur}[a\ dstit]\phi$.

6. $\vDash [a\ iit]\phi \rightarrow \mathbf{K}_i^{dur}[a\ iit]\phi$.

7. Let ϕ be a formula expressing one of the formulas of responsibility above for the agent a. $\vDash \phi \rightarrow K_i^{post}\phi$.

Proof. I sketch a proof of 1.,2., and 7. 1. follows from the fact that the three relations are equivalence relations. 2. follows from the fact that the identity relation is a sub relation of the relation given by the partition of the actions, and this is a sub relation of the universal relation. 7. follows from the fact that an agent has total knowledge of any formula ϕ after the situation. Let o be an outcome and ϕ be any formula , s.t. $M, o \vDash \phi$. Since o is identical to o and only identical to o, we have $M, o \vDash \mathbf{K}_i^{post}\phi$. □

More interesting concepts of knowledge could be introduced, for instance, the uncertainty after the situation can be increased by loosening the restriction that $\sim_i^{post}$ are the singleton sets. This uncertainty is purely epistemic, not ontological. I leave the evaluation of these options to somebody else. The main purpose of introducing these epistemic concepts here was to facilitate a comparison with Kanger's interesting concepts of responsibility.

3.9.1 Kanger on responsibility

Kanger's definition of individual responsibility was presented in Kanger (1971). His definition of a_i being responsible for ϕ is that a_i is either blameworthy or praiseworthy for ϕ. Thus he defines responsibility as always already containing an evaluative component. This is in opposition to our strategy, which assumes that there are both agentive and intentional concepts of responsibility (and guilt), which are morally or legally neutral. It seems that agents are responsible for quite a few things that are neither forbidden or obligatory, for example, I am responsible for writing this chapter right now. Blame and praise are evaluations based on an added normative component. Here this added normative component is the utilities of outcomes. In the following I transpose Kanger's definition of being blameworthy into the models presented in this thesis. Apart from agentive and alethic concepts it involves deontic and epistemic concepts. In the following I use the Chellas stit.

Definition 3.20 (Kanger)**.** An agent a_i is blameworthy for ϕ if,

1. $\bigcirc\neg\phi$,
2. $[a_i\ cstit]\phi$,

3. $\neg\mathbf{A}[a_i\ cstit]\phi$,
4. $\mathbf{EK}_i \bigcirc \neg\phi$
5. $\mathbf{EK}_i[a_i\ cstit]\phi$
6. $\mathbf{EK}_i\neg\mathbf{A}[a_i\ cstit]\phi$

The first three parts of the definition consist of a deontic condition, an action condition and an ability condition. An agent a_i is blameworthy for ϕ, if ϕ is wrong, a_i has done ϕ, and it is not the case that a_i cannot avoid doing ϕ, (equivalent in our system to it is possible for a_i to allow $\neg\phi$). The last three parts of the definition consist of a sort of 'epistemic closure' of the first three parts. It is required that it is possible that a_i knows that it is wrong, it is possible that a_i knows what he did, and it is possible that he knows, he did not have to do it. To be praiseworthy replace $\bigcirc\neg\phi$ with $\bigcirc\phi$ in parts 1 and 4 of the definition. Now for the comparison with the theory presented in this chapter. Part 2 and part 3 of the definition together imply the dstit operator (if we had used the dstit operator to begin with the third clause would be superfluous, since it is already a tautology). Kanger does not make the distinction between *ought to be* and *ought to do* observed in this thesis. Let us restrict our attention to *ought to do*. Thus, we replace Kanger's first condition with $\odot[a_i\ cstit]\neg\phi$. After that the three first parts corresponds to the definition given of blameworthiness in Definition 3.11, leaving out the intentions. It is not obvious from Kanger's paper whether we should consider the knowledge of the agent before during or after a situation. However, it seems natural to require that the agent knows before the situation that she ought not to do ϕ. Further, she also knows beforehand that there are other choices available to her where she does not see to it that ϕ. Finally, she should know what she is doing during the situation. In fact, the fulfilment of these three requirements follows from the first three parts of the definition with the way knowledge is represented here. The fourth part follows from the validity $\odot[a_i\ cstit]\neg\phi \rightarrow \mathbf{K}_i^{pre}\odot[a_i\ cstit]\neg\phi$. The fifth part follows from the validity $[a_i\ cstit]\phi \rightarrow \mathbf{K}_i^{dur}[a_i\ cstit]\phi$. The sixth part follows from the validity $\mathbf{E}\langle a_i\ cstit\rangle\neg\phi \rightarrow \mathbf{K}_i^{pre}\mathbf{E}\langle a_i\ cstit\rangle\neg\phi$. Thus, when restricting ourselves to knowledge as presented above, Kanger's epistemic closure conditions are already met, or actually something stronger since we do not have to add the possibility operators in front of the epistemic operators. These outermost

possibility operators on the conditions follow from the validity $\phi \rightarrow \mathbf{E}\phi$. It is possible that Kanger by adding these possibility operator wanted to add epistemic uncertainty for the observer of the situation i.e. it is possible to the observer that the agent had this knowledge and thus the observer is justified in *assigning* blame or praise. Here, the focus has been not on epistemic justifications of assigning blame but on *being* blameworthy and praiseworthy in an 'objective' sense. The reason for the quotes around objective is that all is relative to a given representation of a given situation and a certain set of values. When that is assumed, I think that the defined concepts capture quite well an intuitive sense of objective fairness, also in the sense of giving sufficient and necessary reasons for being justified in assigning blame according to a certain set of values. If all the conditions for an agent to be blameworthy obtain, then it is fair to assign blame to him. If one of the conditions does not obtain then it is not fair to assign blame to him. For instance, if ϕ did not happen but might have happened with the actual choice of an agent, the agent could and ought to have prevented ϕ and the agent did not intend it that ϕ, it is not fair to hold that agent responsible for attempting ϕ, although it is fair to hold the agent responsible for risking ϕ. In other words, the theory reinstates a fruitful distinction between *being* responsible (guilty, etc.) and merely being *assigned* responsibility.

Chapter 4

Ability modalities and the metaphysics of agency

4.1 Introduction

This chapter contains a systematic analysis of ability modalities, also called *dynamic* modalities. The distinctions provided by this conceptual analysis are applied in a metaphysical discussion of ability. The focus of this chapter is the ability modality *can* and some related ability modalities. Thus I do not consider e.g. deontic and quantificational *can*(as in 'screwdrivers can be dangerous'). A distinction between agent capability *can* (general ability of an agent, as in 'she can swim') and situation ability *can* (the ability of a specific agent in a particular situation) is also relevant to this analysis. Vaguely stated, the connection between the two is that with the preconditions for a certain event in place, agent capability normally implies situation ability. E.g., An agent may be capable of swimming in general, but since a normal precondition for swimming is the presence of a body of water of some size, she might not have situation ability to swim in a given situation. This distinction is not the focus of this chapter, where the aim is exclusively to analyze situation ability, concerning what particular agents can (and may, might or must) do in specific situations.

4.2 The Brown-Horty double modality analysis of ability

In this section I summarize the discussion leading to the Brown-Horty definition of ability following Brown (1988), Brown (1992), (Horty; 2001, pp. 2-24), see also (Portner; 2009, Chapter 4). The double modality definition of ability given by Brown and Horty provides the starting point for the more general and systematic analysis of ability modalities within the stit framework presented here. Kenny argued that a normal modal logic possibility operator cannot capture a natural concept of ability, see Kenny (1976). This is so because the validities $\phi \rightarrow \Diamond\phi$ and $\Diamond(\phi \vee \psi) \rightarrow (\Diamond\phi \vee \Diamond\psi)$ contradict our intuitions about this concept. From the fact that an agent is lucky enough to hit the bull's-eye it does not follow that she has an ability to do it. From the fact that an agent is able to hit the top half or the bottom half of the dart board it does not follow that she has the ability to hit the bottom half or she has the ability to hit the top half. The latter might require more control over the dart than she has. The first objection can be met by not requiring that

the accessibility relation for the modality is reflexive. The second objection applies to all normal modal logics (for a definition of a normal modal logic, see e.g. Chellas (1980)). As a consequence, Brown responded to Kenny by a move to non-normal modal logics. He introduced an ability operator in possible worlds models with the following semantics: A set of propositions are singled out as relevant - they represent actions. The worlds that are elements of each action/set represent outcomes of that action. We can now both quantify over actions and over outcomes. The ability operator corresponds to existential quantification our actions and universal quantification over outcomes : For some action all outcomes make ϕ true. The dual of this gives us universal quantification over actions and existential quantification over outcomes. This Brown refers to as a concept of *might*: for any action, some outcome makes ϕ true. An even weaker concept of *might* is existential quantification over actions and existential quantification over outcomes: with some outcome of some action, ϕ is true. The strongest concept is universal quantification over actions and universal quantification over outcomes: With all outcomes of any action, ϕ is true. This Brown calls a concept of *will*.

In Brown (1992) the analysis of ability is recaptured with two normal modalities one alethic possibility operator and one action or *brings it about that* operator. The other ability modalities (will, might) are not investigated in that paper. It is further investigated (following a suggestion made to Brown by David Lewis) that the action operator could be an S5 operator. Horty captures Brown's ability operator in the specific stit setting by the composite formula $\mathbf{E}[a_i\ cstit]\phi$.[1] The most rigid translation into natural language of this Brown-Horty double modality definition of ability is 'it is possible that the agent sees to it that ϕ.' However, given that the choices partition the outcomes, and so any outcome is a member of some choice, it is also fair to translate this: 'the agent has the choice to see to it that ϕ.'. This comes fairly close to 'the agent is able to ϕ.' from this there is not far to 'the agent can ϕ.' At least, if we restrict our attention to what an agent can do in a specific situation. In the following we will transpose the rest of Brown's ability modalities to the stit setting using the universal modality and the 'sees to it that' and 'allows it that' modalities.

[1] A minor difference between Horty's and Brown's analysis is that Brown considers the alethic possibility relation to be only reflexive, whereas Horty considers it the universal (an equivalence) relation.

ability modal	Formula	Rigid translation
The agent must ϕ	$\mathbf{A}[a_i\ cstit]\phi$	It is necessary that the agent sees to it that ϕ
The agent can ϕ	$\mathbf{E}[a_i\ cstit]\phi$	It is possible that the agent sees to it that
The agent may ϕ	$\mathbf{A}\langle a_i\ cstit\rangle\phi$	It is necessary that the agent allows it that ϕ
The agent might ϕ	$\mathbf{E}\langle a_i\ cstit\rangle\phi$	It is possible that the agent allows it that ϕ

Fig. 4.1: Ability modalities

4.3 Ability *must, can, may* and *might*

The following four different concepts connected to ability will be considered: *must*, *can*, *may*, *might*. They are presented in Figure 4.1.

There is no doubt that the rigid translations of the formulas are fine, given that the definitions of *seeing to it that* and *allowing it that* are fine. There is also no doubt that the ability modalities occur in natural language. The strongest of these I call ability *must*. This name replaces Brown's *will*. *Will* seems to be more connected to intentions than to ability. However, this is not an important matter, the semantics is the same as the one suggested for *will* by Brown. Although ability *must* is rare, it does occur, e.g. in the question 'must you breathe so loud?' Instead of using the word *might* in two different ways, I introduce *may* for the concept which is the dual of *can* and reserve *might* for the dual of *must*. For *can* and *may* the following dialogue will serve as an example.

Example 4.1.

Hector: 'Bob can run the ten kilometers.'

Andrea: 'No, he is not in very good shape, so he may stop after five kilometers.'

Here, Andrea takes Hector to mean that Bob has the ability to ensure that he runs the distance. She contradicts this by claiming that he does not have this ability in view of his poor physical shape. Are these statements really contradictory? I shall argue that they are, given the following proviso

regarding the distinction between *may* and *might*. The distinction that I have made between *may* and *might* is clearly artificial and does not occur in natural language. The underlying conceptual difference underlying this distinction was observed by Brown and it is obviously relevant. With the present theory, 'He may ϕ', means that no matter what choice he makes (how or what he does), ϕ is a possible consequence. We may assume this is what Andrea meant by *may* in the previous example, although she easily could have used *might* to get at the same meaning. With our theory, 'He might ϕ', refers to Brown's weaker *might*. It means that with one of his choices, ϕ is a possible consequence. The meaning of this latter is perhaps captured in a dialogue like the following.

Example 4.2.

Hector: 'Should I throw the dart?'

Andrea: 'Yes, you might hit the bull's eye.'

The meaning of the latter 'might' seems to be that given that Hector chooses to throw the dart there is a possibility that he hits the bull's eye. This is what we shall mean by 'might' in this chapter. I don't consider it a very serious objection that the distinction between ability may and ability might is not present in natural language, since my prior concern is not to make a theory of natural language, but rather to analyze interesting modal concepts within an already given theory of agency. I consider it sufficient reason to fix the distinction that these different concepts seem to underlie different uses of the ability modalities in natural language. Another more serious objection will be discussed towards the end of the chapter.

4.4 Logical relations between ability modalities

It is well known that for any modal logic with at least a serial accessability relation (i.e. also including reflexive relations as for the stit modality and the universal modality) the traditional square of opposition holds, see e.g Fitting and Mendelsohn (1998). For the *sees to it that* and *allows it that* operators, it looks as in Figure 4.2. Two formulas are contradictory iff they cannot both be true and they cannot both be false. Two formulas are contraries iff they cannot both be true. Two formulas are subcontraries iff they cannot both be false. A formula is a subaltern of another iff it is implied by it but not the other way around. Since there are 8 ability modalities, it is not possible to

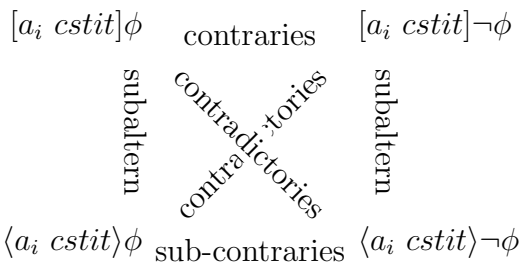

Fig. 4.2: Square of opposition for action modalities

make a square of opposition for these. For a combination of two modalities, a *cube* of opposition can be constructed, as in Figure 4.3 (to make simpler, explicit reference to the agent is omitted.) I will just give one example of how to test these logical relations. Assume, $M, o \vDash \mathbf{A}[a_i\ cstit]\phi \wedge \mathbf{E}[a_i\ cstit]\neg\phi$, (i.e. the agent a_i must ϕ and can $\neg\phi$.) From the second conjunct, there is an outcome o', such that $M, o' \vDash [a_i\ cstit]\neg\phi$, hence $M, o' \vDash \neg\phi$ From the first conjunct, $M, o' \vDash [a_i\ cstit]\phi$, hence $M, o' \vDash \phi$, contradiction.

The corner that cannot be seen contains **may**$\neg\phi$. Imagine the cube tilted forward 45 degrees so that the corners with **must**ϕ and **must**$\neg\phi$ are on top. The diagonal planes are hidden in the figure. The diagonal plane constituted by **must**, **might** and their negations constitutes a regular square of opposition **must**ϕ and **might**$\neg\phi$ are contradictories and so on. The other diagonal plane, constituted by **can**ϕ, **may**ϕ and their negations, is extracted in Figure 4.4.

The only logical relation from the traditional square of opposition that holds between these duals and their negations is that a formula is the contradictory of its dual with the complement negated. If the two top formulas are both true, then, of course, the two bottom ones are both false. In this case the agent has what we called *regulative control* over ϕ in the previous chapter. If the two bottom formulas are both false, ϕ is what we may call *totally contingent* for the agent, which means that it is completely out of control of the agent. No matter what she does, it is ϕ is true with some outcome and it is false with some other outcome.

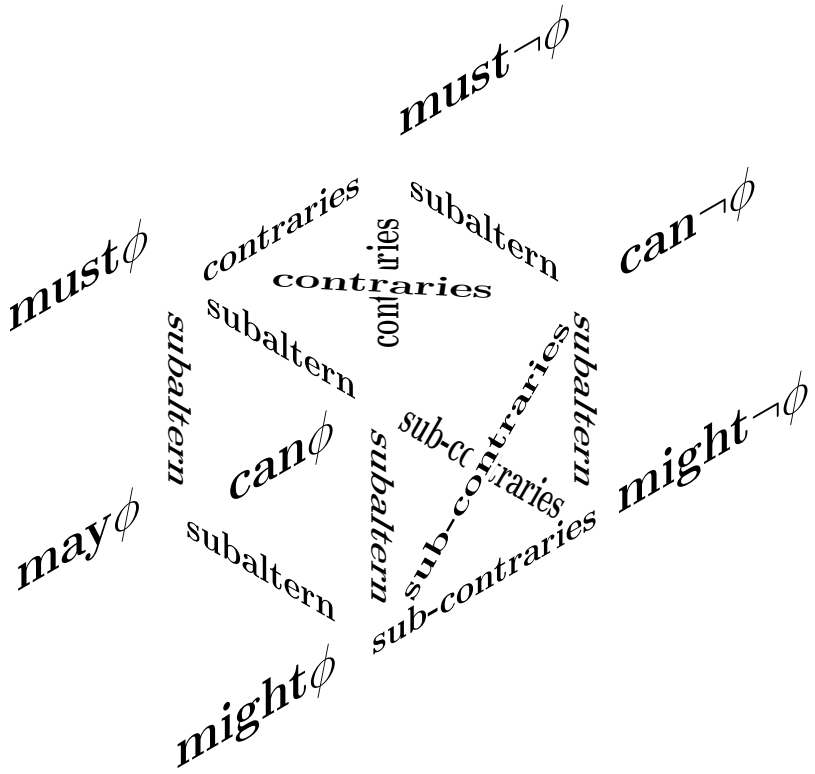

Fig. 4.3: Cube of opposition for ability modalities

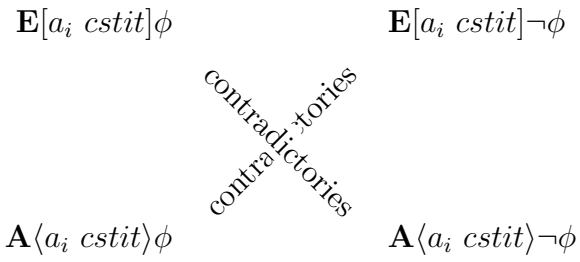

Fig. 4.4: Square of opposition for *can* and *may*

4.5 The Metaphysics of agency

The formal models used in this thesis are mainly meant as idealized representations of situations that might actually occur. Since the work is purely conceptual (no experiments, no serious appeal to any science other than mathematics) large parts of this thesis may be considered metaphysical.[2] However, so far I have limited myself to what was called metaphysics with a small 'm' in (Belnap et al.; 2001, p. v). The way I see it, rather than being a precise distinction, this indicates a commitment to everyday intuitions when introducing primitives, defining concepts, and so on. In this section, I momentarily free myself of this commitment and digress into loftier areas of the Metaphysical realm, spelled with a capital 'M'. I even have a proposition about that most elusive of beings to the effect that if there is an omnipotent god it is unique and solely responsible for everything.[3]
First, however, I would like to note that in situation models ability *must* and *might* are somewhat trivial, since ability *must* is equivalent to alethic *must* and ability *might* is equivalent to alethic *might*. The following are valid.

$$\mathbf{A}[a_i\ cstit]\phi \leftrightarrow \mathbf{A}\phi \text{ and } \mathbf{E}\langle a_i\ cstit\rangle\phi \leftrightarrow \mathbf{E}\phi.$$

Further, if we use the deliberative stit and formalize **must** ϕ as $\mathbf{A}[a\ dstit]\phi$ then there is nothing any agent must do, because $\neg\mathbf{A}[a\ dstit]\phi$ is valid. Either there is some indeterminacy or there is no deliberative action. Some might think this is okay, there is no event you must see to, you always have the choice to refrain (making it at least possible that the event does not occur), if this event is count as the result of an action. Others might use this fact to question the presented concept of ability *must*. That in turn might lead them to question the stit formalization of choice as a partition of all outcomes. If they want to keep the general idea they could move to a variant of the more liberal theory presented in Brown (1988), where choices are still subsets of outcomes but do not partition the outcomes. That would be a way to distinguish ability *must* from alethic *must*. I will not explore this option here. Instead, I will take the chapter in a rather different direction.

[2] Chapter 7 is an exception with its appeal to ordinary linguistic intuitions. That chapter can be considered independently, as contributing to the philosophy of language.

[3] Chapter 6 contains another kind of digression from metaphysics with a small 'm'. In that chapter, I do not consider unrealistic agents in order to say something about agency, but unrealistic situations in order to say something about responsibility.

The conceptual distinctions offered by the cube of opposition present an interesting way of characterizing various possible philosophical positions on the relation between agency and contingency, which will now be investigated.

