UiT
THE ARCTIC UNIVERSITY OF NORWAY

Faculty of Humanities, Social Sciences and Education

Formalization of the logic of appropriateness

—

Andreas Fjellstad

Master’s thesis in Master of Business Administration June 2018

Summary

This thesis presents a theoretical investigation into March and Olsen's logic of appropriateness in organizational theory using methods from philosophical logic. The topic is formalization of the logic of appropriateness, and it focuses on the following research questions:

(1) Which aspects of the logic of appropriateness should a logical-mathematical model aim at representing?

(2) Is there a general logical-mathematical framework suitable for representing the features described in the answer to (1)?

Applying the method of explication, it clarifies the relationship between the concepts involved in the logic of appropriateness, and develops one preliminary suggestion for a logical-mathematical model for the logic of appropriateness based on neighbourhood semantics. As part of the presentation of the logic of appropriateness, it also argues that the logic of appropriateness is fruitfully understood as a Weberian ideal type as opposed to, or in addition to, as empirical hypothesis.

It concludes with observations on how the model can be improved and questions for future research.

Preface

This master's thesis is part of an attempt at an interdisciplinary project involving philosophical logic and organizational theory, and is submitted for a Master of Business Administration at UiT The Arctic University of Norway.

I would like to thank the program coordinator Hanne C Gabrielsen for the opportunity to pursue this - from the perspective of the degree program - rather unusual project, my supervisor Jarle Weigård for very helpful comments on two drafts, and Knut Mikalsen for comments and fruitful discussion at the seminar presentation. In addition, I would also like to mention all the fruitful discussions on the topic with Tuukka Tanninen at the University of Helsinki over the last months.

Contents

Chapter 1

Introduction

Within organizational theory one can, following for example Røvik (2007) and Christensen et al. (2015), distinguish between two paradigms for explaining the functioning of organizations, roughly between rational and institutional explanations. Following March and Olsen (1989), it is furthermore common, at least from within the institutional perspective, to associate the rational perspective with a logic of consequences according to which decisions are made based on their expected utility. The institutional perspective, one the other hand, is associated with a logic of appropriateness which is based on "a vision of actors following internalized prescriptions of what is socially defined as normal, true, right or good, without, or in spite of, calculation of consequences and expected utility"(March and Olsen, 2011, p.690). The central insight is accordingly that agents decide on which actions to take in a situation by matching the situation with rules associated with their institutional role rather than basing the decision on the expected utility of each option.

We find various logical-mathematical models for the logic of consequences, most importantly variations of expected utility theory for the case in which only one agent is making a decision and game theory for scenarios with multiple agents. With help aid of such tools one can make formally precise arguments and analyses employing rational choice theory to explain organizational behaviour. In the case of the logic of appropriateness, however, the literature does not offer any logical-mathematical models through which claims involving the logic of appropriateness can be made precise. Since the logic of appropriateness is supposed to be a *logic* of action and an alternative approach to understanding decision-making according to March and Olsen (1989), it is arguably a theoretical disadvantage that no such models are available for the logic of appropriateness. This dissertation is a first attempt at rectifying this situation.

The overall contention behind the project to which this dissertation belongs is that we can through the development of a logical-mathematical model for the logic of appropriateness get a clearer picture of the logic of appropriateness over and above seemingly repetitive and programmatic statements by March and Olsen (1989, 1996, 1998, 2011). This, the hope is, can advance our understanding of not only the logic of appropriateness itself but also its use in theorizing within social sciences and organizational theory in particular.

This dissertation is in that regard intended merely as a preliminary investigation for a larger project which seeks to develop such a logical-mathematical model in detail. It prepares the grounds for the larger project by clarifying which features of the logic of appropriateness such a logical-mathematical model should represent, and by presenting an *Ansatz* towards a logical-mathematical model that aims at representing these features. This dissertation presents therefore an investigation into the following questions:

(1) Which aspects of the logic of appropriateness should a logical-mathematical model aim to represent?