4.5.1 Four reductionist positions about being and ability

Stit theory offers many ways of contrasting how agents may relate to contingent events and the cube of opposition offers one way of comparing them. It is, for instance, an interesting exercise to investigate different metaphysical views on agent ability. In order to make the discussion more interesting we might establish the common ground among these views that agents cannot do anything about tautologies and contradictions, and so the discussion will center around the status of contingent formulas and their negations (also contingent.) Moving from the top down in the cube, we first consider the position that everything that you do you must do or must not do i.e. for an agent a and for any contingent p, either $\mathbf{A}[a_i\ cstit]p$ or $\mathbf{A}[a_i\ cstit]\neg p$. We call such an agent *determined*. As we have just seen, determinism about action implies alethic determinism and so this view trivializes agent choice. There is no moral responsibility for determined agents, neither positive or negative, using the definitions of these concepts offered in Chapter 3. For the next positions it seems most interesting to consider conjunctions of ability formulas. Another possible position is that for an agent, and for any contingent p, $\mathbf{E}[a_i\ cstit]p$ *and* $\mathbf{E}[a_i\ cstit]\neg p$ holds, i.e. that the agent has regulative control over any contingent fact. Such an agent I call *omnipotent*. An omnipotent agent is responsible for *everything* that happens, since she either sees to it or could have prevented it! In most everyday situations, considering any agent to be omnipotent seems to imply a serious confusion of the real and the imaginary, and further it implies a certain kind of solipsism as shown below, since there can be at most one omnipotent agent in any given situation. Hence if two agents regard themselves omnipotent, at least one of them is wrong. A more subtle and interesting position is the one contrary to this. This is the position, for an agent a that she can never see to anything or prevent anything in the strong sense of stit theory. Let us call a person adopting this view an *ability sceptic*.[4] No matter what you say an agent sees to, the ability sceptic will respond that it was a possibility that things did not turn out this way. Further, whatever contingent event did *not* occur,

[4] I will briefly return to this view in Chapter 8.

it could have happened no matter what you did to try to prevent it. The ability scepctic holds that for any p, $\mathbf{A}\langle a_i\ cstit\rangle p$ and $\mathbf{A}\langle a_i\ cstit\rangle \neg p$. We call such an agent *very weak*. Like a determined agent, a very weak agent is also not responsible for anything in either the negative or the positive sense of the definitions in the previous chapter. The last position is indeterminism without control. We take this position to reduce all contingencies to $\mathbf{E}p$ and $\mathbf{E}\neg p$. An agent for which this holds we call *undetermined*. Whereas this does not rule out any of the other positions except determinism, indeterminism is not this is not *enough* to establish moral responsibility in either of the two senses. For example, an undetermined agent might be very weak and thus not responsible for anything.

In my view, none of these positions are very appealing since they all rule out important aspects of ability. Ability determinism rules out the possibility that agents can control some things, which seems to be contradicted by everyday human experience. On the other hand, indeterminism in itself is not enough to establish this control. It seems obviously wrong to consider human beings omnipotent, there are things we cannot either prevent or guarantee, such as the weather on distant planets. At any rate the existence of omnipotent agents is strictly contradicted by *naturalist realism*, which claims that there are some contingencies which are totally contingent for any agent. It is hardest to respond to the claim that all agents are very weak in the sense defined above. In fact, I am not sure that I can give a completely convincing reply within the present framework. In Chapter 8 a different definition of ability *can* is given, which is not affected by this objection. Somebody making this claim will grant that we have *some* control, we just can't ever 100 percent guarantee or prevent any contingency. Take an assertion about the result of a simple action, such as 'my right hand is lifted.' The position does not rule out that this is true because of my decision to lift my right hand a moment ago, i.e. I have causal responsibility for the event. What is claimed is that it is always the case that given my choice to raise my hand the sentence *might* in principle have turned out false, say by some freak quantum event disintegrating my entire body. If I had decided not raise my hand a sudden brain malfunction might have made me do it anyway. Thus just enough is claimed to undermine responsibility in the two agentive senses defined in the last chapter. Since the objection trivializes the theory I will offer a defence against it. The defence is pragmatic in nature. It will enable us to keep using the theory in face of the scepticism outlined above. In any discourse, certain

kinds of factors may be considered relevant or otherwise. For instance, in a court of law, low probability quantum events of the kind described above are not considered relevant. In a philosophical discourse more factors may be considered relevant. When modeling situations, we should restrict our attention to factors and agents, which may be considered relevant to the kind of situation we are modeling. Local Scepticism about specific contingencies in specific situations will play a big role in the chapter Frankfurt Examples. An interesting elaboration of the exercise is to consider that some agents are different from others. Let us be traditional and call an agent who is omnipotent in *every* situation a 'god.' We can logically establish that there can be only one god, so let us spell it with a capital G. Further, the existence of such an agent implies that all other other agents are *very weak*. Thus, if God has regulative control over all contingent facts, it implies that no human being is responsible for anything. It further implies that God is responsible for everything. For simplicity contingencies are represented by propositional variables. Also an explicit theory of dependencies between contingent facts (e.g. a person cannot have brown eyes and blue eyes at the same time) will not be presented . Such dependencies are not essential to this application. It will simply be assumed that the omnipotent being can enforce or prevent each individual contingent fact. Further, for this application, it is assumed that agents are allowed an infinite number of choices. [5]

Fact 4.1. The existence of a god, *God*, implies its uniqueness and that any other agent is very weak. Further *God* is responsible for any contingent fact and no other agent is responsible for anything.

Proof. Let *God* be an agent, such that for any $p \in \Phi$ and any situation M, *God* has regulative control over p. Assume that there is an agent b different from *God* and a $p \in \Phi$ and a situation M such that $M \vDash \mathbf{E}[b\ cstit]p$ or $M \vDash \mathbf{E}[b\ cstit]\neg p$. If $M \vDash \mathbf{E}[b\ cstit]p$, then this taken together with $M \vDash \mathbf{E}[God\ cstit]\neg p$ violates independence of agents, contradiction. (It follows that b is not positively responsible for p). If $M \vDash \mathbf{E}[b\ cstit]\neg p$, then $M \vDash \mathbf{E}[God\ cstit]p$ again violates independence of agents (and b cannot be negatively responsible for p.) It follows that $M \vDash \mathbf{A}\langle b\ cstit\rangle p \wedge \mathbf{A}\langle b\ cstit\rangle \neg p$.

[5] The restriction to finite choices in Chapter 2 was made to simplify the semantics of the deontic operators. Deontic operators are not central to this chapter. Further, by complicating the semantics of Horty's ought to do operator it can be extended to infinite choices, , see (Horty; 2001, Chapter 4) for details

Therefore any agent different from *God* is very weak in every situation and not responsible for anything. Further, if p is true with some outcome o of some situation M then *God* could have prevented it by $M, o \vDash \mathbf{E}[God\ cstit]\neg p$, so *God* is negatively responsible for p. If $\neg p$ is true with some outcome o of some situation M, then $M, o \vDash \mathbf{E}[God\ cstit]\neg\neg p$ makes *God* negatively responsible for $\neg p$. Therefore, *God* is responsible for everything. □

The conclusion of this argument is that either there is no god (who is omnipotent in every situation) or no other agent (including any human being) is responsible for anything in the agentive sense. However, causal responsibility for other agents is not ruled out by this argument. Also, while it is ruled out that other agents have control over anything, it is not ruled out that God leaves some things to chance. It is the fact that God *could* have seen to it that it turned out otherwise, which makes him responsible. As a corollary we have that if any agent has regulative control over any ϕ in any situation then any other agent in that situation is very weak with regard to ϕ.

The strength of stit theory lies in the fact that we do not necessarily have to accept any of the reductions represented by the positions (or any combination of them for different agents). On the contrary we can model subtle differences between kinds of contingencies. There are some contingencies that must be this way (presupposed in the situation, such as the presence of oxygen in most situations involving humans). Agents can have regulative control or only positive or negative control over other contingencies. Still other contingencies may depend on a specific choice and there can be total contingencies, which are independent of any choice.

4.6 Objections to the theory

How natural are the ability modalities presented here? One obvious objection is that like the deontic modalities *must* and *may* the ability modalities seem to operate on action types not sentences expressing propositions. Rather than saying something about which events we enforce they say something about our actions. One way of interpreting this is that an agent *can* t (where t is an action type) in a given situation if he somehow has an action instantiating t available to him in that situation. In Chapter 8, ability *can* will be interpreted along such lines. This provides an alternative to the theory presented in this chapter.

Chapter 5

Group responsibility

5.1 Joint agency

In this chapter a concept of group responsibility is defined. Further, individual responsibility for members of groups is considered. In Horty (2001) and Belnap et al. (2001) the definition of individual agency encountered in Chapter 2 is extended to group agency.

Let M be a utilitarian strategic model as defined in Chapter 2. To facilitate the definitions of group agency, the following *accessibility relation* R_i is defined for each agent.

Definition 5.1. Let $a_i \in Agent$. Let $o, o' \in dom(M)$.
oR_io' iff $o, o' \in A_{ij}$, for some $A_{ij} \in Choice_i$.

R_i is an equivalence relation. We now define accessibility relations for groups of agents,

Definition 5.2. Let $\Gamma \subseteq Agent$.

1. $R_\emptyset = dom(M) \times dom(M)$.
2. $R_\Gamma = \bigcap_{a_i \in \Gamma} R_i$, for $\Gamma \neq \emptyset$.

Because of *independence of agents*, R_Γ is an equivalence relation for each $\Gamma \subseteq Agent$. Truth conditions for various modal operators are now given in terms of these relations. The names of the operators are as follows. **A** is a universal modality. $[a_i\ cstit]$ is a single agent Chellas stit operator. $[a_i\ dstit]$ is a single agent deliberative stit operator. $[\Gamma\ cstit]$ is a joint Chellas stit operator. $[\Gamma\ dstit]$ is a joint deliberative stit operator. $[\Gamma\ stit]$ I call a joint Belnap/Perloff/Xu stit operator. It is a deliberative stit version of their joint achievement stit operator, see (Belnap et al.; 2001, p. 283). It is a joint Chellas stit operator with a negative condition added to it.

Definition 5.3. Let $a_i \in Agent$ and $\Gamma \subseteq Agent$.

1. $M, o \vDash \mathbf{A}\phi$ iff for any $o' \in dom(M)$, $M, o' \vDash \phi$.
2. $M, o \vDash [a_i\ cstit]\phi$ iff for any o', such that oR_io', $M, o' \vDash \phi$.

3. $M, o \vDash [a_i\ dstit]\phi$ iff for any o', such that oR_io', $M, o' \vDash \phi$ and for some $o'' \in dom(M)$, $M, o'' \nvDash \phi$.

4. $M, o \vDash [\Gamma\ cstit]\phi$ iff for any o', such that $oR_\Gamma o'$, $M, o' \vDash \phi$.

5. $M, o \vDash [\Gamma\ dstit]\phi$ iff for any o', such that $oR_\Gamma o'$, $M, o' \vDash \phi$ and for *no* Δ such that $\Delta \subset \Gamma$, $M, o \vDash [\Delta\ dstit]\phi$.

6. $M, o \vDash [\Gamma\ stit]\phi$ iff for any o', such that $oR_\Gamma o'$, $M, o' \vDash \phi$ and for some $o'' \in dom(M)$, $M, o'' \nvDash \phi$.

The clause 'for some $o'' \in dom(M)$, $M, o'' \nvDash \phi$', in the third part of the definition is called the *negative condition*. The clause ' for *no* Δ such that $\Delta \subset \Gamma$, $M, o \vDash [\Delta\ dstit]\phi$' in the fifth part of the definition is called the *generalized negative condition*. When Γ is a singleton it corresponds to the negative condition for the single agent deliberative stit operator.

5.2 Joint strict agency

As noted by Horty and by Belnap, Perloff, and Xu, the joint Chellas stit operator and the joint Belnap/Perloff/Xu stit operator validate the following formulas.

Fact 5.4. Let $\Gamma, \Delta \subseteq Agent$.

1. $\vDash [\Gamma\ cstit]\phi \rightarrow [\Delta\ cstit]\phi$, where $\Gamma \subseteq \Delta$.

2. $\vDash [\Gamma\ stit]\phi \rightarrow [\Delta\ stit]\phi$, where $\Gamma \subseteq \Delta$.

Agents in Δ but not in Γ are called *free riders*. Intuitively, a free rider is *inessential* for the proposition expressed by ϕ to obtain. In Belnap et al. (2001) the following definition of a member of a group being essential is presented.[1]

Definition 5.5. An agent $a_i \in \Gamma$ is *essential* for ϕ at an outcome o of M iff $M, o \vDash [\Gamma\ stit]\phi$ and $M, o \nvDash [\Gamma - \{a_i\}\ stit]\phi$.

[1] Strictly speaking, they define it for the achievement stit, here it is transposed to the joint Belnap/Perloff/Xu stit.

The same definition of an agent being *essential* can be given for the joint deliberative stit operator replacing everywhere *stit* with *dstit*. The authors propose a *joint strict agency operator* [Γ *sstit*] with the following truth condition, see (Belnap et al.; 2001, p. 287).[2]

Definition 5.6. $M, o \vDash [\Gamma\ sstit]\phi$ iff $M, o \vDash [\Gamma\ stit]\phi$ and for any set Δ such that $\emptyset \neq \Delta \subset \Gamma, M, o \nvDash [\Delta\ stit]\phi$

We have the following validities.

Fact 5.7. 1. $\vDash [a_i\ cstit]\phi \leftrightarrow [\{a_i\}\ cstit]\phi$.

2. $\vDash [a_i\ dstit]\phi \leftrightarrow [\{a_i\}\ dstit]\phi \leftrightarrow [\{a_i\}\ sstit]\phi$.

3. $\vDash [\emptyset\ dstit]\phi \leftrightarrow \mathbf{A}\phi$.

4. $\vDash [\Gamma\ dstit]\phi \leftrightarrow [\Gamma\ sstit]\phi$, for $\Gamma \neq \emptyset$.

In view of the first two of these validities, it is natural to identify the single agent stit operators with the corresponding joint stit operators for singletons. Further, in view of the last validity, the joint strict agency operator can be identified with the joint deliberative stit operator. In the following, I will therefore restrict my attention to the joint deliberative stit operator. The following validities ensure that every agent in the group is essential for the joint deliberative stit operator to hold and that it is not possible to add any free riders to the group.

Fact 5.8. Let Γ be a non-empty group of agents. Let $a_i \in \Gamma$ and let $\Gamma \subset \Delta$.

1. $\vDash [\Gamma\ dstit]\phi \rightarrow \neg[\Gamma - \{a_i\}\ dstit]\phi$.

2. $\vDash [\Gamma\ dstit]\phi \rightarrow \neg[\Delta\ dstit]\phi$.

Proof. 1. Assume $M, o \vDash [\Gamma\ dstit]\phi$. Since $\Gamma - \{a_i\} \subset \Gamma$, $M, o \nvDash [\Gamma - \{a_i\}\ dstit]\phi$. 2. Assume $M, o \vDash [\Gamma\ dstit]\phi$. To obtain a contradiction, further assume $M, o \vDash [\Delta\ dstit]\phi$. Since $\Gamma \subset \Delta$, $M, o \nvDash [\Gamma\ dstit]\phi$., contradiction. □

For this operator, the following validities are also worth mentioning.

[2] Again, they define it for the achievement stit, and I transpose the definition to the strategic situation framework.

Fact 5.9. 1. $\vDash [\Gamma\ dstit]\phi \leftrightarrow [\Gamma\ cstit]\phi \wedge \bigwedge_{\Delta\subset\Gamma} \neg[\Delta\ dstit]\phi$.

2. $\vDash [\Gamma\ dstit]\phi \rightarrow \phi$.(**T**)

3. $\vDash [\Gamma\ dstit]\phi \rightarrow [\Gamma\ dstit][\Gamma\ dstit]\phi$.(**4**)

Proof. 1. Obvious from the truth condition of the joint deliberative stit operator. 2. Use reflexivity of the cstit operator. 3. If Γ is empty, the result boils down to $[\emptyset\ cstit]\phi \rightarrow [\emptyset\ cstit][\emptyset\ cstit]\phi$, which holds by transitivity of the $R_\emptyset$ relation, $W \times W$. So assume that Γ is non-empty and that $M, o \vDash [\Gamma\ dstit]\phi$ (so $M, o \vDash [\Gamma\ cstit]\phi$) and $M, o \nvDash [\Gamma\ dstit][\Gamma\ dstit]\phi$. So either a_1) $M, o \nvDash [\Gamma\ cstit][\Gamma\ dstit]\phi$ or a_2) $M, o \vDash [\Delta\ dstit][\Gamma\ dstit]\phi$ for some $\Delta \subset \Gamma$. If a_1) there is some o', with $oR_\Gamma o'$, and $M, o' \nvDash [\Gamma\ dstit]\phi$. So either b_1) $M, o' \nvDash [\Gamma\ cstit]\phi$ or b_2) $M, o' \vDash [\Delta\ dstit]\phi$ for some $\Delta \subset \Gamma$. If b_1) there is some o'', with $o'R_\Gamma o''$ and $M, o'' \vDash \neg\phi$. By transitivity of R_Γ, $oR_\Gamma o''$, so $M, o'' \vDash \phi$, contradiction.
If a_2), we prove that this entails $M, o \vDash [\Delta\ dstit]\phi$. Let o' be such that $oR_\Delta o'$. It follows that $M, o' \vDash [\Gamma\ dstit]\phi$ and by **T** that $M, o' \vDash \phi$. Thus $M, o \vDash [\Delta\ cstit]\phi$. Let $\Sigma \subset \Delta$. Since $\Sigma \subset \Gamma$, $M, o \nvDash [\Sigma\ dstit]\phi$. It follows that $M, o \vDash [\Delta\ dstit]\phi$, contradiction.
If b_2), $M, o' \vDash [\Delta\ cstit]\phi$ and by the obvious $[\Delta\ cstit]\phi \rightarrow [\Delta\ cstit][\Delta\ cstit]\phi$ (**4** for the joint Chellas stit operator), $M, o' \vDash [\Delta\ cstit][\Delta\ cstit]\phi$. By symmetry of R_Γ, since $oR_\Gamma o'$, $o'R_\Gamma o$. Since $R_\Gamma = \bigcap_{a_i\in\Gamma} R_i \subseteq \bigcap_{a_j\in\Delta} R_j = R_\Delta$ it follows that $o'R_\Delta o$. Hence $M, o \vDash [\Delta\ cstit]\phi$. Let $\Sigma \subset \Delta$. Since $\Sigma \subset \Gamma$, $M, o \nvDash [\Sigma\ dstit]\phi$. It follows that $M, o \vDash [\Delta\ dstit]\phi$ contradicting $M, o \vDash [\Gamma\ dstit]\phi$. □

The principle **5**, $\neg[\Gamma\ dstit]\phi \rightarrow [\Gamma\ dstit]\neg[\Gamma\ dstit]\phi$ however, is not valid. To see this consider Figure 5.1. ϕ is true with outcome o_1, so we must have that $\{a, b\}$ is not preventing ϕ there that is
$M, o_1 \vDash \neg[\{a, b\}\ dstit]\neg\phi$. Since we have this with o_2 as well, and $M, o_5 \vDash [\{a, b\}\ dstit]\neg\phi$, we have $M, o_1 \vDash [a_i\ dstit]\neg[\{a, b\}\ dstit]\neg\phi$, which implies, by the impossibility of free riders proved above, $M, o_1 \nvDash [\{a, b\}\ dstit]\neg[\{a, b\}\ dstit]\neg\phi$.

5.3 Joint refraining

Horty considers two plausible definitions of individual refraining, see (Horty; 2001, p. 26).

a

K_1	o_1: ϕ	o_2: ϕ
K_2	o_3:ϕ o_4:$\neg\phi$	o_5: $\neg\phi$
	K_3	K_4

b

Fig. 5.1: A counter model to **5**

Definition 5.10. 1. $\neg[a_i\ dstit]\phi \wedge \mathbf{E}[a_i\ dstit]\phi$.

2. $[a_i\ dstit]\neg[a_i\ dstit]\phi$.

As Horty notes, the following validity makes the two definitions interchangeable.

Fact 5.11. $\vDash (\neg[a_i\ dstit]\phi \wedge \mathbf{E}[a_i\ dstit]\phi) \leftrightarrow [a_i\ dstit]\neg[a_i\ dstit]\phi$

The following is a naive attempt at lifting the first concept of refraining to groups.

A group Γ refrains from ϕ with an outcome o of a situation M iff
$M, o \vDash \neg[\Gamma\ dstit]\phi \wedge \mathbf{E}[\Gamma\ dstit]\phi$.

However, with this definition the equivalence to the other obvious definition, $[\Gamma\ dstit]\neg[\Gamma\ dstit]\phi$, does not hold and this for a conceptually good reason. It happens when a subgroup of Γ is preventing Γ from seeing to it that ϕ, in which case we would not say that Γ is refraining from ϕ. Consider the following situation, where a father is talking to a friend, while his son is standing by.

Example 5.1. 1. Dad: 'We refrained as a family from buying a car.'