(2) Is there a general logical-mathematical framework suitable for representing the features described in the answer to (1)?

To answer these questions, we begin in chapter 2 with a more detailed presentation of the logic of appropriateness based primarily on March and Olsen (1989) but also March and Olsen (1998) and March and Olsen (2011). Following that presentation, we elaborate in chapter 3 on a meta-scientific perspective underlying the above contention and introduce the method of explication which is employed both to clarify which aspects one should aim to represent and to evaluate a logical-mathematical model. Finally, we present in chapter 4 first a comparison between logic of consequences and logic of appropriateness within game theory to illustrate what we should be interested in representing before we then elaborate on these aspects and finally proceed to present one approach to modelling these features using a general logical-mathematical framework called neighbourhood semantics.

Chapter 2

Logics of consequences and appropriateness

2.1 Acting appropriately in institutional contexts

March and Olsen (1989) develop a perspective within organizational theory and political science which stresses "the role of institutions and institutionalization in the understanding of human actions within an organization, social order, or society"(March and Olsen, 1998, p.948). Their starting-point is that politics is based around institutions understood roughly as "a relatively stable collection of practices and rules defining appropriate behavior for specific groups of actors in specific situations"(March and Olsen, 1998, p.948). These rules and practices "define the framework within which politics take place"(March and Olsen, 1989, p.18). Under rules and practices, they include "routines, procedures, conventions, roles, strategies, organizational forms and technologies around which political activity is construed"(March and Olsen, 1989, p.22).

As the logic according to which agents act within a political institution, they present the logic of appropriateness. (March and Olsen, 1989, p.38)'s proposal is that agents, or "political actors", fit their actions "to situations by their appropriateness within a conception of identity" provided by the institution. Being part of an institution amounts to taking a role within it consisting of rules for appropriate behaviour which are then applied to decide which action to take in various situations. Acting appropriately amounts therefore to fulfilling one's institutional role regardless of what that role consists in. Acting appropriately is not the same as acting morally since "rules of appropriateness underlie atrocities of action, such as ethnic cleansing and blood feuds, as well as moral heroism."(March and Olsen, 2011, p.479).

The logic of appropriateness is presented as an alternative to what they call "the logic of consequences" according to which actions are made on the basis of expected utility or anticipated consequences along the lines of rational choice theory as opposed to on the basis of obligations arising from their adopted institutional role. To distinguish the logics, (March and Olsen, 1989, p.23) present two lists of questions that need to be answered within each logic in order to choose which action to take:

Anticipatory action based on logic of consequences

(1) What are my alternatives?

(2) What are my values?

(3) What are the consequences of my alternatives for my values?

(c) Choose the alternative that has the best consequences

Obligatory action based on logic of appropriateness

(1) What kind of situation is this?

(2) Who am I?

(3) How appropriate are different actions for me in this situation?

(c) Do what is most appropriate.

To illustrate the use of the templates in a decision process, consider for example a situation in which an important public figure with bodyguards is being attacked and the bodyguards must decide on their action. Based on the first template, their alternatives are to either run in safety (and thus be more likely to survive) or stay put and protect their object (and thus be less likely to survive). While they value their life, they possibly also value their job, i.e. role as bodyguard, and escaping from the situation can mean that they will loose their job as bodyguards and loose esteem among their colleagues and others. A relevant further question is thus whether they can live with that consequence.

According to the second template, on the other hand, the bodyguards (hopefully) identify the situation as an attack on their object, and furthermore identify themselves as their object's bodyguards. They conclude that their role requires them to protect their object. This is the appropriate action: "Action involves evoking an identity or role and matching the obligations of that identity or role to a specific situation."(March and Olsen, 1998, p.951).