2. Son interrupts: 'That is not correct. You decided that we weren't going to buy it, because we couldn't afford it.'

Let us assume that there are just the father and the son in the family. The situation can be modeled as in Figure 5.2, where ϕ means that a car is bought for the family. With outcome o_1, the group consisting of father and son, does not see to it that ϕ. Also, there is a choice the group can make, the one resulting in outcome o_5, where the group sees to it that ϕ. So with the first definition of joint refraining, the other condition is met. However,

Father	K_1	o_1: $\neg\phi$	o_2: $\neg\phi$
	K_2	o_3:ϕ o_4:$\neg\phi$	o_5: ϕ
		K_3	K_4
		Son	

Fig. 5.2: The choices of father and son

since the father is vetoing ϕ at o_1 it is not intuitively correct to say that the group is refraining. And for exactly the same reason the group is not seeing to it that, the group is not seeing to it that ϕ, so the two obvious definitions of refraining do not coincide.

In order to rule out these cases I propose the following definition of joint refraining.

Definition 5.12. A group Γ refrains from ϕ with an outcome o of a model M iff $M, o \vDash \neg[\Gamma\ dstit]\phi \wedge \bigwedge_{\Delta\subset\Gamma} \neg[\Delta\ dstit]\neg[\Gamma\ dstit]\phi$

In words, a group refrains from ϕ, if and only if, the group does not see to it that ϕ, and no subgroup of the group is preventing the whole group from seeing to it that ϕ. It is immediate from the definition that this concept is equivalent to the first option given by Horty in the single agent case that is

Fact 5.13. An agent a_i refrains from ϕ with o iff $M, o \vDash \neg[a_i\ dstit]\phi \wedge \mathbf{E}[a_i\ dstit]\phi$.

Furthermore, we get the following more general equivalence result.

Fact 5.14. $\vDash (\neg[\Gamma\ dstit]\phi \wedge \bigwedge_{\Delta\subset\Gamma} \neg[\Delta\ dstit]\neg[\Gamma\ dstit]\phi) \leftrightarrow [\Gamma\ dstit]\neg[\Gamma\ dstit]\phi$

Proof. Left to right. Assume $M, o \vDash \neg[\Gamma\ dstit]\phi \wedge \bigwedge_{\Delta\subset\Gamma} \neg[\Delta\ dstit]\neg[\Gamma\ dstit]\phi$ and $M, o \nvDash [\Gamma\ dstit]\neg[\Gamma\ dstit]\phi$. Then, from the first conjunct, $M, o \nvDash [\Gamma\ dstit]\phi$. Further, from the second conjunct, either a_1) $M, o \nvDash [\Gamma\ cstit]\neg[\Gamma\ dstit]\phi$ or a_2) $M, o \vDash [\Delta\ dstit]\neg[\Gamma\ dstit]\phi$, for some $\Delta \subset \Gamma$. If a_1) there is an o', with $oR_\Gamma o'$, and $M, o' \vDash \neg\neg[\Gamma\ dstit]\phi$. Because of symmetry of the choice relation, $o'R_\Gamma o$, and because of **4**, $M, o' \vDash [\Gamma\ dstit][\Gamma\ dstit]\phi$, so we have $M, o \vDash [\Gamma\ dstit]\phi$, contradiction.
If a_2) we have $M, o \vDash \neg[\Delta\ dstit]\neg[\Gamma\ dstit]\phi$, contradiction.
Right to left. Assume $M, o \vDash [\Gamma\ dstit]\neg[\Gamma\ dstit]\phi$ and $M, o \nvDash \neg[\Gamma\ dstit]\phi \wedge$

$\bigwedge_{\Delta\subset\Gamma} \neg[\Delta\ dstit]\neg[\Gamma\ dstit]\phi$. Then, from the second conjunct, either a_1) $M, o \vDash [\Gamma\ dstit]\phi$ and, from the first conjunct, $M, o \vDash \neg[\Gamma\ dstit]\phi$, contradiction. Or a_2) $M, o \vDash [\Delta\ dstit]\neg[\Gamma\ dstit]\phi$ for some $\Delta \subset \Gamma$ and, by the generalized negative condition, $M, o \vDash \neg[\Delta\ dstit]\neg[\Gamma\ dstit]\phi$, contradiction. □

5.3.1 Arendt on collective responsibility

In everyday life, we are certainly held (personally) responsible for things we could only achieve with other people. To put it a bit differently, we are held responsible *as members of groups*. How can we account for that within our theory of agency? To begin with, there are certain pitfalls given our account of joint agency. Consider the following inference chain.

Example 5.2. 1. 'Some Germans saw to the Holocaust.'

2. 'All Germans saw to the Holocaust.' (From 1.)

3. 'All Germans are responsible for the Holocaust.' (From 2.)

We will call the above *Arendt's fallacy*, because Hannah Arendt discusses it in several places, see e.g. Arendt (1964). Clearly, the joint agency property poses a problem, if we want to lift the concept of responsibility from individuals to groups. If I see to ϕ, then we see to ϕ (in a weak sense, which permits free riders), but we are not both necessarily responsible for ϕ. If we lifted individual responsibility to groups by the Chellas stit, then when somebody is responsible, all are, and the concept is trivialized. As Hannah Arendt puts it:

> Where all are guilty, nobody is.(Arendt; 1968, p.147)

In the following I only consider agentive aspects of positive joint responsibility. The supposedly non-trivial work of integrating joint intentions into the framework as well as negative responsibility for groups is left for somebody else.

5.4 Positive responsibility of groups

As Arendt sees it, a useful concept of collective responsibility really requires a concept of *active participation*. Therefore, although Arendt's Fallacy is about collective responsibility, it seems natural to block it at the level of

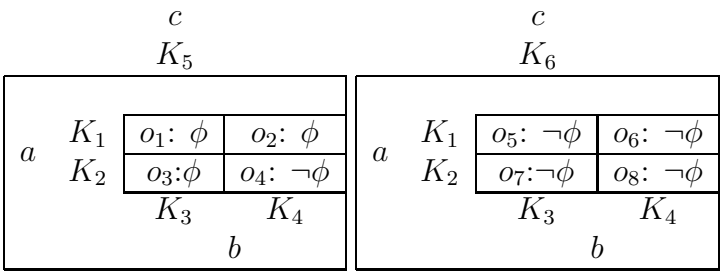

Fig. 5.3: The choices of 3 agents

joint agency, i.e. by resisting the inference from 1. to 2 in the fallacy above. However, if we simply generalize the concept of responsibility with the the dstit operator, we already have a concept of active participation built into our definition of responsibility. This idea corresponds exactly to holding a group Γ responsible for ϕ at an outcome o iff. $M, o \vDash [\Gamma\ dstit]\phi$. It follows immediately that Γ is responsible for ϕ only if every nonempty $\Delta \subseteq \Gamma$ is essential for ϕ.

5.4.1 Holding members of groups personally responsible

Presumably, an individual can be held responsible for ϕ if he is a member of some group, Γ, which deliberatively sees to ϕ. The following definition captures that intuition.

Definition 5.15. Let $a_i \in Agent$. a_i can be held personally responsible for ϕ, iff $[\Delta \cup \{a_i\}\ dstit]\phi$, for some $\Delta \subseteq Agent$.

In words, to substantiate that you cannot be held responsible for ϕ on account of denial of active participation, it is not enough that you are a free-rider in some group, which sees to ϕ. You will have to be a free-rider of any subgroup of that group as well. Conversely, to check if a_i can be held personally responsible for ϕ start with the group containing just a_i, and go through any subset $\Delta \subseteq Agent$. If $[\Delta \cup \{a_i\}\ dstit]\phi$ is true then a_i can be held personally responsible. To illustrate this idea consider the following example.

> In the Third Reich, at any rate, there was only one man who did and could make decisions and hence was politically fully responsible. That was Hitler himself who, therefore, not in a fit of megalomania but quite correctly once described himself as the only man in all Germany who was irreplaceable. (...)Does this mean that nobody else could be held personally responsible?(Arendt; 1968, p. 30)

I do not wish to interpret Arendt or Hitler as such, the reason for giving the above quote is to point to something more general.[3] People will sometimes argue that because somebody else might have done what they do, they are not responsible. To illustrate this fallacy, consider Figure 5.3. We here have three agents, $Agent = \{a, b, c\}$. Each agent faces two choices each containing 4 outcomes, e.g. a must choose between $K_1 = \{o_1, o_2, o_5, o_6\}$ and $K_2 = \{o_3, o_4, o_7, o_8\}$. There will then be a single outcome when everybody has chosen. Let ϕ mean some crime, to keep with our example the supreme crime of genocide, and let c be a dictator, who can in fact prevent ϕ by choosing K_6, where ϕ is false with any outcome. None of the other agents have this power to prevent, and c is as such necessary. He is the only agent, who is individually (negatively) responsible for ϕ. Now, let us look at what groups can achieve ϕ. Since either the cooperation of a or b is enough to ensure ϕ each of them is 'replaceable' for achieving ϕ. However, c needs *somebody* to cooperate with him. If a and b choose K_2 and K_4, respectively, then ϕ will be prevented. Further, they may be personally responsible depending on their choice. Consider first outcome o_3. The choices made are K_2,K_3, and K_5. We have here that $[\{b\} \cup \{c\}\ dstit]\phi$, so b and c can both be held personally responsible for ϕ. However with this outcome there is no $\Gamma \subseteq Agent$, such that $[\{\Gamma \cup \{a\}\ dstit]\phi$, so a cannot be held responsible for ϕ. Consider now o_1. Here $\neg[\{b\} \cup \{c\} \cup \{a\}\ dstit]\phi$, because either a or b could be taken out and ϕ would still be true. However, the fact that an agent is not needed to achieve an event, does not mean that she cannot be held responsible. This is determined by her active participation in achieving it. In this chapter active participation is understood as the existence of some subgroup, which you

[3] To mention just one point, in the quoted essay Arendt makes an intricate distinction between being politically responsible and personally responsible, which I have no use for at present. Among other things, political responsibility involves considering the whole history of a nation.

belong to, which deliberatively sees to it that ϕ. In fact both a and b belong to such a subgroup, since $[\{c\} \cup \{a\}\ dstit]\phi$ and $[\{c\} \cup \{b\}\ dstit]\phi$ are true at o_1, and since these groups deliberatively see to it that ϕ, the agents can be held personally responsible. So, on this account personal responsibility for active participants is compatible with one person being a negative dictator that is having the sole executive power to personally veto ϕ. Furthermore, the fact that others might have done what an agent did even if the agent had refrained so that the agent was not not strictly needed for the event to obtain, does not free that agent from responsibility.

Chapter 6

Frankfurt examples

6.1 Introduction

In this chapter, some examples of how the theory of individual responsibility may be used are presented. Also, some connections to informal meta-ethics are established. The chapter is centered around a certain cluster of thought experiment known as *Frankfurt examples* or *Frankfurt-Style examples*. Frankfurt examples play an important role in the ongoing debate about incompatibilism vs. compatibilism. In 'folk' philosophy this kind of thought experiment is sometimes used instead of a theoretical argument; the Frankfurt examples *show* that moral responsibility and determinism are compatible, it is claimed. However, the intuitive understanding of thought experiments is not in itself certain enough to establish any such conclusion. Thought experiments without an explicit theory to go with them are just stories, and it is not the task of philosophy to tell stories, cf. Plato *The Sophist*. Any serious compatibilist or incompatibilist will have a theory, such as e.g. van Inwagen (1978) or Fischer and Ravizza (1998). Given such a theory, it seems a legitimate task to test it on various 'hard' cases. The Frankfurt examples may be considered hard cases: it *seems* that the agent is determined and morally responsible in the thought experiment. How can that be explained? The strategy of this chapter will be to analyze these thought experiments within the theoretical framework already presented in Chapter 3. It will then become clear that they do not pose any particular problem to libertarian incompatibilism as understood in this thesis. Further the examples help us understand various concepts related to agency, such as overdetermination.

6.2 Factors

The thought experiments to be discussed involve certain variables in the environment of the agent, which are important to the overall situation facing the agent. For instance, in Locke's example there is a 'room', which may be 'locked' or 'unlocked'; in the Frankfurt example there is a 'device', which may be 'implanted' or 'not implanted' into the brain of the agent. Thus, to adequately represent the examples discussed in this chapter, it will be necessary to extend the models with a set of non-intentional, non-normative 'agents.' These will be called *factors*. The factors represent unknowns that

might influence the final outcome of a situation. An example is the weather. Imagine an agent considering going to the beach the next day. The weather might be good (not too windy, not too hot, ...) or bad (too hot or too windy or ...) If the weather is good, the trip will be pleasant, if the weather is bad, it will be unpleasant. The 'possible values' for the weather, 'good' or 'bad', correspond to the choices of an agent. They will be represented accordingly, as a partition of the outcomes. The equivalence classes of the partition will be called the *alternatives* of the factor. As a common name for agents and factors, the term *influences* will be used. Formally, the factors are just like agents, except they do not have intentions. It is also assumed that values do not apply to factors. Thus, although it is possible to hold factors responsible (in an agentive sense), it is not possible to hold them guilty, blameworthy or praiseworthy. It is now possible to represent examples such as the following, which include non-agentive unknowns.

Example 6.1 (Assassin). An assassin is about to decide whether to shoot at a politician from afar. She knows that the wind and the way her hands tremble will influence the final outcome of her choice. The assassin does not have a licence to kill, and the murder or attempted murder is bad from a legal and from a moral perspective.

6.2.1 Formalizing the assassin example

The example is formalized with the model represented in Figure 6.1. The choices, outcomes and values of this situation are *given* in the informal story and represented formally. There is an element of creativity involved in such a formalization. Intuitively, the representation may be more or less faithful. An explicit understanding of this relation between the stories and formal models is beyond the scope of this thesis. There is an assassin, denoted a_1, who is confronted with the choices 'Shoot' and 'Don't Shoot.' The atomic propositional variable D expresses the event that the politician is shot to death. The non-agentive factors are the wind and the hands of the assassin, which may be steady or unsteady. Precisely how the wind factor plays a role is not specified, so the alternatives of this factor do not have suggestive names. The alternatives of the wind factor are simply called K_1 and K_2. If the assassin chooses to shoot, then her intention is to kill the politician. Formally, the intention set of that action is o_1, $I_{Shoot} = \{o_1\}$. We assume that she is indifferent between the outcomes, if she does not shoot, $I_{Don't\ Shoot} =$

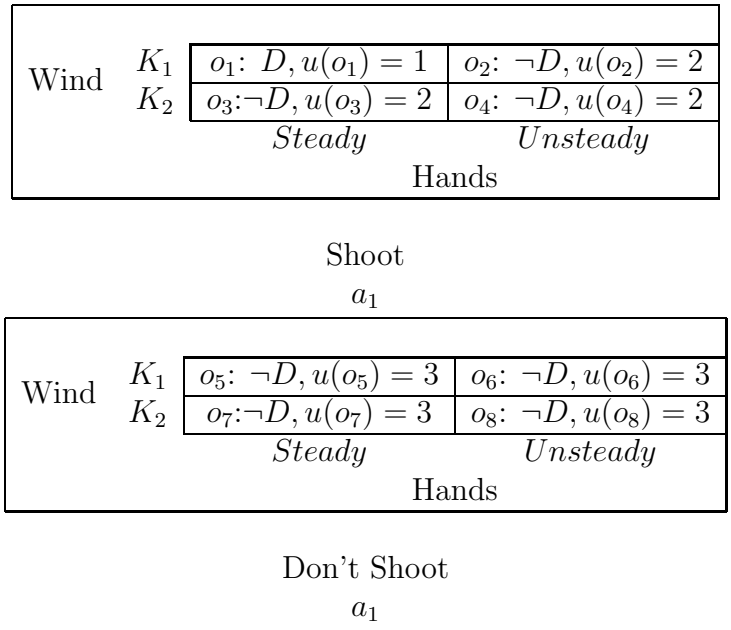

Fig. 6.1: The choice of the assassin

$\{o_5, o_6, o_7, o_8\}$. Further, the utilities represent the legal or moral value of the outcomes, not the private utility of the outcomes for the assassin. From the legal or moral perspective considered here, killing the politician is worst, risking his life is bad, and not shooting is best. This is so, even if the assassin were to get away with murder, collect the prize of two million dollars for the murder and live happily ever after on a tropical island. According to these values, the choice to not shoot strictly dominates the choice to shoot. As a consequence, the assassin ought not to kill the politician, $M \vDash \bigodot[a_1\ cstit]\neg D$. It can also be observed, for instance, that if the assassin shoots and the wind behaves as in alternative K_1, and her hands are steady, then the final outcome of the situation is o_1. It is also easy to determine the various kinds of responsibility, guilt, and blameworthiness at the different outcomes.

- The assassin is blameworthy for D at o_1.
- The assassin is blameworthy for attempting D at o_2, o_3, o_4 (since she is guilty of attempting D and D is forbidden).

- The assassin is praiseworthy for $\neg D$ at $\{o_5, o_6, o_7, o_8\}$

it may be objected that not very much is required of the agents to make them praiseworthy. Normally, we would not praise anybody for not assassinating somebody. However, the concepts are relative to the outcomes of the given situation. In any number of particular situations, it can be hard enough to do the right thing.

The factors can also be used to reason about the circumstances that played a role in determining the final outcome. For instance, the assassin did not succeed in killing the politician with o_2, because her hands were unsteady. Had her hands not been unsteady she would have killed him.

6.3 Causal responsibility, agentive responsibility, overdetermination

It is useful to distinguish between *causal responsibility* for an event and *agentive responsibility* for an event. An agent is causally responsible for an event, if the (physical) agent is the actual cause of the event. The concepts of responsibility (guilt, blame and so on) defined in Chapter 3 are all agentive concepts. Causal and agentive responsibility are logically independent. For instance, an agent may accidently bump into somebody and cause him to fall, without agentively seeing to it that he falls. Conversely, and more controversially, an agent arguably might see to it that an event obtains, although she is not the cause of the event with the actual outcome of the situation. Corresponding to these two kinds of responsibility there are two ways an event can be overdetermined.

1. An event E may be *causally* overdetermined.

2. An event E may be *agentively* overdetermined.

Agents may be agentively responsible for an event, although they are not causally responsible for it. This happens in case of agentive overdetermination. These will often also be cases of causal overdetermination. (Belnap et al.; 2001, p. 290) consider only what is here called agentive overdetermination.[1] They define overdetermination as distinct agents seeing to the same event, and the same definition will be adopted here.

[1] To be precise, the authors define overdetermination for the so-called achievement stit and not the deliberative stit as considered here. The definition is easily transposed to cover the deliberative stit, however.

Agent 1			
	Press red	o_1: Bomb o_2: Bomb	o_3: Bomb
	Press green	o_4: Bomb	o_5: No bomb
		Press red	*Press green*
		Agent 2	

Fig. 6.2: Red button or green button

Definition 6.1. An event expressed by a formula ϕ is *agentively overdetermined* with an outcome o of a situation M iff there are more than one *distinct* influences $a_1, \ldots a_n$, such that

$$M, o \vDash \bigwedge_{1 \leq i \leq n} [a_i \ dstit] \phi$$

.

We thus also speak of agentive overdetermination, even if one or more of the agents $a_1, \ldots a_n$ are actually factors, not agents. Although agentive overdetermination might free an agent from causal responsibility (i.e. the agent might not be the actual cause of the event), it does not free an agent from agentive responsibility. This is of course formally speaking immediately clear from the definition. The following example is meant to provide some intuitive justification.

Example 6.2 (Red button or green button). Two agents each has to press either a red button or a green button. If an agent presses the green button, nothing happens, except that pressing the red button afterwards will now have no effect. If an agent presses the red button, the multi million city Metropol is blown to pieces. The agents press their buttons independently. The bomb is released by the first agent who presses his button, but if they both press it around the same time, there is no way of telling which button actually caused the release of the bomb.

This example is represented by Figure 6.2.

The atomic formula *Bomb* expresses the event that Metropol is blown to pieces. If both agents press the red button, the outcome will be in the top left box. Here, with outcome o_1 and o_2, the event that Metropol is blown to pieces is agentively overdetermined. Either agent could have made his other choice, and the city would still have been blown to pieces. However, there is

only one actual cause of the event, say Agent 1's button with o_1, and Agent 2's button with o_2. With the concept of positive responsibility both agents are responsible for the city being blown up. The justification is that they both acted so as to ensure that the city is blown up. So, although nobody might ever know 'who actually caused it', it is still possible to place agentive responsibility. Further, even if there is no such uncertainty, and it *is possible* to place causal responsibility, the theory still holds both agents agentively responsible. Thus, ultimately, it is the choices of the agents that they are held responsible for. The independence of the actions is really important. For instance, if one agent knows that the other has pressed the red button, then he will not be responsible for the bomb, even if he presses the red button afterwards.
In other examples of agentive overdetermination it may be ontologically impossible to determine an actual cause. Two agents each simultaneously poison a person with a dose sufficient for killing him . The amount of poison actually killing the person is a mix of the two doses.

6.4 The philosophical context of the Frankfurt examples

As another example of how to apply this theory we consider Frankfurt examples, originally proposed by Harry Frankfurt in Frankfurt (1969), see also van Inwagen (1978), Hunt (2000), Fischer (2005a), Lippert-Rasmussen (2005). Frankfurt examples aim at showing that alternative choices are not a prerequisite for moral responsibility. At the heart of Frankfurt examples lie the following two apparently contradictory claims.

Principle 6.2 (PAP)**.** There can be no moral responsibility without alternative choices.

Frankfurt There are situations, where an agent is responsible for ϕ, although circumstances which in no way influence the agent's choice, nonetheless make it impossible for that agent to refrain from doing ϕ.

It is claimed that there are situations as described by *Frankfurt*. Furthermore, this is claimed to rule out *PAP* (principle of alternate possibilities). We have a theory which gives a precise meaning to the crucial concepts of these principles: Circumstances, alternative choices, situations, refraining, responsibility. It turns out that it is possible to consistently claim that there

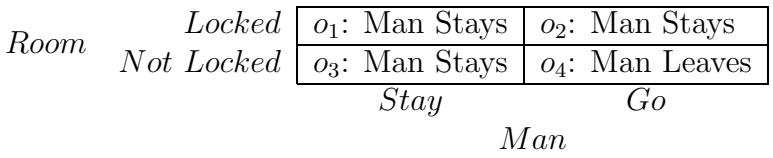

Fig. 6.3: Man in a locked room

are situations as described in *Frankfurt*, while maintaining that moral responsibility is dependent on alternative choices. The analysis will show that a libertarian research strategy has a lot of explanatory power, even when it comes to examples specifically devised to refute that very strategy. First, consider the following example by John Locke (see also Hunt (2000), Fischer (2005a)):

Example 6.3 (Man in a Locked Room). ...suppose a man be carried, whilst fast asleep, into a room, where is a person he longs to see and speak with; and be there locked fast in, beyond his power to get out: he awakes, and is glad to find himself in so desirable company, which he stays willingly in, i.e. prefers his stay to going away. I ask, is not this stay voluntary? I think nobody will doubt it; and yet being locked fast in, 'tis evident he is not at liberty not to stay, he has not freedom to be gone. (Locke; 1690, Bk II,ch. XXI)

This thought experiment is often cited as a starting point for the compatibilist position that agents can do something of their own free will without being able to do otherwise. For now, it is important to note the following features of the example. There is a man, facing a choice either to stay in a room or to go. Further, the exits out of the the room might be locked or not. This circumstance is an important factor in the situation, which will influence the outcome of his choice. The example is represented by Figure 6.3. If the man chooses to leave, the intended outcomes of his choice are the ones where he leaves, $I_{Go} = \{o_4\}$. If he chooses to stay the intended outcomes are the ones where he stays, $I_{Stay} = \{o_1, o_3\}$. In other words, the intention of the man going is to leave the room and that only happens with the outcome o_4, which only becomes actual if the room is not locked. The utilities of the outcomes are not important to this example, so they are left unspecified.

Figure 6.3 seems to provide a fair representation Locke's example. The agent can choose between attempting to go and staying and the room is

locked or not. The situation is modeled so that the room being locked or not in no way influences the agent's choice, i.e. his choice is independent of the alternative of this factor.

6.4.1 Reasoning about outcomes

Now it is possible to analyze the four possible outcomes of this situation in turn. It is easy to see that with o_1, o_3 the man is guilty of staying, with outcome o_2 he is guilty of attempting to leave, but not of staying, and with outcome o_4 he is guilty of leaving. His guilt at o_1 (the outcome singled out by Locke) is quite independent of the fact that the event of him staying is overdetermined with this outcome.

o_1 The agent decides to stay, the door is locked. With this outcome the agent is (positively) responsible for staying since he sees to it that he is staying. The fact that the door is locked is independent of this and does not matter to his responsibility. Him staying is overdetermined, but this does not free him from responsibility. This is the outcome described in Locke's example.

o_2 Here the man attempts to go, but he is forced to stay, because the room is locked. That he is guilty of attempting to leave is clear, since it was his intention to leave, $M, o_2 \vDash [Man\ iit]Man\ leaves$,it did not happen, but it could have happened with his choice . If the room were unlocked he would have left, so he does not see to it that he is staying. It follows that he is not positively responsible for staying. Furthermore, he could not have prevented staying by making his other choice. As a consequence, he is not negatively responsible for staying either.

o_3 The man deliberatively sees to it that he stays, so he is (positively) responsible. With this outcome the door is open.

o_4 The man tries to leave and the door is unlocked so he succeeds. He could have unconditionally prevented leaving by making his other choice, so he is (negatively) responsible for leaving.

Now, let us consider a version of the original Frankfurt example.

Example 6.4 (Frankfurt example). A man, Jones, must decide whether to murder or not to murder another man, Peter. A third man, Black, has built

a device into Jones' head, so that in case Jones decides to not murder Peter, he will do so anyway. The device is only activated, if Jones decides not to kill Peter. As it happens, Jones decides on his own to murder Peter, so the device is never used. Jones is thus responsible for the murder, although he could not have refrained from the murder.

6.4.2 Informal approaches to the Frankfurt examples

Most responses to Frankfurt's examples have been informal. They can roughly be divided into compatibilist responses, defending Frankfurt, and incompatibilist responses, attacking him. Kane summarizes the typical incompatibilist response as follows.

> ... Suppose that A is the action that the controller wants Jones to perform ... and suppose Jones does A on his own without Black interfering. Many responses to Frankfurt... have taken the following general form. If Jones is responsible in this case, it is because he *did A on his own*(i.e., of his own free choice, without interference from Black). But Jones *could* have done other than that: he could have done other-than-A-on-his-own by not choosing or trying to do A and forcing Black to intervene. If Black intervened, to be sure, Jones would still have done A, but he would not have done A-on-his-own. So, responsibility and could-have-done-otherwise are not disconnected after all. Where Jones is responsible (for doing-A-on-his-own), he could have done other than that. And where he could not have done otherwise..., he is not responsible.(Kane; 1998, p. 41)

The analysis to follow agrees in many ways with the type of response presented by Kane.[2] The main difference is that the formalism makes parts of the analysis stand out more clearly.

6.4.3 Lewis' causal responsibility approach

David Lewis gives a semi-formal account of the Frankfurt example in Lewis (2000). The paper gives a counterfactual account of causality in light of

[2] Kane thinks the response has merit, but he does not fully agree with it. I can not go into his objections here. I refer to his book, Kane (1998)

the problems posed by overdetermination, or as Lewis calls it, *preemption*. With regard to the Frankfurt example, his main concern is to explain why Jones' choice causes the event of Peter's death, although Jones is not able to exercise any influence on that event. According to Lewis' definition, one event D influences another E, when there is a range of changes in D, which result in changes of E. To keep the present terminology we might think of events as sets of outcomes and suppose a linear temporal ordering of events. The dependence of outcomes in one event on another event can be represented by functions from events to events, $I_D : D \to E$, and so on. Imagine a temporal ordering of three events, D, E, F. Assume that D influences E and E influences F. Transitivity of influence might break down. This happens when any outcome $o \in E$ in the range of the function I_D, is mapped to the same single outcome $s \in F$ by I_E. No change in D will effect a change in F, so D does not influence F. According to Lewis, the Frankfurt example is a case in point. Jones' choice, (D), does influence Black's behavior, (E): if he chooses to murder Peter, Black will not interfere; if he chooses to refrain, Black will interfere. Further, the behavior of Black influences the murder, (F), since we also might assume possibilities where Black does not interfere although Jones refrains. However, no matter what Jones chooses, Peter will be murdered, hence D does not influence F. Therefore Lewis suggests that we take causality to be not simply influence but rather the ancestral (reflexive, transitive closure) of the relation given by events influencing each other. According to that definition of causality Jones' choice does cause the event of the murder, although his choice did not influence it. The question remains whether this causation is really enough to establish responsibility. Lewis thinks it is.

> The moral of the story is that preemptive causation, without dependence, suffices to confer ownership and responsibility of ones actions.
> (Lewis; 2000, p. 193)

We might agree that Lewis' definition gives reasonable necessary and sufficient conditions for establishing *causal responsibility* for an event, and that Lewis' theory is very elegant for that purpose. However, when it comes to moral responsibility or in general *agentive* responsibility, the theory is inadequate. The reason is this. Also with the outcomes, where Jones decides not to murder Peter and Black interferes, he is responsible on Lewis' account.

This is as it should be, if we talk about causal responsibility, since Jones certainly did cause the murder of Peter. He was the one firing the gun (or whatever) causing Peter's death. However, we would not say that Jones is *morally* responsible with this chain of events. He decided not to murder Jones, but Black interfered with his brain and made him do it. Obviously, Jones is not to blame for the murder. Thus, if we are to assume that Lewis' theory is to cover everything there is to say about moral responsibility it gives wrong predictions with some outcomes of the Frankfurt example. Alternatively, and this is what I prefer, we can take Lewis' theory to be about causal responsibility, which must be distinguished from agentive responsibility. In that case it gives the right predictions, it seems to me. However, when it comes to the Frankfurt example, it is not mainly causal responsibility, we are mainly interested in , but *agentive* responsibility. A typical way out for the compatibilist at this point would be to require an additional internal component in the agent to establish moral responsibility. Thus, Jones is responsible with the outcomes where he causes the event and his choice is based on the process of an appropriate inner mechanism causing him to decide, see e.g. Fischer and Ravizza (1998). This does work well in the Frankfurt examples. However, the incompatibilist solution presented below, does not require the introduction of such additional inner mechanisms.

6.4.4 Approaches within stit theory

Also within stit theory itself, the Frankfurt examples have been treated. This has been done in (Belnap et al.; 2001, Chapter 9) and in the paper Paprzycka (2002). Here I will focus on the latter.

Paprzycka's strategy for dealing with the Frankfurt examples contains two important components.

1. Weakening the requirement of regulative control for moral responsibility.

2. Making *stit* a necessary condition for moral responsibility.

Paprzycka argues that compatibilists take incompatibilists to require regulative control over an event in order to establish moral responsibility. Recall that an agent a has regulative control over ϕ in a situation M, when $M \vDash \mathbf{E}[a\ dstit]\phi \wedge \mathbf{E}[a\ dstit]\neg\phi$. Clearly Jones does not have this kind of

control, he cannot prevent the murder. Thus, with this definition of responsibility libertarians would have to claim that Jones is not responsible, even if he chooses to murder Peter on his own. Certainly, Inwagen (one incompatibilist) has been taken by Frankfurt (one compatibilist) to think this, or rather something even stronger than this.

> A person is fully responsible, then, for all and only those events or states of affairs which come about because of what he does and which would not have come about if he did otherwise. . . . Perhaps. . . he [Inwagen] construes moral responsibility in this strong sense. (Frankfurt; 2005, p. 280)

It will be shown below that the reason why Inwagen denies agents responsibility for states of affairs in the Frankfurt example is quite different from what Frankfurt suggests. It should also be noted that Inwagen's necessary condition for moral responsibility is weaker than the one suggested by Frankfurt. Inwagen requires that the agent has the ability to prevent an event to be responsible for it. However, this does not preclude that the agent has the further choice to leave the event up to chance. In that case the event *might* have come about if he did otherwise, so he is not responsible on Frankfurt's definition. Paprzycka, however, suggests to loosen this negative requirement of regulative control altogether. She suggests replacing the ability to prevent in the definition of regulative control with the negative condition of the achievement stit operator. Consequently, she suggests the following necessary condition for responsibility.

Principle 6.3 (Paprzycka)**.** One can be responsible only for what one sees to.

Paprzycka then gives two different interpretations of the Frankfurt examples, one she calls *non-reductive* and one she calls *reductive*. Both of these interpretations can be easily transposed to the deliberative stit, which I will do here. The main idea is that, in order for the negative condition to be fulfilled, there must be an alternative outcome, where the event does not obtain. There are two obvious ways of achieving this: differentiating the *outcomes of the two actions* or *introducing an external agent or factor into the situation*. On the first interpretation, either Peter is murdered by Jones on his own or Peter is murdered by Jones as a consequence of the device.

Device			
	Yes	o_1: Peter murdered, Device off	o_2: Peter murdered, Device on
	No	o_3: Peter murdered	o_4: Peter not murdered
		Murder	*Don't murder*

Jones

Fig. 6.4: Frankfurt example

However, strictly speaking, Jones is not really acting at the outcome where he murders Peter as a consequence of the device. After all he did not intend to murder Jones at this outcome. Therefore, it is not true to say that *Jones murders Peter* at this outcome. Thus the negative condition is fulfilled. I find this analysis problematic. Although, he did not act voluntarily, I think it is correct to say that Peter murders Jones, given his causal responsibility for the event. Similarly, one can say that Peter bumps into Jones, and the like, although it was an accident. However, if that is correct, the negative condition is not fulfilled.

Paprzycka's second interpretation requires treating Black as an agent in the situation. Clearly Black could have chosen not to interfere, so there is a different possibility. This is basically the approach taken below, with some minor differences. I treat the device (not Black) as an independent factor in the situation. I use the deliberative stit instead of the achievement stit, a minor change. The most important difference is that I do not accept Paprzycka's principle of responsibility, since it rules out the possibility of negative responsibility. Negative responsibility is required to get the right prediction with the outcome where the device is not implanted and where Jones chooses not to murder Peter. Here, Jones is negatively responsible for not murdering Jones. Further, I include intentions.

6.4.5 Analysis of a Frankfurt example

The Frankfurt example is represented by Figure 6.4.

The scenario is formally very similar to Locke's example. It is not entirely correct to say that Jones has the choice to murder or not murder Peter, at least he does not have regulative control over the event of Peter being murdered. Rather, he has the choice to murder Peter or *try to* not murder Peter. The purpose of the formulae 'device off', 'device on' is to capture the idea that Black only shows his hand if necessary, the interference is

counterfactual. So, there is both a distinction between the device being there or not, and between the device being used or not resulting in four possible outcomes.

o_1 Jones decides to murder Peter on his own, the device is implanted. The device is not activated. However, we can reason counterfactually as follows. If Jones had decided not to murder Peter, circumstances staying the same (the device is implanted), he would have murdered Peter anyway.[3] Thus, given the circumstances, he could not have done otherwise. The reason that he is responsible, however is the following. It is a necessary consequence of his choice that Peter is murdered, and it is also possible (given different circumstances, i.e. that the device is not there) that he does not murder Jones. Thus he deliberatively sees to it that Peter is murdered. The fact that the device is there and independently ensures this event as well, does not free him from responsibility. The device is also *agentively* responsible for the event. The example is constructed so that the device is not *causally* responsible for the murder. This is the situation described in the Frankfurt example.

o_2 Jones tries not to murder Peter, but he is forced to do so, because the device is there. The murder is not a necessary consequence of his choice, so he is not positively responsible. Furthermore, he could not have prevented the murder by making his other choice, so he is not negatively responsible. As a consequence, he is not responsible for the murder with either of the concepts of responsibility.

o_3 Jones decides to murder Peter on his own and the device is not there. This is a normal case of homicide, Jones is positively responsible.

o_4 Jones decides not murder Peter and there is no device, so he succeeds. He is negatively responsible for not murdering Peter, since he could have seen to the murder by making his other choice.

[3] The account of counterfactual reasoning given here, is made precise in (Horty; 2001, pp. 81-85). Since it is not essential to my argument, I refer to Horty for details. The main idea is the following. When we evaluate a counterfactual statement, such as 'with another choice for agent a ϕ would have been the case', we keep all other circumstances fixed and change only the choice of that particular agent, and then we evaluate ϕ.

The theory holds Jones responsible for murdering Peter with outcome o_1 and outcome o_3, and frees him from responsibility for the murder with outcome o_2 and o_4, and that seems to agree with libertarian intuitions. It is furthermore clear that there is no inconsistency between outcomes of situations, where one could not have done otherwise and the principle of alternative possibilities. The reason is that the 'could not have done otherwise' in the Frankfurt example is arrived at by keeping the circumstances fixed.

6.5 Frankfurt examples and negative responsibility

Compatibilists seem to focus only on positive responsibility for the obvious reason that negative responsibility explicitly requires a 'could have done otherwise.' At the other extreme, Inwagen, in a famous paper, only recognizes negative responsibility, see van Inwagen (1978). He presents principles for responsibility for *actions*, *event particulars*, and *event universals* or states of affairs. Here I will only consider the latter two, since they relate directly to the concepts of responsibility for events considered in this thesis.

6.5.1 Inwagen on event particulars

Inwagen takes *event particulars* to be objects individuated by different causal histories. They correspond roughly to the individual outcomes of situations in this thesis. Inwagen presents the following principle for event particulars.

Principle 6.4 (Inwagen, event particulars)**.** A person is morally responsible for a certain event particular only if he could have prevented it.

Now, regarding the Frankfurt example, Inwagen argues as follows. Since the outcome where Jones murders Peter on his own, and the outcome where he murders him as a consequence of the device have different causal histories they are different event particulars. Let us call them E and D. Hence, where Jones is responsible for E, he could have prevented it by making his other choice, causing D instead. Inwagen's principle is extremely week. With any event particular that an agent brings about he prevents all others. It should be noted, though that since Inwagen's principle only states one necessary condition for responsibility, it does not entail that Jones *is* responsible for the murder at D. Nothing prevents the proponent of Inwagen's theory from requiring further necessary conditions. One of these might be a suitable mental mechanism as mentioned above. Even though the principle is very weak,

Fischer attempts to construct a counter-example to it along the following lines, see (Fischer; 2005a, p. 296).

Example 6.5. If Black detects that Jones is about not to murder Peter, then he will use his machine to destroy Jones' brain and thus kill him instantly.

Fischer then states that then Jones could not have brought about a different event particular. Fischer seems to confuse choices with the known consequences of choices here. Certainly, Fischer must agree that the event where Peter is murdered is very different from the event where Jones' brain is destroyed. Thus, by choosing not to murder he *does* bring about an event different from the one, where he murders. The event looks very different from what it thought it would look like when he made his choice. He does no longer exist. However that is irrelevant for Inwagen's theory. The question remains, though whether such a weak principle is relevant in establishing moral responsibility. Related to this concern is another argument of Fischer's against this sort of response to the Frankfurt example, which he has dubbed a *flicker of freedom* response. In the Frankfurt example, the flicker of freedom consists in the slight difference, which there must always be in the outcome where the device is activated and the one where it is not. If there was not *any* difference, they would be the same outcome. However, Fischer claims, this difference is not enough to base a concept of responsibility on. I agree with Fischer on this. Certainly it is not enough to establish the negative condition for seeing to the murder, which requires that the murder might not have taken place. However, the role the device plays in the Frankfurt examples is as a factor independent of the choice of the agent. It therefore seems reasonable to represent it as such, which will be done below.

6.5.2 Inwagen on event universals

Regarding event universals Inwagen presents the following principle, see (van Inwagen; 1978, p. 210):

Principle 6.5 (Inwagen, event universals)**.** A person is morally responsible for a certain state of affairs only if (that state of affairs obtains and) he could have prevented it from obtaining.

It is clear that Inwagen sees states of affairs as propositions or sets of possible worlds, a view, which seems similar to the one adopted here. Inwagen

does not consider values or intentions. Thus the principle seems to correspond to the definition of an agent being strictly liable as defined in Chapter 3.

However, Inwagen wants to establish the negative conclusion that agents can never be responsible for event universals. In order to strengthen this claim, he presents two examples. One is a Frankfurt example. The other one is like the following.

Example 6.6. A person is stuck on a horse. He cannot make the horse stop, but he can make it turn left or right. However, whether he turns left or right, the horse will end up in Rome.

Inwagen claims, it is obvious that the man is not responsible for ending up in Rome. Indeed it is. Since ending up in Rome is true with any outcome of the situation, the man cannot be responsible for it. This example only establishes the conclusion that some agents are not responsible for some events they cannot prevent, see also Fischer (2005a). The Frankfurt examples are different and require another kind of argument. When it comes to the Frankfurt example, Inwagen's agrees with the stit analysis that the Frankfurt example is a counterexample to his principle of event universals. With the outcomes where Jones sees to it that Peter is murdered, he could not have prevented this event universal. The response given in this thesis is that with that outcome of the Frankfurt example Jones is indeed only positively responsible. Inwagen does not seem to acknowledge a concept of positive responsibility. Instead he claims that it is not possible to be responsible for event universals at all! In conclusion, if we are talking about event particulars, then agents always can prevent them. If we are talking about event universals, agents can never prevent them.
However, Inwagen's argument for the latter is faulty. He argues as follows. Consider an outcome of the Frankfurt example where Jones murders Peter on his own. With these outcomes Jones' choice is a sufficient condition for the event that Peter dies. Jones' choice is also a sufficient condition for the event that Peter is mortal. However, we will not say that Jones is responsible for Peter being mortal. So far, Inwagen's argument is fine and in accordance with the stit analysis. Any of Jones' choices is a sufficient condition for Peter being mortal, for presumably, Peter is mortal with any outcome. However, precisely therefore the negative condition cannot be fulfilled. Thus, Jones is *not* responsible for Peter being mortal. This does not entail that he is not responsible for the murder. Here the negative condition is fulfilled. In

order to establish his conclusion that Jones is not responsible for the event that Peter dies, Inwagen needs more. He continues by claiming that Jones being mortal is arguably the same proposition as the proposition that Jones dies. If this was so, Inwagen would have his conclusion that Jones is not responsible for the event universal that Peter dies. Because by substitution of identicals, it would entail that Jones is responsible for Peter being mortal, which is absurd. However, Inwagen clearly commits an error in semantics here. In any moment of Peter's life it is true that he is mortal. It is only true that he dies in the last moment of his life.[4] Therefore *Peter is mortal* is not the same proposition as *Peter dies* and Inwagen's argument fails. Inwagen presents other argument but they are all based on a similar error in basic semantics. We conclude that Inwagen has not established that agents cannot prevent at least some event universals from obtaining. Where does this leave the concept of negative responsibility?

6.5.3 In defence of negative responsibility

Inwagen's basic intuition is that responsibility is connected to the ability to prevent. However, he fails to give an adequate theory of what that means. Regarding event particulars his principle is too weak to be of any interest, since it is basically true with any action of any agent. Regarding event universals he failed to establish the conclusion he was going for. Thus a concept of negative responsibility for events or event universals, considered as sets of outcomes is still possible. However, one might still speculate whether we really need such a concept. Let us take an example from the same paper by Inwagen.

Example 6.7 (Inwagen's Telephone Example). Suppose I look out the window of my house and see a man being robbed and beaten by several powerful-looking assailants. It occurs to me that perhaps I had better call the police. I reach for the telephone and then stop. It crosses my mind that if I do call the police, the robbers might hear about it and wreak their vengeance on me... So I decide not to get involved... Now suppose also that, quite unknown to me, there has been some sort of disaster at the telephone exchange, and that every telephone in the city is out of order... Am I responsible for failing to call the police? Of course not. (van Inwagen; 1978, pp. 204-205)

[4] For a formal model, one can take any finite linear or branching temporal model and let the states be the moments of Peter's life.

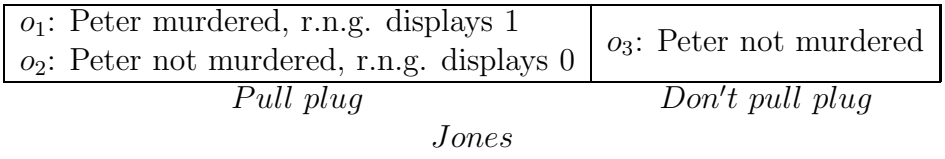

Fig. 6.5: Random number generator 1

In fact, we must model this example exactly like the Frankfurt example, see Figure 6.4. Of course the agent could not have prevented the police from not being called, so he is not negatively responsible. However, he is positively responsible for failing to call the police. The compatibilist might certainly maintain that agents can be responsible for negative facts. They might argue as follows. It was a necessary consequence of your choice that the police was not called. Hence you are responsible. The compatibilist can argue convincingly that the agent is responsible for failing to call the police. Thus, this example is not strong enough to establish that we need a concept of negative responsibility. Establishing such a need requires an example of an agent, who is *only* negatively responsible without being positively responsible. For that purpose, let us assume that there is true randomness in nature, and that Black has access to a random number generator based on such a random physical phenomenon. When the device is activated, the random number generator displays a 1 or a 0.

Example 6.8 (Random Number Generator). Black has hooked up a gun to his random number generator. If the end state is 0, the gun will not fire. If the end state is 1 it will fire. The device will be turned on in a few seconds, unless Jones pulls the plug. He does not pull the plug, the device displays 1, the gun fires and murders Peter. Is Jones responsible for the murder?

With the outcome described in the story the event of Peter's death is in some sense random. It is not correct to say that Jones brings about the event of the murder with his choice, at most one can say that he brings it about that the murder *might* happen. Causally, the murder is due to a truly random physical process initiated by another agent, Black. Jones did not cause the murder anymore than the walls of the room caused it by not collapsing. However, intuitively, Jones *is* responsible for the murder. Why? Because, as a free and accountable agent, he could have easily prevented it by pulling the plug. For a formal representation look at Figure 6.5. We are asking if Jones is responsible for the murder outcome o_1. Clearly he is not

Device		Pull plug	Don't pull plug
	Yes	o_1: Peter murdered, r.n.g. displays 1 o_2: Peter not murdered, r.n.g. displays 0	o_3: Peter murdered, Device on
	No	o_3: Peter murdered, r.n.g. displays 1 o_3: Peter not murdered, r.n.g. displays 0	o_6: Peter not murdered

Jones

Fig. 6.6: Random number generator 2

positively responsible for the murder, it was not a necessary consequence of his choice that Peter was murdered. However, he could have prevented the murder, so he is negatively responsible.
What this example does is undermining positive responsibility as much as possible, while maintaining negative responsibility in order to show the need for the concept.

We should be able to test the claim that the agent is only negatively responsible further by adding a Frankfurt overdetermined enforcer to the situation. Since such an enforcer undermines negative responsibility and since the agent is only negatively responsible, such an enforcer should free the agent of responsibility altogether.

Example 6.9. Black has connected a random device to a gun. If it displays a 1, Peter will be shot. If it displays a 0, Peter will not be shot. Jones must choose whether to pull the plug or not. However, Black has implanted a device into Jones' head. If Jones decides to pull the plug, it will make him pull the trigger manually, ensuring that Peter dies. Is Jones responsible at the outcome where he presses the button and it shows 1?

Is Jones responsible for the murder at outcome o_1 of Figure 6.6. Clearly he could not have prevented it by making his other choice. On the other hand, he did not see to it, since it depended on what the random number generator displays. Formally, there is no doubt he is not responsible. He is neither negatively or positively responsible. What about intuitively? It is fair to say that Jones is in a kind of lottery situation. Since it is in no way up to him whether Peter dies or not, it seems correct that he is not responsible for the murder.

The strategy in this chapter has been to maintain two concepts of responsibility from Chapter 3, a negative and a positive. The theory has then

been put to use in an analysis of the Frankfurt example. It is the purpose of the Frankfurt example to show that *PAP* is not needed for a theory of moral responsibility. We have undercut this claim in two different ways. First, by giving a thorough formal analysis of the Frankfurt example, which shows that there are indeed situations as described by *Frankfurt*. However, these do not show that *PAP* is not needed for moral responsibility. On the contrary, maintaining a theory requiring *PAP*, enables us to distinguish models, where an agent brings about an event, which might have turned out differently under different circumstances (The Frankfurt example), from models where the event was necessary.

I do not agree with Inwagen that negative responsibility is all there is. There is still a strong case to be made for maintaining a concept of positive responsibility, see also (Fischer; 2005a, 294). Agents may be responsible for seeing to an event, although it could not be prevented under the circumstances as in the Frankfurt example. The negative condition is necessary to rule out responsibility for events, which are necessary in the situation. On the other hand I do not agree with Paprzycka that positive responsibility is all there is. In particular, negative responsibility cannot be reduced to positive responsibility for negated propositions, as some authors suggest, see e.g. Fischer and Ravizza (1998). I have argued that there are examples of negative responsibility for an event with no positive responsibility for that event. When these are turned into Frankfurt cases agents are not responsible for the event at all . If this is correct, the claim that negative responsibility is needed, is fully justified.

Chapter 7

Deontic logic for action types

An old tradition in modern deontic logic going back to von Wright's first paper on the topic, applies deontic operators to *action types*. On the proposal presented here, it works in the following way. We have a non-empty set of action tokens, which instantiate various action types. A subset of the action tokens are singled out as *acceptable* or *good*. Informally, an action type is permissable, if there is an instance of the type, which is good or acceptable. An action type is required if all good action tokens are of that type. My aim is to show that a good portion of the problems of deontic logic can be solved by these devices. In order to operationalize this last claim, I will use this introduction to present what have been considered bench mark cases for a deontic logic. These cases are supposed to represent some deontic inferences a fluent speaker probably would make in a natural language context, even if (or perhaps especially if) that speaker has no theoretical knowledge about deontic logic.[1] Since my default belief is that natural language should be respected as far as possible, I elevate these bench mark cases to normative standards that a good deontic theory should meet. Here I appeal to normal linguistic intuitions as a regulative guideline. My view is not that logical theory should be founded on these sorts of empirical observations of natural language phenomena. On the contrary, a good theory gives a plausible theoretical explanation of the inferences it validates. However, a theory should provide a very good theoretical reason, if it diverges from natural language, preferably a reason arising within the theory itself. Philosophical logicians are often too eager to sacrifice natural language to save the logical theory they have become accustomed to working with. I divide the benchmark cases into three categories: consequences (what one should be able to infer), non-consequences (what one should not be able to infer), and equivalences (what should mean the same). To the best of my knowledge, there is no deontic logic which gets all these right that is, validates or proves the consequences, does not validate or prove the non-consequences and validates or proves the equivalences. Most of the examples are in the tables because they play an important role in natural language. However, most of them are also there to highlight specific problems encountered by one or more existing deontic

[1] I have not tested this statistical claim empirically, although it is easy to imagine an experiment to test it, hence the disclaimer 'are supposed to represent.'

	From	Infer
1	You must run or hide	You may run and you may hide
2	You must run	You may run
3	You may run or hide	You may run and you may hide
4	You may run and you may hide	You may run or hide
5	You may run and hide	You may run and you may hide
6	You must run and hide	You must run and you must hide
7	You must run and you must hide	You must run and hide
8		You may run or you may not run
9		not (You must run and you must not run)

Fig. 7.1: Consequences

logics. For instance, Standard Deontic Logic, Horty's stit logic, as well as almost all variants of dynamic deontic logic validate variants of Ross' paradox (1. and 2. of the non-consequences).

7.1 Overview of the chapter

First, I consider some of the bench mark cases, which have caused problems for formal theories. In particular, I consider Ross' paradox (Figure 7.2: 1,2), free choice inferences (Figure 7.1: 1,3,4), and conjunction exploitation (Figure 7.1: 5). While going through the problems I review some important contributions from the literature. After that I say a bit about action types in philosophy and legal theory to motivate the formal theory of action types, which will be presented. The main part of the chapter is a deontic logic of action types. Finally I discuss some issues about expressivity and natural

	From	Do not Infer
1	You must run	You must run or eat
2	You may run	You may run or eat
3	You must run or hide	You must run or you must hide
4	You may run and you may hide	You may run and hide
5	You may eat or not eat	You may run
6	You may eat	You may eat and hit the waiter

Fig. 7.2: Non-consequences

		If and only if
1	You must run	not (you may not run)
2	You may run	not (you must not run)

Fig. 7.3: Equivalences

language.

7.2 Ross' paradox and free choice inferences

Ross' paradox was introduced by Alf Ross, see Ross (1941). In standard deontic logic (and many other deontic logics) the paradox crops up in form of the following validity.

$$\bigcirc\phi \rightarrow \bigcirc(\phi \vee \psi)$$

The paradox has a dual twin.

$$\mathbf{P}\phi \rightarrow \mathbf{P}(\phi \vee \psi)$$

Ross' paradox brings out differences between deontic and epistemic or alethic modalities. For instance, the following justification of Ross' paradox does not work: since a disjunction gives less information than either disjunct it is pragmatically inappropriate to give a disjunctive permission, when you can give a categorial one. This explanation misses the problem it is supposed to explain, which is a problem of semantics, not pragmatics. Can we infer a disjunctive permission from a categorical one at all? There is no doubt that we can infer a disjunctive assertion from a categorical one, the question in pragmatics is why we do not do that very often. However, there is doubt when it comes to permissions. Further, it is incorrect to say that a disjunctive permission gives less information than either disjunct. The reason is that a disjunctive permission is always a free choice permission (see below). This implies that a disjunctive permission behaves much like a conjunction, so that it actually gives the addressee more information about her possibilities to act than either disjunct. Whereas the ideal of theoretical pursuits is to narrow the possibilities down as much as possible to arrive at the truth, this is not the case in practical life. Quite often, we would like more possibilities, which is to say, more freedom. As an example, take an employer who tells an employee that she may work through July. Later on he reveals that it was actually the case that she was allowed to work through July or go on vacation. However, he gave her the first 'more precise information' to respect pragmatic principles. The employer was not lying, it is true that the employee was allowed to work through July. Let us assume that she in fact did work through July and that the employer strongly believed that this would be the case when he uttered the first permission. Still, the employer has hidden

important information about his employee's rights from her. This would not be the case, if he had replaced a disjunctive assertion with either disjunct. For instance, assume that the employee asks him where the stapler is. He cannot be blamed for answering that it is in his office instead of answering that it is in his office or in the photo copying room, if he strongly believes it is in his office and it is in his office. The main reason then, why Ross' paradox must be considered a problem, is the occurrence of *free choice inferences* in natural language. The following is an example of a free choice inference.

Example 7.1. From 'You may send the letter *or* burn the letter' infer 'you may send the letter *and* you may burn the letter.'

Most accounts of deontic logic do not validate this inference. For instance, for a normal modal logic such as standard deontic logic, we have:

$$\not\models \mathbf{P}(\phi \vee \psi) \rightarrow (\mathbf{P}\phi \wedge \mathbf{P}\psi)$$

An inference of deontic logic, which is usually considered to be unproblematic, though, is the following, which we refer to as '*must* implies *may*.'

Example 7.2. From 'you must send the letter' infer 'you may send the letter.'

Or, in general, for any formula ϕ.

$$\bigcirc\phi \rightarrow \mathbf{P}\phi$$

It is now perhaps clear, how we may get problematical reasoning by combining Ross' paradox with free choice reasoning, as sketched in the following chain of informal reasoning.

Example 7.3.

1. 'You must send the letter.' (Given.)
2. 'You must send the letter or burn the letter.' (From 1. by Ross' Paradox.)
3. 'You may send the letter or burn the letter.' (From 2. by *must* implies *may*.)
4. 'You may send the letter and you may burn the letter.' (From 3. by free choice inference.)
5. 'You may burn the letter.' (From 4. by classical logic.)

This is clearly unacceptable. From any obligation to do something, I get the permission to do anything I want! Now, coming from classical logic, one is inclined to explain Ross' paradox away by denying the non-classical free choice inference. The reflex response by now is to deny the free choice inference status as an important problem, indeed it is sometimes called a *pseudo* problem. This is usually followed by the remark that we can express free choice as a conjunction of permissions, $\mathbf{P}\phi \wedge \mathbf{P}\psi$. However, as Hans Kamp pointed out already in 1973, we have a strong intuition that a disjunctive permission does entail the permission to do either disjunct, see Kamp (1973). On a closer inspection, the degradation of the free choice inference to the status of pseudo problem begins to sound a lot like dogmatic clinging to a certain paradigm, i.e. the paradigm of representing deontic reasoning with a normal modal logic or at least a monotonic one, as some of the logics suggested by Chellas, see Chellas (1980). Reconsidering the inference above with natural language as a guide, the problematic step does not seem to be from 2. to 3., but rather from 1. to 2..

7.2.1 Ross' paradox - a problem for stit theory

Ross' paradox may be seen as a problem for stit theory, at least if stit theory is supposed to reflect reasoning about norms in natural language. The deontic logics considered in Horty (2001) as well as the extensions presented in Chapter 2 all succumb to Ross' paradox. We have the following validity.

$$\vDash \bigodot[a\ cstit]\phi \rightarrow \bigodot[a\ cstit](\phi \vee \psi)$$

Of course, the free choice inference is not valid on Horty's account. Still, one might want to block Ross' paradox, and it seems there is an obvious way to do so. Simply introduce a *deliberative* ought operator. Supposedly, the truth condition for such an operator should be as follows, where we limit ourselves to finite choices.

Definition 7.1. $M, o \vDash \bigodot[a_i\ dstit]\phi$ iff $K \subseteq |\phi|_M$ for each $K \in Optimus_i'$ and there is an o', such that $M, o' \nvDash \phi$

It is clear that the validity constituting Ross' paradox is now blocked.

Fact 7.2. $\nvDash \bigodot[a\ dstit]\phi \rightarrow \bigodot[a\ dstit]\phi \vee \psi$

This inference is blocked with an outcome o, where we have $M, o \vDash \bigodot[a\ dstit]\phi \wedge \mathbf{A}(\phi \vee \psi)$. It is easy to construct such a model, M. Let $dom(M) = \{o_1, o_2\}$, $V(p) = \{o_1\}, V(q) = \{o_2\}$. There is just one agent, a. Assume that the two actions of this agent consist of one outcome each and that o_1 has higher utility than o_2. We then have $M, o_1 \vDash \bigodot[a\ dstit]p \wedge \neg \bigodot[a\ dstit](p \vee q)$. However, there is a good reason to reject this as a natural solution to Ross' paradox.[2] The following *is* valid.

Fact 7.3. $\vDash (\bigodot[a\ dstit]\phi \wedge \mathbf{E}(\neg\phi \wedge \neg\psi)) \rightarrow \bigodot[a\ dstit](\phi \vee \psi)$

This makes the following inference possible.

Example 7.4.

1. 'You ought to send the letter.' (Given.)
2. 'Possibly, the letter will neither be sent or burned.'(Given.)
3. 'You ought to send the letter or burn the letter.' (From 1. and 2. by the validity above.)

Since it seems quite appropriate to assume as a possibility that the letter is neither sent or burned, the proposed solution given to Ross' paradox does not work. Incidently, this critique applies also to the account of solving the paradox for imperatives by means of the achievement stit operator presented in (Belnap et al.; 2001, pp. 83-85).

7.3 Strong permission

A tradition in deontic logic, which takes free choice permissions seriously goes back to Anderson (1966). The idea is to define a *strong* permission operator **P**as follows, where $\Box$ is a normal modal operator and **ok** is a propositional constant, which is true iff a world is morally acceptable.

Definition 7.4. $\mathbf{P}\phi \leftrightarrow \Box(\phi \rightarrow \mathbf{ok})$

Anderson's proposal validates the permission versions of the free choice inferences and blocks Ross' paradox. However, there are obvious problems (see also Asher and Bonevac (2005)). For instance, as the reader can easily check, if we interpret the box as a normal **K** operator.[3] Anderson's proposal

[2] The argument laid out in the rest of this section was presented to me by Frank Veltman in a conversation

[3] Should we interpret the box this way? Since the relation between ϕ and **ok** seems to be a causal one, 'ϕ leads to an okay state', we might need to move to conditional logic.

validates $\mathbf{P}\phi \rightarrow \mathbf{P}(\phi \wedge \psi)$ and $\mathbf{P}(\phi \vee \neg\phi) \rightarrow \mathbf{P}\psi$, both counter-intuitive in a deontic context, (see Figure 7.2: 5, 6).

7.3.1 Conjunction exploitation

Moreover, Anderson's strong permission fails to validate $\mathbf{P}(\phi \wedge \psi) \rightarrow \mathbf{P}\phi$ called *conjunction exploitation*, see (Figure 7.1: 5). In a way, this is quite natural, when we think about the semantics of the strong permission, because it makes sense to read it as something like 'it is safe to.' From 'it is safe to jump from a plane and wear a parachute' it does not follow that 'it is safe to jump from a plane' simpliciter - if you do not wear a parachute, for instance, you might get in trouble. However, I do think that the failure to validate this principle shows that Anderson's strong permission operator does not capture what we normally mean by a permission. For instance, the following conjunction seems problematic.

Example 7.5. 'You may invite Beth and Smilla, but you may not invite Beth or you may not invite Smilla.' (?)

How can it be that you may invite Beth and Smilla, but there is one of them you may not invite? In other words, it seems that deontic logic should obey the following principle of conjunction exploitation.

$$\mathbf{P}(\phi \wedge \psi) \rightarrow (\mathbf{P}\phi \wedge \mathbf{P}\psi)$$

The converse, on the other hand should not hold. It does not follow, from the fact that you may take an apple and you may take a pear, that you may take an apple and take a pear.

A recent update of Anderson's proposal, Asher and Bonevac (2005) fails to validate conjunction exploitation. Moreover, on any such account permissions and obligations fail to be duals - or we must have both a strong and a weak permission operator. There are certainly strength differences in natural language. For instance, 'have to' is stronger than 'should', as can be seen by comparing the sentences, 'you should wash your hands but you do not have to' which is okay to 'you have to wash you hands but you shouldn't', which seems problematic at least when the 'have to' comes from the same normative source as the 'should not', see also von Fintel and Iatridou (2008). However, it is not clear that natural language actually contains the distinction suggested by the theory of strong permission.

7.4 Dynamic deontic logic

There are at least three different traditions, which all deserve the name *dynamic deontic logic*. One body of work starts with Meyer (1988), and continues with e.g. van der Meyden (1996), (see also Meyer and Veltman (2007) for more references.)

One inspiration for this work is the following validity of propositional dynamic logic, see e.g. (Goldblatt; 1992, p. 111).

$$[\alpha \cup \beta]\phi \leftrightarrow [\alpha]\phi \wedge [\beta]\phi$$

This is valid because the relation for the modality $[\alpha \cup \beta]$ is the union of the two relations for $[\alpha]$ and $[\beta]$. Since $\alpha \cup \beta$ is read 'do either α or β', it looks a lot like a free choice inference. The sentence may be read 'after you do either α or β, ϕ is true, if and only if, after you do α, ϕ is true and after you do β, ϕ is true.

In the papers mentioned above variations of dynamic logic are developed to handle deontic inferences. Even though these logics agree with the one presented here that modal operators should be applied to action types, Ross' paradox is not blocked and free choice inferences are not allowed in the first paper mentioned. I will not treat it any further. In van der Meyden (1996), however, Ross' paradox is blocked and free choice inferences are allowed. Broadly speaking, van der Meyden's logic is in the tradition of strong permission. An execution of an action is represented as a sequence of states. An action kind or type is a set of such sequences. Some of these sequences or executions are *green* or permitted. A benefit of this framework is of course that one may speak of sequential actions, a topic not treated in this thesis. However, in other ways the expressivity of the logic is limited. It is not clear how to talk about doing actions of different kinds simultaneously. This problem is related to the objection to the truth condition of the free choice permission operator stated below. A free choice permission of an action a holds, when all executions of a are permitted. Obligation is treated as a dual to a (traditional) weak permission, giving us some of the problems of standard deontic logic for these operators, including Ross' paradox. Thus, we have the problem of two different kinds of permissions, without a well-founded conceptual distinction between the two. A simple objection is that empty actions (with no possible executions) are always permitted, but as van der Meyden points out this can be remedied by a non-emptiness condition. However, there is

a more serious objection, which is that the universal style truth condition is too strong to model permission. The problem is not so easy to detect in van der Meyden's logic because it is not clear how to represent simultaneous actions, or as we might say, action tokens (sequences) instantiating several action types. Intuitively, though, an execution s is of two types α and β if it is an element in both sets representing the types. However, if that is a correct way of seeing it, the following problem follows.[4]

There are many different ways an agent may eat a pear. With some of these action executions the agent also kills a person, since the agent eats a pear and kills the person at the same time. However, (we may assume) it is not permitted for the agent to kill a person. Hence this particular action execution cannot be green. According to van der Meyden's semantics, it is not permitted for the agent to eat a pear!

This seems very counterintuitive, but it is a natural consequence of the requirement that *every single execution of an action* must be green, for the action to be permitted. In view of the discussion about action tokens and action types in the following section, it seems ontologically correct to say that a single action token may instantiate several action types. Further, it should be enough that *some* ways of performing an action type is enough to say that this action type is permitted. However, this *existential* nature of permission is blocked in van der Meyden's semantics. As van der Meyden points out, Meyer's logics do not have this problem, as he has an existential style truth condition for permission. However, in view of the above, it also seems correct to keep van der Meyden's intuition that free choice permissions should be allowed and Ross' paradox blocked.

Another more recent body of work is in the tradition of dynamic epistemic logic, see van Benthem and Liu (2007). Here the focus is on changing the preference relation between worlds dynamically with normative utterances, see also Yamada (2006). The implications for deontic logic are sketched, but since the focus is on performative aspects of deontic sentences, I will not treat it further.

[4] This objection to this kind of semantics, to be precise a similar proposal of my own, as well as the following example was presented to me by Frank Veltman.

A third body of work starts with the work on dynamics in logical linguistics, see Veltman (1996). As with the second body of work, the focus is on the context changing potential of deontic sentences. Paul Portner writes:

> ... there seems to be no linguistically-oriented discussion of dynamic properties of deontic modals... (Portner; 2009, p. 12)

From a general conceptual perspective, however, the paper van der Torre and Tan (1999) does approach the issues. Further, Nauze's thesis, Nauze (2008), fills the gap mentioned by Portner in giving deontic modalities a linguistically-oriented treatment. The main focus of this work is on performative aspects of deontic language and it is thus also beyond the scope of this thesis.

7.5 Other related approaches

In the dynamic deontic logics discussed above, as in dynamic logic in general, there are two symbols for disjunction, one for between sentences and one for between actions. Thus natural language *or* gets two different formal meanings. This is also the approach taken in this chapter. However, it has also been suggested that disjunction *always* should be interpreted to allow the free choice inference. In Zimmermann (2000) a disjunction is taken to be a list of epistemic possibilities.

The idea to use lists is close to the one presented below, where complex actions are considered unordered tuples. An agent may do A_1 or... or A_n, iff the agent may do all the things on the list $A_1, \ldots, A_n$.

However, Zimmermann acknowledges that the theory has a problem with deontic *must* as in the following sentence.

Example 7.6. 'Mr. X must take a taxi or a bus.'

The only reading his theory can get of the above implies that either Mr. X must take a taxi or Mr. X must take a bus. This is clearly wrong. The above means (interpreted deontically) that there are two options, each permitted, but that Mr. X must choose *one of them*. The sentence should imply that Mr. X may do either, but *not* that he must do one of them. This is the case in the logic to be presented. An important concern of Zimmerman's is that free choice inferences may be blocked as in the following sentence.

Example 7.7. 'Detectives may go by bus or boat - but I forget which.'

According to Zimmermann and others, see e.g. Asher and Bonevac (2005), this sentence does not imply that detectives may go by bus and may go by boat, but only that they may do one of them. I conjecture that the disjunction in this sentence is under the scope of a hidden epistemic operator, so that the sentence could be paraphrased as follows.

Example 7.8. 'I know that detectives may go by bus or that they may go by boat, but I forget which.'

If that is the case, it is not strange that there is no free choice effect. However, nothing in this chapter depends on this conjecture, because of the following. In order to be true the sentence requires an epistemic uncertainty about the permitted actions, which it is beyond the scope of this thesis to cover. If we assume (as we have so far and shall again in the next chapter) that agents know their own permitted actions and the permitted actions of other agents, sentences such as the one above will always come out false.

Another closely related theory was developed in Jackson (1985). In that paper, an action is always a member of a set of actions, the available alternative actions. The following truth condition of 'it ought to be that A' is given, where A is an action.

Definition 7.5. It ought to be that A out of $\{A, A_1, \ldots, A_n\}$, iff. what would be the case were A true is better than what would be the case were A_i true, for all $1 \leq i \leq n$.

Jackson's main intuition that there are a set of available action tokens some of which are permitted is also the main intuition behind the theory presented here. The main difference between Jackson's approach and the one taken here, is perhaps the explicit distinction between action types and action tokens introduced below. On Jackson's proposal, deontic operators are applied to action tokens, whereas they will be applied to action types in the logic to follow.

7.6 Action types and action tokens

According to von Wright norms apply to *action types* or as he calls them, *acts*, which he takes to cover *act qualifying properties*.

> We shall say that theft, murder smoking, etc. are acts. The individual cases that fall under theft, murder, smoking, etc. we

> shall call act-individuals. It is of acts and not of act-individuals that deontic words are predicated.
> (von Wright; 1951, p. 2)

Peter Geach has criticized deontic logic from moving away from actions and towards applying deontic operators to formulas expressing arbitrary propositions, see Geach (1982). Stit theory is a rather successful attempt at taking Geach' s objections seriously, see also (Horty; 2001, p. 4). Horty's distinction between *ought to do* and *ought to be* is very illuminating. However, stit theory in general has been criticized for not representing actions directly.

> No author in the Anselm-Kanger-Chellas line up through Belnap - Davidson belongs to a different tradition - has countenanced the existence of actions in logic: action talk, yes; ontology of actions, no.
> (Lindström and Segerberg; 2007, p. 1199)

The authors explicitly say that this remark does not apply to Horty. However, even Horty himself states that he has no way of representing action types. Thus stit theory at most appears to be suited for treating *action tokens*, cf. Horty:

> These actions are only action tokens, however - individual concrete actions. There is no such thing as the action type of "opening a window", for example. There are individual, concrete openings of individual windows, but nothing to group them together.
> Horty (2001)

In the philosophy of action the view of treating actions as objects is especially associated with Donald Davidson and with the event calculus. As such it is sometimes considered opposed to the modal view on action and agency advocated by stit theory. In my view this contrast is just a matter perspective. We can take different perspectives on the same thing at different times. In this chapter I take the perspective of treating actions as objects. In general, talk about action tokens and action types is a typical example of our mental capacity for reification through abstraction. The distinction between action tokens and action types seems quite natural. There are many different particular ways an agent may brush her teeth, but all of them are of

the 'teeth brushing kind.' When actions are reified in this way, they can be treated like objects with several properties. An action token may instantiate several types, e.g. the same action token may instantiate 'going to the beach', 'going for a run', 'getting some exercise', 'meeting a friend', and so on. Some other references for this view are, Ross (1930), Hage and Brouwer (2000). Hage and Brouwer elaborate on the relation between action types and action tokens in the following way (they call the latter 'act tokens').

> Deontic sentences of the ought-to-do type typically refer to action types. If James is not allowed to shoot John Doe, there is not one particular act that James is not allowed to perform. It is rather the case that James is not allowed to perform any act of the type 'shooting John Doe'. The temptation might arise to say that the action type 'shooting John Doe' is the object of the prohibition. But this would be wrong. James is forbidden to perform individual acts, not to perform action types. Assuming that action types are forbidden would involve a category mistake. Actors do not perform action types, but act tokens. Nevertheless, the prohibition refers to the type. This does not mean that the type is forbidden, but rather that acts that are instances of this type, are pro tanto forbidden.
> Hage and Brouwer (2000)

At another place the authors write the following.

> It is not impossible to say that an act token was obligatory or forbidden, but then the natural meaning is that the token is an instance of a type which was (for the actor) obligatory, allowed, or forbidden.
> Hage and Brouwer (2000)

It seems that the authors waver a bit between, whether it is action tokens or action types that are forbidden. On a perhaps uncharitable reading they say:

1. Action types cannot be forbidden.
2. The natural meaning of an action token being forbidden is to be an instance of a forbidden action type.

I will now state informally, how action types and action tokens are connected to deontic operators in this thesis. Von Wright's basic intuition is followed, since deontic operators only apply to action types. However, unlike von Wright's theory, this is not a primitive syntactic relationship between operators and types. Rather the truth conditions of deontic sentences are given via the good or acceptable action tokens instantiating these types. An *action token* is a single object. However, on this theory there are also *complex actions*. The basic idea is that the complex actions should be constructed from the action tokens. Here, they are unordered tuples of such action tokens. There is a countable set of basic action types such as 'having a drink' and 'eating a pear.' Basic types, and inductively defined conjunctions and negations of these, are instantiated by single action tokens. Thus, for instance, action types such as 'not going to the beach' or 'eating an apple and going skiing' may only be instantiated by single action tokens (only simultaneous actions are considered, not sequential actions of the type 'eat an apple and *then* go skiing.') On the other hand, actions of the disjunctive type 'going to the beach or having a drink' may also be instantiated by (in this case 2-) n-tuples of action tokens, so that there is one action token of the type 'going to the beach' and one action token of the type 'having a drink.' This is why, (when talking about simultaneous action) it makes sense that an action may be of the type of 'going to the beach or going to the cinema', because different action tokens may instantiate either disjunct, but not of the type 'going to the beach and going to the cinema', given that there is no single action instantiating both of these types at once. This idea that the basic parts of action disjunctions are simultaneously instantiated by what may be different action tokens is what makes the disjunction behave in many ways like a conjunction in sentences about actions and norms. This assumption about the relation between actions and the types they instantiate is very substantial and may be disputed. However, if or when this is granted, the rest of the material follows rather naturally, given one more substantial assumption. The other substantial assumption is that there is always a non-empty set of *good* or *acceptable* n-tuples of action tokens for any arity n. On purpose, I say nothing about where this set comes from, as the very general framework presented here may be applied to specific ways of arriving at the good. It is now possible to define operators **may** and **must** on action *types* as follows. **must**[t] holds if and only if, *every* good action is of type [t]. **may**[t] holds if and only if *some* good action is of type [t].

To sum up, all possibilities for developing a deontic logic for action types have not been explored, especially not in connection with the stit framework. The focus of this chapter is mainly to present such a deontic logic and to show how well it deals with the deontic inferences presented in the introduction. In the next chapter I apply the logic to the situation models which have been considered in the rest of this thesis. The purpose of that chapter is to show that the theory merges with the stit framework in a natural and useful way. I will now go into more detail with the logic.

7.7 Logic

I am going to present a syntax and a semantics for deontic operators on action types.

7.7.1 Syntax

Symbols

The logic of action types that will be presented here is meant to be applied to reasoning about actions and norms. It is generated from a countable set of basic action types or action predicates , $T = (T_1, T_2 \ldots)$. We also have the usual boolean connectives $\vee, \neg, \rightarrow$ and so on, as well as the complex action connectives $\cup$, $\cap$, $-$ (for action disjunction, conjunction and negation), left and round and square parentheses, and deontic operators **may** and **must**

Action type terms

An expression of the language is just a string of symbols. We are mainly interested in well-formed action type terms and well-formed formulas. We first define the set of *single action type terms* $SATT$. These will represent the action types instantiated by *single* action tokens. These consist of the basic action type terms, the conjunctive action type terms and the negated action type terms.

Definition 7.6. $SATT$ is the smallest set satisfying the following conditions:

1. For any $T_i \in T$, $[T_i] \in SATT$ (basic action type terms BATT).
2. If $[t], [s] \in SATT$, then $[(t \cap s)] \in SATT$.

3. If $[t] \in SATT$, then $[-t] \in SATT$.

The *complex action type terms* will represent action types instantiated by complex actions. These consist only of the *disjunctive action type terms* $DATT$.

Definition 7.7. $DATT$ is the smallest set obeying the following condition:

$$\text{If } [t_1], \ldots, [t_n] \in SATT, \text{ then } [t_1 \cup \ldots \cup t_n] \in DATT.$$

The well-formed action type terms are just the members of $DATT$ (note that $SATT \subseteq DATT$, set $n = 1$ in the definition above). The *arity* of an action type term is defined as (the number of $\cup$ occurring in it)+1. Thus, e.g. $\text{arity}([(T_1 \cap -T_2) \cup T_2]) = 2$. The main thing to notice is that action type terms are constructed on two levels. On the first level, we construct basic action type terms, as well as negations and conjunctions of these basic action type terms. On the second level we construct disjunctions of these.

Formulas

The set of well-formed formulas WFF is defined as follows.

Definition 7.8. WFF is the smallest set obeying the following conditions.

1. If $[t] \in DATT$ then $\mathbf{may}[t] \in WFF$ and $\mathbf{must}[t] \in WFF$.
2. If $\phi, \psi \in WFF$, then $\neg\phi, (\phi \wedge \psi) \in WFF$

The rest of the propositional connectives are defined as usual.

7.7.2 Semantics

We take action tokens as primitive objects in our ontology and we consider only the actions of a single agent.

Models

An *action type model* is a structure $M = \langle W, \{good^i \mid 1 \leq i, i \in \mathbb{N}\}, V\rangle$, where W (also written $dom(M)$) is a countable, non-empty set called the domain of M. Informally, the domain is the set of available action tokens of an agent. V is a function which gives to each basic action type term, $[T_i]$, with $T_i \in T$, a subset of the domain, i.e. $V : BATT \to \mathrm{P}(W)$. $good^1 \subseteq W$ is a non-empty set of action tokens, which we call the *basic good actions*. $good^i$ is defined as follows for each $i > 1$.

Definition 7.9. $good^i$ is a nonempty set of unordered tuples $(a_1, \ldots, a_i)$, such that $a_1 \in good^1, \ldots a_i \in good^1$.

The following property is a consequence of this definition.

Grounded in basic good actions $(a_1, \ldots, a_i) \in good^i$ implies $a_1 \in good^1, \ldots, a_i \in good^1$.

We also require the following property to hold.

Complete with respect to basic good actions For each $n-tuple\ a_1 \in good^1, \ldots, a_n \in good^1$ of *distinct* basic good actions, (i.e. $i \neq j$ implies $a_i \neq a_j$ for $1 \leq i \leq n, 1 \leq j \leq n$), $(a_1, \ldots, a_n) \in good^n$.

V is extended to arbitrary action types of any arity as follows. Note that the value of the valuation of an action type of arity n higher than 1 is a set of unordered tuples of length n: the tuples made from tokens of the type of the disjuncts.

Definition 7.10.
1. $V([(t \cap s)]) = V([s]) \cap V([t])$.
2. $V([-s]) = W \backslash V([s])$.
3. $V([t_1 \cup \ldots \cup t_n]) = \{(a_1, \ldots, a_n) \mid a_1 \in V([t_1]), \ldots, a_n \in V([t_n])\}$

In the definition above $(a_1, \ldots, a_n)$ is an unordered n-tuple. In the metalanguage, we sometimes write $[t]a$ for $a \in V([t])$ with the intuitive meaning that a is of type $[t]$.
For single action type terms, $[t], [s]$, the following hold.

Fact 7.11.
1. $[-t]a$ iff not $[t]a$.
2. $[t \cap s]a$ iff $[t]a$ and $[s]a$.
3. $[t \cup s](a, b)$ iff ($[t]a$ and $[s]b$) or ($[t]b$ and $[s]a$) .

Evaluation of formulas

The truth of a formula in an action type model M is defined as follows.

Definition 7.12. Let $[t]$ be an action type of arity n

1. $M \vDash \mathbf{must}[t]$ iff for any $a \in good^n$, $[t]a$.
2. $M \vDash \mathbf{may}[t]$ iff for some $a \in good^n$, $[t]a$.
3. $M \vDash \neg\phi$ iff $M \nvDash \phi$.
4. $M \vDash \phi \wedge \psi$ iff $M \vDash \phi$ and $M \vDash \psi$.

Validity and logical consequence are defined as follows.

Definition 7.13. A sentence is valid if it is true in all models. A sentence is satisfiable if it is true in some model. We also say that a set Γ of sentences is true in a model, denoted $M \vDash \Gamma$, if for every $\phi \in \Gamma$, $M \vDash \phi$. We say that a set of sentences Γ is (jointly) satisfiable, if there is a model M, such that $M \vDash \Gamma$. A sentence ϕ is a logical consequence of a set of sentences Γ, if for any model M such that $M \vDash \Gamma$, we have $M \vDash \phi$.

Consider the following informal argument. At all times you wish to perform one of the (legally or morally) acceptable actions available to you. If one of these actions instantiates the type of, say, eating an apple, then you may eat an apple. On the other hand, if you may eat an apple, then there has to be an acceptable action of that type. If all of the acceptable actions are of a certain type, e.g. of not-killing, then you must not-kill. On the other hand, if you must not-kill, then it cannot be that there is an acceptable action of the killing type (we are thus not considering defeasible norms.) These sort of considerations underlie the definitions above.

7.7.3 Validities

Here are some validities for this logic.

Fact 7.14. 1. $\vDash \mathbf{must}[c \cup d] \rightarrow (\mathbf{may}[c] \wedge \mathbf{may}[d])$.

2. $\vDash \mathbf{must}[t] \rightarrow \mathbf{may}[t]$.

3. $\vDash \mathbf{may}[c \cup d] \rightarrow (\mathbf{may}[c] \wedge \mathbf{may}[d])$.

4. $\vDash (\mathbf{may}[c \cap -d] \wedge \mathbf{may}[d \cap -c]) \rightarrow \mathbf{may}[c \cup d]$.

5. $\vDash \mathbf{may}[(c \cap d)] \rightarrow (\mathbf{may}[c] \wedge \mathbf{may}[d])$.

6. $\vDash \mathbf{must}[(c \cap d)] \rightarrow (\mathbf{must}[c] \wedge \mathbf{must}[d])$.

7. $\vDash (\mathbf{must}[c] \wedge \mathbf{must}[d]) \rightarrow \mathbf{must}[(c \cap d)]$.

8. $\vDash \mathbf{may}[c] \vee \mathbf{may}[-c]$.

9. $\vDash \neg(\mathbf{must}[c] \wedge \mathbf{must}[-c])$.

Proof. 1. Follows from 2. and 3. by propositional logic.
2. Assume $M \vDash \mathbf{must}[t]$, where t has arity n. Then for any $(a_1, \ldots a_n) \in good^n$, $(a_1, \ldots a_n) \in V([t])$, and since $good^n \neq \emptyset$, $M \vDash \mathbf{may}[t]$.
3. Assume $M \vDash \mathbf{may}[c \cup d]$. Then there are $(a, b) \in good^2$, s.t.$[c \cup d](a, b)$. Hence $[c]a$ and $[d]b$. By groundedness, $(a, b) \in good^2$ implies $a, b \in good^1$, so $M \vDash \mathbf{may}[c] \wedge \mathbf{may}[d]$.
4. Assume $M \vDash (\mathbf{may}[c \cap -d] \wedge \mathbf{may}[d \cap -c])$. Then there are $a, b \in good^1$, s.t. $[c]a$ and $\neg[d]a$ and $[d]b$ and $\neg[c]b$. It follows that $a \neq b$ (otherwise $[c]a$ and $\neg[c]a$), so $(a, b) \in good^2$ by completeness. Since $[c]a$ and $[d]b$, $[c \cup d](a, b)$, so $M \vDash \mathbf{may}[c \cup d]$.
The rest of the validities can be verified in a similar way. □

7.7.4 Non-validities

Here are some non-validities.

Fact 7.15. 1. $\nvDash \mathbf{must}[t] \rightarrow \mathbf{must}[t \cup s]$

2. $\nvDash \mathbf{may}[t] \rightarrow \mathbf{may}[t \cup s]$.

3. $\nvDash \mathbf{must}[t \cup s] \rightarrow (\mathbf{must}[t] \vee \mathbf{must}[s])$.

4. $\nvDash (\mathbf{may}[c] \wedge \mathbf{may}[d]) \rightarrow \mathbf{may}[c \cap d]$.

5. $\nvDash \mathbf{may}[c \cup -c] \rightarrow \mathbf{may}[t]$.

Proof. Let $W = \{a, b, c\}$, $good^1 = \{a, c\}$, $good^2 = \{(a, c), (c, a)\}$, and $good^n = \{$all n-tuples of $good^1\}$, for $n > 2$. Let $V(T_1) = \{a, c\}, V(T_2) = \{b\}, V(T_3) = \{a\}$, $V(T_4) = \{c\}$.
1. + 2. We have $M \vDash \mathbf{must}[T_1]$, and $M \vDash \mathbf{may}[T_1]$, but neither $M \vDash \mathbf{must}[T_1 \cup T_2]$, or $M \vDash \mathbf{may}[T_1 \cup T_2]$.
3. We have $M \vDash \mathbf{must}[T_3 \cup T_4]$, but neither $M \vDash \mathbf{must}[T_3]$ or $M \vDash \mathbf{must}[T_4]$.
4. We have $M \vDash \mathbf{may}[T_3] \wedge \mathbf{may}[T_4]$, but not $M \vDash \mathbf{may}[(T_3 \cap T_4)]$.
5. We have $M \vDash \mathbf{may}[T_3 \cup \neg T_3]$, but not $M \vDash \mathbf{may}[T_2]$. □

7.7.5 Equivalences

Finally, here are some logical equivalences.

Fact 7.16. 1. $\vDash \mathbf{must}[c] \leftrightarrow \neg\mathbf{may}[-c]$.

2. $\vDash \mathbf{may}[c] \leftrightarrow \neg\mathbf{must}[-c]$.

3. $\mathbf{may}[(d \cap c)] \leftrightarrow \mathbf{may}[(c \cap d)]$.

4. $\mathbf{may}[s \cup t] \leftrightarrow \mathbf{may}[t \cup s]$.

5. $\mathbf{must}[(d \cap c)] \leftrightarrow \mathbf{must}[(c \cap d)]$.

6. $\mathbf{must}[s \cup t] \leftrightarrow \mathbf{must}[t \cup s]$.

Proof. 1. Left to right. Assume $M \vDash \mathbf{must}[c]$ and $M \vDash \mathbf{may}[-c]$. Then there is $a \in good^1$, s.t. $a \in W \setminus V([c])$, but since $a \in good^1$, $a \in V([c])$, contradiction.
Right to left. Assume $M \vDash \neg\mathbf{may}[-c]$. So for no $a \in good^1$, $a \in W \setminus V([c])$. Hence for all $a \in good^1$, $a \in V([c])$, so $M \vDash \mathbf{must}[c]$.
2. is similar to 1.
3. $M \vDash \mathbf{may}[(d \cap c)]$ iff. there is $a \in good^1$, s.t. $a \in V([d])$ and $a \in V([c])$ iff. there is $a \in good^1$, s.t. $a \in V([c])$ and $a \in V([d])$ iff. $M \vDash \mathbf{may}[(c \cap d)]$.
The rest of the equivalences are similar. □

7.8 How intuitively adequate is the logic?

As can be seen from the previous section the logic gets the consequences of Figure 7.1 right with one exception. It gets all of the non-consequences of Figure 7.2 and all of the equivalences of Figure 7.3 right. The exception is the

following. The fourth of the validities above, $(\mathbf{may}[c \cap -d] \wedge \mathbf{may}[d \cap -c]) \rightarrow \mathbf{may}[c \cup d]$, suggests that we only have the fourth of the consequences in Figure 7.1 in a restricted form. We would expect the consequence, $(\mathbf{may}[c] \wedge \mathbf{may}[d]) \rightarrow \mathbf{may}[c \cup d]$. The reason that we do not have this is that we are not guaranteed that any $2-tuple$ of elements of $good^1$ are in $good^2$. The condition $(\mathbf{may}[c \cap -d] \wedge \mathbf{may}[d \cap -c])$ guarantees that there are distinct good actions of type $[c]$ and $[d]$, and thus the conclusion follows, by the completeness condition. Requiring that all $n-tuples$ of the basic good actions are in $good^n$ is too strong. From this e.g. $\mathbf{must}[c \cup d]$ would imply $\mathbf{must}[c] \wedge \mathbf{must}[d]$. From e.g. 'you must marry or flee the country' it would follow that 'you must marry and you must flee the country', which is clearly not intuitive.

Thus, as far as I can see, we can only get the above restricted form and there is still room for improvement of this logic, at least in this respect.

7.9 Natural language and expressivity

The **must** operator is meant to represent a strong deontic modality, equivalent to *have to*, but weaker than *ought to*. As in von Wright's original system (and in natural language, except for pragmatic emphasis, e.g. 'you must, must, must come tonight!') iterations of **must** do not occur. A more serious limitation on expressivity is that we can only represent conditionals with our operators in the following ways.

1. $\mathbf{must}[t] \rightarrow \mathbf{must}[s]$.
2. $\mathbf{must}[t] \rightarrow \phi$.
3. $\phi \rightarrow \mathbf{must}[s]$.

1. is a special case of 2. and 3. An obvious criticism is the lack of ability to express directly some of the following sentences, where a conditional is within the scope of a must.

1. 'You must, if it rains, use an umbrella.'
2. 'You must, if it rains, *not* use an umbrella.'
3. 'You *must* not, if it rains, use an umbrella.'

The following is a list of suggestions as to how we may first transform these particular sentences before we perform the translation into the formal language, which follows.

1. 'If it rains, you must use an umbrella.' $\phi \rightarrow \mathbf{must}[s]$.
2. 'If it rains, you must *not* use an umbrella.' $\phi \rightarrow \mathbf{must}[-s]$.
3. 'If it rains, you *must* not use an umbrella.' $\phi \rightarrow \neg\mathbf{must}[s]$.

Arguably, the second set of sentences is more natural than the first set. If this is so, it strengthens the appeal of the formalizations given here. In the second sentence the scope of the negation is 'use an umbrella', i.e. it is an internal action type negation, in the third sentence the scope of the negation is 'you must use an umbrella.' In spoken natural language we often use emphatic intonation to indicate this scope distinction, here indicated *must* and *not*. These considerations are similar for **may**. The two-fold reason for the lack of expressivity is the following.

1. I wanted to apply **must** and **may** to action types only.
2. I do not believe that there are conditional action types, types of the form of 'if running, then wearing shoes', do not seem to make any intuitive sense, and I have refrained from giving them an artificial formal meaning.

The first of these reasons, although sometimes longed for in deontic logic, does not match up with natural language. Also, given that I do think **must** and **may** can be applied to conditionals, but that I do not think action types can be conditional, this assumption would be the one to relax. For instance, an obvious exception to 1. seems to be *de re* norms about objects or persons as in the following sentence.

Example 7.9.

The president of the united states must be 31 years old, when elected for office.

Another obvious criticism concerns contrary-to-duty obligations and defeasible obligations. Clearly, the semantics is not fit to deal with these cases. There is hope, though that we can combine the logic presented here with

techniques already known from deontic logic. Such investigations might also shed light on W.D. Ross' famous theory on prima facie oughts vs. all out norms, which he defined exactly in terms of action *types*.

A third rather obvious criticism is that the natural language word 'or' has two distinct formal meanings on this theory, one as it occurs between sentences and one as it occurs in action types. This is the price we pay for keeping classical logic on the sentence level and getting a more intuitively correct logic for permissions and obligations as applied to action types.

Despite these limitations, the logic of norms applied to action types presented in this chapter solve a good deal of the problems found in the literature on deontic logic in a natural way. No doubt it has other problems of its own, which will come to light eventually but such is life in logic. I leave the integration of the present logic with logics for defeasible and contrary-to-duty obligations for somebody else. Also, I leave the problem of providing the logic with a complete proof system.

Chapter 8

Action types in strategic situations

In this chapter some connections are established between the deontic logic for action types from the previous chapter and the utilitarian strategic models presented in the rest of the thesis. This will make it apparent how to use the deontic logic for action types. Further, it will make it possible to compare this logic with the logic used elsewhere in the thesis. Finally, it will enable us to apply this logic to a problem concerning intentions encountered in Chapter 3. The basic idea is to let the actions of agents in utilitarian strategic models with intentions play the role of action tokens. The complex actions will then be unordered tuples of these. The basic good actions of an agent is her optimal actions.
The main difference between the logic in this chapter and in the previous chapter is that the operators will be relativized to an agent to allow for several agents. Also an ability *can* operator is added to the logic.

8.1 Syntax

I will be rather informal. To the symbols used in Chapter 2 we add square brackets and the action negation, conjunction and disjunction signs, as well as three operators $\mathbf{may}_a$, $\mathbf{must}_a$, and $\mathbf{can}_a$ for each $a \in Agent$. The basic action types T is a non-empty subset of the propositional variables, $T \subseteq \Phi$. The well formed action type terms are defined as in the previous chapter.

Definition 8.1. The well-formed formulas WFF is the smallest set such that:

1. If $[t] \in DATT$ then $\mathbf{may}_i[t]$, $\mathbf{must}_i[t], \mathbf{can}_i[t] \in WFF$, $i \in \mathbb{N}$.
2. If $\phi, \psi \in WFF$, then $\neg\phi, (\phi \wedge \psi) \in WFF$ as well as a well-formed formula prefixed by one of the modal operators used elsewhere in the thesis.

8.2 Semantics

A *utilitarian strategic action type model* , M, is a structure $\langle M_S, Choice$, for each $a_i \in Agent$ a set $\{good_i^n \mid 1 \leq n, n \in \mathbb{N}\}, V_T\rangle$, where:[1]

[1] For some applications it might easier just to take $good_i^1$ as primitive, as a non-empty subset of the actions of the agent, but here the goal is to emphasize the connection to stit

1. M_S is a utilitarian strategic model with intentions.
2. $Choice = \bigcup_{a_i \in Agent} Choice_i$.
3. $V_T : BATT \rightarrow \mathcal{P}(Choice)$.
4. $good_i^1 = Optimus_i'$, for each $a_i \in Agent$.
5. $good_i^n$ for $n > 1$ is defined as in the previous chapter for each $a_i \in Agent$ and to fulfill the condition that it is complete with respect to the basic actions.

Note that since T is just a subset of the propositional variables of the model these will be used in two ways: they represent propositions in the model and (enclosed in square brackets) they represent action types.

8.2.1 Evaluation of formulas

In order to establish connections with other modal operators, all formulas will be evaluated with outcomes of situations. Conceptually this will not make a significant difference, as we shall see. All the propositional operators and the modal operators are evaluated as elsewhere in the thesis. The new operators get the following truth conditions.

Definition 8.2. Let $[t]$ be an action type of arity n.

1. $M, o \vDash \mathbf{must}_i[t]$ iff for each $A \in good_i^n$, $A \in V_T([t])$.
2. $M, o \vDash \mathbf{may}_i[t]$ iff for some action $A \in good_i^n$, $A \in V_T([t])$.
3. $M, o \vDash \mathbf{can}_i[t]$ iff for some unordered tuple $A = (A_{i1}, \ldots, A_{in})$, such that each $A_{ij} \in Choice_i$ for $1 \leq j \leq n$, $A \in V_T([t])$.

The above operators are settled in the situation, i.e. either they are true with all outcomes or false with all outcomes. This is why it does not make a conceptual difference whether they are evaluated with an outcome or in the model as a whole. The motivation for the deontic operators is the same as in the previous chapter. For example, an agent may perform an action of a certain type in a situation if and only if there is a good action of that type

models.

available to her in the situation. The motivation for the ability *can* operator $\mathbf{can}_i$ is as follows. An agent can t in a situation if and only if, there is an action of type t available to her in the situation. Thus the truth condition for ability *can* is just like the truth condition for deontic *may* except there is no normative requirement on ability *can*. We immediately get the validities $\mathbf{must}_i[t] \to \mathbf{can}_i[t]$ and $\mathbf{may}_i[t] \to \mathbf{can}_i[t]$.

So far there is no real connection between actions and the outcomes of actions. This connection will now be established. There are two components to the connection. The *corresponding proposition* of an action establishes a general connection between an action type and what the action achieves in a situation. A *principle of intentions* gives a necessary condition of how an action of a specific type is related to the intended outcomes of that action: the corresponding proposition must be true with the intended outcomes.

8.3 Action types and corresponding propositions

One criterion for identifying an *action* is that it is done for a *reason*. Davidson defines an intentional action as one done for a reason, Davidson (1963). Here the reason for an action will be identified with the goal that the agent wishes to achieve with the action, cf. Neth and Müller (2008). Thus it will be assumed that to each action type there corresponds a certain state of affairs in the world. For instance, to the action type 'close the door' corresponds the state of affairs or proposition that the door is closed. How are the action types and their corresponding propositions connected to the outcomes of actions? On the one hand it is too strong to require that the corresponding proposition is true with every possible outcome of that action. After all, an agent may fail to achieve the goal of the action. The previous sentence, however, gives a clue as to with which outcomes it *is* natural to require that the corresponding proposition is true. The goal will at least be achieved with the intended outcomes of the action. If an agent closes the door, then the door is closed with the intended outcomes of that action. This minimal condition will be required fulfilled in the following. To make it explicit let us first be precise about what the corresponding proposition of an action is. We already have the tools built into the syntax. All that remains is to connect the action types to the propositional formulas. This will be done via the following translation, $corr$.

Definition 8.3. 1. For a basic type $T_i \in T$, $corr([T_i]) = T_i$.

2. $corr([-t]) = \neg corr([t])$.

3. $corr([(t \cap s)] = corr([t]) \wedge corr([s])$.

4. $corr([t_1 \cup \ldots \cup t_n]) = corr([t_1]) \vee \ldots corr([t_n])$.

Thus, for instance, $corr([(T_1 \cap -T_2) \cup T_3]) = (T_1 \wedge \neg T_2) \vee T_3$. We call $corr([t])$ the corresponding proposition of the type $[t]$.

8.4 Intentions and action types

We are going to assume the following *intention principle*, which connects the action types with the intended outcomes of the action tokens of that type through the corresponding proposition of the action type. The principle reflects that the corresponding proposition of an action type is the goal of the action of that type.

Principle 8.4. For an atomic action $A_{ij} \in Choice_i$ of type $[t]$, $I_{A_{ij}} \subseteq |corr([t])|_M$

In words, for an action of type $[t]$, the proposition corresponding to $[t]$ must be true with the intended outcomes of $[t]$. For instance, if an action is of the type 'close the door', 'the door is closed' will be true at least with all intended outcomes of that action.

It is easy to see that this account of action does not succumb to the particular objection put forth by the ability sceptic of Chapter 4. Recall that this objection says that the account of ability given by stit theory is too demanding. Since it is always possible to imagine circumstances under which an action does *not* succeed, no agent ever sees to anything with the stit definition of ability. However, the account of action presented in this chapter is not inconsistent with this ability scepticism. There might always be a possible outcome of each action token of each action type, with which the goal of that action type is not achieved. However, it will be required that an action type at least succeeds with the *intended outcomes* of each action token of that type.

8.4.1 The right way of eating a pear - is not killing somebody while doing it!

It is time for an example. Recall that van der Meyden's dynamic deontic logic presented in van der Meyden (1996), has the following problem. For

$o_1 : K, P, u(o_1) = 1$ $o_2 : P, u(o_2) = 2$	$o_3 : K, u(o_3) = 1$ $o_4 : u(o_4) = 2$	$o_5 : P, u(o_5) = 3$	$o_6 : u(o_6) = 3$
$A_{11} : [K], [P]$	$A_{12} : [K]$	$A_{13} : [P]$	A_{14}

a_1

Fig. 8.1: Killing and eating an pear

an action to be permitted, every possible execution of it must be permitted. However, there are obviously right ways of executing an action of a certain type and wrong ways of executing an action of the same type. The following provides an example of how the present theory may be applied to a specific situation.

Example 8.1. An agent a_1 has to decide whether to kill, $[K]$, a person or not. Simultaneously a_1 has to decide whether to eat a pear, $[P]$, or not. The agent may fail to murder the person (if she decides to do it), but the agent will succeed eating the pear (if she decides to eat it).

The situation is represented in Figure 8.1. We only represent the basic action types an action token instantiates, since the other action types of that action and the action types of complex actions can be calculated from the basic action types. Formally, let $T = \{K, P\}$ and M be the following.
$Agent = \{a_1\}$, $W = \{o_1, \ldots, o_6\}$.
$Choice_1 = \{A_{11}, \ldots, A_{14}\}$, $A_{11} = \{o_1, o_2\}$, $A_{12} = \{o_3, o_4\}$, $A_{13} = \{o_5\}$, $A_{14} = \{o_6\}$.
$I_{A_{11}} = \{o_1\}$, $I_{A_{12}} = \{o_3\}$, $I_{A_{13}} = \{o_5\}$, $I_{A_{14}} = \{o_6\}$.
$V_T([K]) = \{A_{11}, A_{12}\}$, $V_T([P]) = \{A_{11}, A_{13}\}$, $V(K) = \{o_1, o_3\}$, $V(P) = \{o_1, o_2, o_5\}$.
$u(o_1) = u(o_3) = 1$, $u(o_2) = u(o_4) = 2$, $u(o_5) = u(o_6) = 3$.

Since A_{13} and A_{14} strongly dominate the two other actions and they do not strongly dominate each other, we have $Optimus'_1 = \{A_{13}, A_{14}\} = good^1_1$. We define $good^2_1 = \{(A_{13}, A_{14})\}$, and $good^n_1$ as the set of all n-tuples of $good^1_1$ for $n > 2$. Since all actions in $good^1_1$ are of the type $[-K]$, it follows that $M \vDash \mathbf{must}_1[-K]$, the agent must not kill. We also have $M \vDash \mathbf{may}_1[P]$ and $M \vDash \mathbf{may}_1[-P]$ and $M \vDash \mathbf{may}_1[P \cup -P]$. We have that the agent *can* kill, $\mathbf{can}_1[K]$. We do not have it that the agent may kill or eat a pear, $M \nvDash \mathbf{may}_1[K \cup P]$. Some differences to the stit framework can now be

$o_1 : P, L$	$o_2 : \neg P, \neg L$
$A_{11} : [L]$	$A_{12} : [-L]$
a_1	

Fig. 8.2: Turning on the light

highlighted. First, we have $M \vDash \bigodot[a_1\ cstit]\neg K$, the agent ought to see to it that she does not kill. However, it follows that $M \vDash \bigodot[a_1\ cstit](\neg K \vee P)$, the agent ought to see to it that she does not kill or eats a pear. The deliberative stit operator suggested in the previous chapter does not help here. We have $M \vDash \bigodot[a_1\ dstit]\neg K$. Further, with outcome o_3, the person is murdered and the pear is not eaten, so we have $M \vDash \mathbf{E}(\neg(\neg K \vee P))$. Hence, $M \vDash \bigodot[a_1\ dstit](\neg K \vee P)$.

8.5 Intended and unintended consequences of actions

In Chapter 3 a problem regarding the representation of intentions was mentioned. The problem is that the intention operator $[a\ iit]$ is closed under logical consequence. Thus agents will intend the *double effects* of their actions. The present framework presents an opportunity to deal with this problem. Intuitively, given the intention principle there is a natural way of distinguishing the*intended* consequences of an action from its *unintended consequences*. We only define intended consequences of *single action types* (basic action types and their conjunctions and negations). They will be defined relative to outcomes and agents.

Definition 8.5. Let $a_i \in Agent$. ϕ is an *intended consequence* with an outcome $o \in A_{ij}$ of some action $A_{ij} \in Choice_i$ iff $A_{ij} \in V_T([t])$ and $\phi = corr([t])$.

8.5.1 Davidson's prowler example revisited

Consider Davidson's example.

Example 8.2. A man, a_1 has to either turn on the light $[L]$ or not $[-L]$. A prowler will be alerted, P, if he turns on the light.

The example is represented by Figure 8.2. Formally, let $T = \{L\}$ and let M be as follows. $Agent = \{a_1\}$, $W = \{o_1, o_2\}$, $Choice_1 = \{A_{11}, A_{12}\}$, $A_{11} =$

$\{o_1\} = I_{A_{11}}$, $A_{12} = \{o_2\} = I_{A_{12}}$, $V_T([L]) = \{A_{11}\}$, $V(L) = V(P) = \{o_1\}$.[2] This situation is represented in Figure 8.2. The outcomes where the light is turned on are exactly the outcomes where the prowler is alerted, P, so we have $|P|_M = |L|_M$. However, although alerting the prowler is a necessary consequence of turning on the light, and although, in this situation, it is the same event as turning on the light, it is *not* and intended consequence of turning on the light. It is not something the agent *does* it is just something that happens.

[2] The utilities are not relevant to this example so they are suppressed.

Summary

This thesis presents new tools and improvements of existing tools for reasoning about actions and norms. The theoretical setting of the work is the multi agent logic *stit theory*, a formal theory in the tradition of modal logic. In this thesis, stit theory is limited to cover only strategic situations. This basic framework is then extended in several ways throughout the thesis, but the two most important extensions are intentions and actions types.
Intentions are an important component of informal reasoning in ethics and law but they have not been part of stit theory so far. In this thesis intentions are represented via subsets of outcomes of individual actions. An intention operator, the *iit* operator, is added to the language.
Many natural language modalities operate on expressions denoting *action types*. Until now, there has been no way to talk about action types in stit theory. A proposal of how to do that is developed in this thesis.

In Chapter 1, some philosophical presuppositions and intuitions that have guided the theorizing in later chapters are laid out. It is discussed, e.g. how one can think about agents, situations and values. In Chapter 2, the basic formal definition of a situation is presented. The specific contribution of the chapter is a generalization of Horty's *ought to do* operator by means of the game theoretical device of simultaneously removing all dominated actions of all agents in an iterative process. In Chapter 3, intentions are added and definitions of various concepts of individual responsibility relative to outcomes of specific situations are presented. The knowledge of agents in situations is also considered. Chapters 4, 5, and 6 treat special topics related to preceding chapters. Chapter 4 takes a closer look at ability modalities and the metaphysics of agency. A cube of opposition for ability modalities is presented and different kinds of agents are defined relative to this cube. It is shown that if an omnipotent agent (called God) exists it is unique and solely responsible for everything. In Chapter 5, concepts of group responsibility and responsibility of individual members of groups are suggested. In Chapter 6, the theory is applied to a discussion of Frankfurt examples. Overdetermination of events and the differences between causal responsibility and agentive responsibility are discussed. Chapter 7 breaks with the framework provided by stit theory and used in the rest of the thesis. In this chapter, the starting point is a foundational discussion of deontic logic considering e.g. *Ross' paradox* and

free choice inferences and the role action types play in informal reasoning. A new deontic logic for action types is then presented where deontic *must* and *may* operators are applied to action type terms. In Chapter 8, it turns out that the break with the stit framework can be mended by introducing action types into stit theory. It is shown how to reason with action types using stit models and how to reason about *double effects* of actions.

Resumé

Denne afhandling giver nye redskaber og forbedrer eksisterende redskaber til at ræsonnere om handlinger og normer. Den teoretiske ramme om projektet er multiagentlogikken *stit teori*, en formel teori i den modallogiske tradition. I denne afhandling begrænses stit teorien til kun at dække strategiske situationer. Denne grundlæggende teori udvides på flere måder i løbet af afhandlingen, men de to vigtigste udvidelser er intentioner og handlingstyper.
Intentioner udgør en vigtig del af informel ræsonnering indenfor etik og jura, men de har indtil videre ikke været en del af stit teori. I denne afhandling bliver intentioner repræsenteret via delmængder af udfald af individuelle handlinger. En intentionsoperator, *iit* operatoren, tilføjes sproget.
Mange modaliteter i dagligsproget opererer på udtryk, der refererer til *handlingstyper*. Indtil nu har der ikke været mulighed for at tale om handlingstyper i stit teori. Et forslag, til hvordan man kan gøre dette, bliver udviklet i afhandlingen.

I kapitel 1 præsenteres nogle filosofiske antagelser og intuitioner, som har været ledende i forhold til det teoretiske arbejde i afhandlingen. Det diskuteres bl.a., hvordan man kan tænke om agenter, situationer og værdier. I kapitel 2 gives den grundlæggende formelle definition af en situation. Kapitlets specifikke bidrag er en generalisering af Hortys *bør gøre* operator via det spilteoretiske greb, der består i samtidigt at fjerne alle dominerede handlinger for alle agenter i en iterativ proces. I kapitel 3 tilføjes intentioner, og der gives definitioner af forskellige begreber om individelt ansvar relativt til udfald af specifikke situationer. Agenters viden behandles også. Kapitel 4, 5 og 6 behandler specielle emner relateret til de foregående kapitler. Kapitel 4 tager et nærmere kig på abilitive modaliter og metafysik angående agenter. En logisk kubus for abilitive modaliter præsenteres og forskellige slags agenter defineres relativt til denne kubus. Det vises, at hvis en omnipotent agent (kaldet Gud) eksisterer, så er den unik og alene ansvarlig for alting. I kapitel 5 foreslås begreber om ansvar for grupper og for individuelle medlemmer af grupper. I kapitel 6 anvendes teorien på Frankfurteksempler. Overdeterminering af begivenheder og forskellen mellem kausalt ansvar og agentivt ansvar diskuteres. Kapitel 7 bryder med den ramme, som er givet af stit teori, og som bruges i resten af afhandlingen. I dette kapitel er udgangspunktet en grundlagsdiskussion af deontisk logik, der blandt andet tager højde for *Ross' paradoks* og *frit*

valg slutninger, og den rolle handlingstyper spiller i informel ræsonnering. En ny deontisk logik præsenteres, hvor deontiske *must* og *may* operatorer appliceres på handlingstypetermer. I kapitel 8 viser det sig, at bruddet med stit teori kan heles ved at introducere handlingstyper i stit teori. Det vises, hvordan man kan ræsonnere med handligstyper i stitmodeller, og hvordan man kan ræsonnere om *dobbelteffekter* af handlinger.

Bibliography

Anderson, A. R. (1966). The formal analysis of normative systems, *in* N. Rescher (ed.), *The Logic of Decisions and Actions*, University of Pittsburgh Press.

Arendt, H. (1964). Personal responsibility under dictatorship, *in* J. Kohn (ed.), *Responsibility and Judgement*, Schocken Books, pp. 17–48.

Arendt, H. (1968). Collective responsibility, *in* J. Kohn (ed.), *Responsibility and Judgement*, Schocken Books, pp. 147–158.

Asher, N. and Bonevac, D. (2005). Free choice permission is strong permission, *Synthese* **145**: 303–323.

Balbiani, P., Herzig, A. and Troquard, N. (2008). Alternative axiomatics and complexity of deliberative stit theories, *Journal of Philosophical Logic* **37**: 387–406.

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

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

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

Broersen, J., Herzig, A. and Troquard, N. (2006a). From coalition logic to stit, *Electronic Notes in Theoretical Computer Science* pp. 23–35.

Broersen, J., Herzig, A. and Troquard, N. (2006b). A stit-extension of atl, *Lecture Notes in Computer Science* pp. 69–81.

Brown, M. (1988). On the logic of ability, *Journal of Philosophical Logic* **17**: 1–26.

Brown, M. (1992). Normal bimodal logics of ability and action, *Studia Logica* **51**: 519–532.

Castañeda, H.-N. (1981). The paradoxes of deontic logic: The simplest solution to all of them in one fell swoop, *in* R. Hilpinen (ed.), *New Studies in Deontic Logic*, D. Reidel Publishing Company, pp. 37–85.

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

Chisholm, R. M. (1964). Contrary-to-duty imperatives and deontic logic, *Analysis* **24**: 33–36.

Chisholm, R. M. (1970). The structure of intention, *The Journal of Philosophy* **67**: 633–647.

Davidson, D. (1963). Actions, reasons, and causes, *The Essential Davidson*, Oxford University Press, pp. 23–36.

Devlin, K. (1992). *The Joy of Sets*, second edn, Springer.

Duff, A. (1982). Intention, responsibility and double effect, *The Philosophical Quarterly* **32**: 1–16.

Ebbinghaus, H.-D., Flum, J. and Thomas, W. (1996). *Mathematical Logic*, second edn, Springer.

Fagin, R., Halpern, J. Y., Moses, Y. and Vardi, M. Y. (1995). *Reasoning about Knowledge*, The MIT Press.

Fischer, J. M. (2005a). Frankfurt-type examples and semi-compatibilism, *in* J. M. Fisher (ed.), *Free Will*, Vol. 3 of *Critical Concepts in Philosophy*, Routledge, pp. 289–298.

Fischer, J. M. (ed.) (2005b). *Free Will*, Critical Concepts in Philosophy, Routledge.

Fischer, J. M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press.

Fitting, M. and Mendelsohn, R. L. (1998). *First-Order Modal Logic*, Kluwer Academic Publishers.

Forrester, J. (1984). Gentle murder, or the adverbial samaritan, *The Journal of Philosophy* **81**: 193–197.

Frankfurt, H. (2005). What we are responsible for, *in* J. M. Fisher (ed.), *Free Will*, Vol. 3 of *Critical Concepts in Philosophy*, Routledge, pp. 280–288.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility, *The Journal of Philosophy* **66**: 829–839.

Geach, P. (1982). Whatever happened to deontic logic, *Philosophia* **11**: 1–12.

Goldblatt, R. (1992). *Logics of Time and Computation*, Lecture Notes, second edn, Center for the Study of Language and Information.

Greve, V. (2004). *Det Strafferetlige Ansvar*, Jurist- og Økonomforbundets Forlag.

Hage, J. and Brouwer, B. (2000). Action types and act tokens in deontic logic of the ought-to-do type, *DEON 00*.

Halpern, J. Y. (2003). *Reasoning about Uncertainty*, The MIT press.

Hansen, J., Pigozzi, G. and van der Torre, L. (2007). Ten philosophical problems in deontic logic, *Normative Multi-agent Systems*, Dagstuhl Seminar Proceedings.

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

Hilpinen, R. (ed.) (1971). *Deontic Logic: Introductory and Systematic Readings*, D. Reidel Publishing Company.

Hilpinen, R. (ed.) (1981). *New Studies in Deontic Logic*, D. Reidel Publishing Company.

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

Hunt, D. P. (2000). Moral responsibility and unavoidable actions, *Philosophical Studies* **97**: 195–227.

Jackson, F. (1985). On the semantics and logic of obligation, *Mind* pp. 177–195.

Jørgensen, J. (1937). Imperatives and logic, *Erkentniss* **7**: 288–296.

Kamp, H. (1973). Free choice permission, *Proceedings of the Aristotelian Society* **74**: 57–74.

Kane, R. (1998). *The Significance of Free Will*, Oxford University Press.

Kanger, S. (1971). Law and logic, *Theoria* **38**: 105–132.

Kaplan, D. (2004). Demonstratives, *Semantics*, Oxford University Press, pp. 749–798.

Kenny, A. (1976). Human abilities and dynamic modalities, *in* J. Manninen and R. Tuomela (eds), *Essays on Explanation and Understanding*, D. Reidel Publishing Company, pp. 209–232.

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

Kratzer, A. (2007). Situations in natural language semantics, *Stanford Encyclopedia of Philosophy*.

Lewis, D. (2000). Causation as influence, *The Journal of Philosophy* **97**: 182–197.

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

Lippert-Rasmussen, K. (2005). *Deontology, Responsibility, and Eguality*, Department of Media, Cognition and Communication, University of Copenhagen.

Liu, F. (2008). *Changing for the Better: Preference Dynamics and Agent Diversity*, ILLC - Dissertation Series, Institute for Logic, Language and Computation.

Locke, J. (1690). *An Essay Concerning Human Understanding*.

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

McNamara, P. (2006). Deontic logic, *Stanford Encyclopedia of Philosophy*.

Mele, A. R. (2006). *Free Will and Luck*, Oxford University Press.

Meyer, J.-J. (1988). A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic, *Notre Dame Journal of Formal Logic* **29**: 109–136.

Meyer, J.-J. and Veltman, F. (2007). Intelligent agents and common-sense reasoning, *in* P. Blackburn, J. van Benthem and F. Wolter (eds), *Handbook of Modal Logic*, Elsevier, pp. 991–1029.

Müller, T. (2005). On the formal structure of continuous action, *Lecture Notes in Computer Science* pp. 210–221.

Müller, T. (2006). A question of trust: Assessing the fulfillment of commitments in terms of strategies, *Lecture Notes in Computer Science* pp. 210–221.

Myerson, R. B. (1997). *Game Theory - Analysis of Conflict*, Harvard University Press.

Nauze, F. (2008). *Modality in Typological Perspective*, ILLC - Dissertation Series, Institute for Logic, Language and Computation.

Neth, H. and Müller, T. (2008). Thinking by doing and doing by thinking, *Proceedings of the thirtieth annual meeting of the Cognitive Science Society*.

Nute, D. (ed.) (1997). *Defeasible Deontic Logic*, Kluwer Academic Publishers.

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

Paprzycka, K. (2002). Flickers of freedom and frankfurt-style cases in the light of the new incompatibilism of the stit theory, *Journal of Philosophical Research* **27**: 553–565.

Pauly, M. (2001). *Logic for Social Software*, ILLC - Dissertation Series, Institute for Logic, Language and Computation.

Portner, P. (2009). *Modality*, Oxford University Press.

Prakken, H. and Sergot, M. (1997). Dyadic deontic logic and contrary-to-duty obligations, *in* D. Nute (ed.), *Defeasible Deontic Logic*, Kluwer Academic Publishers, pp. 223–262.

Prior, A. (1954). The paradoxes of derived obligation, *Mind* **63**: 64–65.

Rao, A. S. and Georgeff, M. P. (1997). Modelling rational agents within a bdi-architecture, *in* M. N. Huhns and M. P. Singh (eds), *Readings in Agents*, Morgan Kaufmann, pp. 317–328.

Ross, A. (1941). Imperatives and logic, *Theoria* **7**: 53–71.

Ross, W. (1930). *The Right and the Good*, Clarendon Press.

Roy, O. (2008). *Thinking before Acting*, ILLC - Dissertation Series, Institute for Logic, Language and Computation.

Segerberg, K. (1992). Getting started: Beginnings in the logic of action, *Studia Logica* **51**: 347–378.

Segerberg, K., Meyer, J.-J. and Kracht, M. (2009). The logic of action, *Stanford Encyclopedia of Philosophy*.

Selective Bibliography in the Logic of Action (1992). *Studia Logica* **51**: 579–589.

Smullyan, R. M. (1968). *First-Order Logic*, revised edition, dover 1994 edn, Springer Verlag.

Star Wars Episode IV: A New Hope (1977).

Troquard, N., Trypuz, R. and Vieu, L. (2006). Towards and ontology of agency and action: From stit to ontostit+, *Formal Ontology in Information Systems*.

van Benthem, J. (2007). Rational dynamics and epistemic logic in games, *International Game Theory Review* **9**: 13–45.

van Benthem, J. and Liu, F. (2007). Dynamic logic of preference upgrade, *Journal of Applied Non-Classical Logics* **17**: 129–155.

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

van der Meyden, R. (1996). The dynamic logic of permission, *Journal of Logic and Computation* **6**: 465–479.

van der Torre, L. W. and Tan, Y.-H. (1999). An update semantics for deontic reasoning, *in* P. McNamara and H. Prakken (eds), *Norms, Logics and Information Systems*, IOS Press.

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

van Inwagen, P. (1978). Ability and responsibility, *Philosophical Review* **87**: 201–224.

van Inwagen, P. (1983). *An Essay on Free Will*, Oxford University Press.

Veltman, F. (1996). Defaults in update semantics, *Journal of Philosophical Logic* **25**: 221–261.

von Fintel, K. and Iatridou, S. (2008). How to say ought in foreign: The composition of weak necessity modals, *in* J. Guron and J. Lecarme (eds), *Time and Modality*, Springer, pp. 115–141.

von Wright, G. H. (1951). Deontic logic, *Mind* **60**: 1–15.

Wansing, H. (2006). Tableaux for multi-agent deliberative-stit logic, *Advances in Modal Logic*, Vol. 6, College Publications.

Xu, M. (1998). Axioms for deliberative stit, *Journal of Philosophical Logic* **27**: 505–552.

Yamada, T. (2006). Acts of commanding and changing obligations, *Proceedings of the Seventh International Workshop on Computational Logic in Multi-Agent Systems (CLIMA VII)*.

Zimmermann, T. E. (2000). Free choice disjunction and epistemic possibility, *Natural Language Semantics* **8**: 255–290.

Index

Agent, 22
BATT, 108
Choice, 22
$Choice_i^M$, 29
$Choice_i^M(o)$ (function), 29
$Choice_i$, $i \in Agent$, 22
$Choice_\Gamma, \Gamma \subseteq Agent$, 22
DATT, 109
$Optimal_i$, 24
SATT, 108
Select, 22
$State_i$, 22
can$_i$ operator, 118
may operator, 111
may$_i$ operator, 118
must operator, 111
must$_i$ operator, 118
corr, 119
cstit operator, 29
  joint, 62
deliberative ought operator, 98
dstit operator, 37
  joint, 62
factors, 72
  alternatives of, 73
$good^i$, 110
iit operator, 39
influences, 73
joint Belnap/Perloff/Xu stit operator, 62
joint strict agency operator, 64
knowledge operators, 47
ought to be operator, 29
ought to be that A operator (Jackson), 104
ought to do operator, 30
strong permission operator, 99

ability
  and normal modal logic, 51
  Brown-Horty double modality analysis of, 51
  can, 53
  individual, 37
  may, 53
  might, 53
  modalities, 53
  must, 53
  scepticism, 58, 120
accessibility relation, 62
action
  complex positive, 25
  profile, 26
  basic good, 110
  removing strongly dominated, 27
  types, 102, 105
action type model, 110
action type term, 108
  arity of, 109
action types
  and corresponding propositions, 119
  and stit theory (Horty), 105
agent
  determined, 58
  essential for an event, 63
  omnipotent, 58
  undetermined, 59
  very weak, 59

allows it that operator, 36
Anderson, A.R., 99
Arendt's fallacy, 68
Arendt, H., 68

basic good actions, 110
    complete with respect to, 110
    grounded in, 110
BDI logic, 38, 41
Belnap, N., 2, 63
Benthem, J.v., 25, 102
blameworthiness, 44
    because of neglect, 44
    of attempt, 44
    of risking, 44
Brouwer, B., 106
Brown, M., 51

Chellas, B., 13, 98, 105
conjunction exploitation, 100
consequences
    intended, 122
    unintended, 122
control
    regulative, 43
corresponding proposition, 119
cube of opposition for ability modalities, 55

Davidson, D., 40, 122
deontic paradoxes, 15
dominance ordering
    strict, 24
    strong, 23
    weak, 23
double effects, 40, 122
dynamic deontic logic, 101
dynamic epistemic logic, 27, 102

event, 3, 22
    expressed by a formula, 28
    particular(Inwagen), 86
    universal(Inwagen), 87

finite choice condition, 22
flicker of freedom, 87
Frankfurt example, 79
Frankfurt, H., 77, 83
free choice inference, 97
free riders, 63

Geach, P., 105
God, 60
guilt, 42
    negative, 43
    of attempt, 43

Hage, J., 106
histories, 2
Horty, J., 2, 6, 9, 15, 17, 23, 51, 60, 63, 85, 98
    and action types, 105
    and intentions, 40

independence of agents, 4, 22
intention principle, 120, 122
intentions, 39, 120
Inwagen, P.v., 86

Jackson, F., 104
Jørgensen's dilemma, 12

Kanger, S., 46
Kenny, A., 51
knowledge
    implicit vs. explicit, 41

Lewis, D., 52, 80

Lindström, S., 105
linguistic intuitions
    appeal to, 93
Liu, F., 102
Locke, J., 78
logical consequence, 29, 111
Luhmann, N., 19

may (deontic) implies can, 119
Meyden, R.v.d., 101, 120
model
    action type, 110
    restriction to an action profile, 26
    utilitarian strategic, 22
    utilitarian strategic action type, 117
    utilitarian strategic with intentions, 39
    utilitarian strategic with intentions and knowledge, 47
must (deontic) implies can , 119
must (deontic) implies may, 97

natural language, 115
Nauze, F., 103
negative condition
    generalized, 63

ought implies can, 31, 119
outcomes, 3
    elementary, 3
    intended, 39, 120
    partitioned by choices, 4
overdetermination
    agentive, 76
    causal, 75

Paprzyscka, K. , 83
Perloff, M, 2
Perloff, M., 63
Plato, 72
Portner, P., 103
praiseworthiness
    negative, 45
    of attempt, 46
    positive, 44
preference ordering
    strict, 24
    strong, 23
    weak, 23
preventing
    individual, 38
principle
    of alternate possibilities, 77
proposition, 3
propositional dynamic logic, 101

refraining, 65
    individual, 38
    joint, 67
regulative control, 43
removing strongly dominated actions, 27
responsibility
    causal, 75
      and D. Lewis, 81
    for individual members of groups, 69
    joint, 69
    Kanger's definition, 48
    negative
      liability for risking, 42
      strict liability, 42
    positive, 42

Ross' paradox
    agentive , 37
    deontic, 96
Ross, A., 96
Roy, O., 38

Segerberg, K., 13, 105
standard deontic logic, 94
strong permission, 99
sure thing principle, 23

truth
    in a model, 29
    with the outcome of a model, 28

universal modality, 29
unordered tuple, 110
utilitarian strategic action type model, 117
utilitarian strategic model, 22

validity, 29, 111
Veltman, F., 99, 102

Wright, G.H.v., 13, 93
    and acts, 104

Xu, M., 2, 63

Zimmermann, T., 103