None of the three questions in the template for obligatory action are idle. Instead, it can be the case that in some contexts "actors have problems in resolving ambiguities and conflicts among alternative concepts of the self, accounts of a situation, and prescriptions of appropriateness"(March and Olsen, 2011, p.482), that is, they have multiple roles in potentially different institutions. It might also be the case that they, in addition to "struggle with how to classify themselves and others—who they are, and what they are", find it unclear "what these classifications imply in a specific situation"(March and

Olsen, 2011, p.482). It might thus be unclear what a role requires in a certain situation, presumably because the norms either under- or overdetermine what the appropriate action is, or because the situation itself is unclear. For March and Olsen (1989), this process of fitting a rule to a situation based on ones role or identity within an institution is thus more analogous to legal than economic reasoning: ”In establishing appropriateness, rules and situations are related by criteria of similarity or difference and through reasoning by analogy and metaphor”(March and Olsen, 1989, p.25).

Importantly then, both in the case of an action based on the logic of consequences and an action based on the logic of appropriateness, we are dealing with a reasoned action as opposed to a habitual action, but we are distinguishing between what is more or less an economic and a juridical reason for acting along the lines of Eckhoff and Jacobsen (1960). Furthermore, it is also important to, as pointed out by Goldmann (2005), to distinguish between the utilitarian reasoning underlying logic of consequences and the deontological reasoning underlying logic of appropriateness on the one hand, and a distinction between opportunism and altruism on the other hand despite occasional conflations by March and Olsen (1996). After all, the role determining appropriateness might very well prescribe opportunism and the preferences of the agent could be of an altruistic nature.

By introducing an institutional approach based on the logic of appropriateness to understand human action, they aim to present an alternative to ”linking action exclusively to a logic of consequences” and thus economic reasoning which ”seems to ignore the substantial role of identities, rules, and institutions in shaping human behavior”(March and Olsen, 1998, p.951). According to the perspective they offer, behaviour within institutions should be understood as driven by rules rather than considerations about consequences, and decisions as made according to the logic of appropriateness.

2.2 Empirical hypothesis versus ideal type

While March and Olsen (1989) present a perspective on behaviour within institutions, we may still ask about the empirical adequacy of the logic of appropriateness formulated as an empirical hypothesis about agents’ reasoning. A study that investigates the empirical adequacy of the logic of appropriateness is presented by Persson (2007). The study tests which of two models, one based on rational calculations (logic of consequences) and one based on normative considerations (logic of appropriateness), better explain the Swedish government’s choices when adjusting to European Union in the case of administrative reorganization. The study revealed that both normative considerations and rational calculations were involved in the three cases of reorganization under study, but that, in two of the cases, the decisions were based more on rational calculations that normative considerations, whereas in the third case, the decision was based more on normative considerations.

Persson (2007)'s study supports (March and Olsen, 1998, p.952) claim that "the two logics are not mutually exclusive" interpreted as a claim about an agent's reasoning and thus that "political action generally cannot be explained exclusively in terms of a logic of either consequences or appropriateness [; a]ny particular action probably involves elements of both". However, as pointed out in Goldmann (2005), this stands in sharp contrast to claims such as that "the simple behavioral proposition is that, most of the time humans take reasoned action by trying to answer three elementary questions: What kind of a situation is this? What kind of a person am I? What does a person such as I do in a situation such as this?"(March and Olsen, 2011, p.479), especially if the occasions "human take reasoned action" based on logic of appropriateness are such that the logic of consequences is not invoked. We are also told that "Action is often based more on identifying the normatively appropriate behaviour than on calculating the return expected from alternative choices."(March and Olsen, 1989, p.22) Again, this seems to be an empirical claim.

One way out of this incoherence is, following a suggestion by Goldmann (2005), to consider each logic not as empirical claims but as Weberian ideal types. This will permit us to keep them conceptually distinct while at the same time acknowledging the messy reality in which they are intertwined.

Ideal types were introduced by Weber (1978) as heuristic devices to explore a (social) phenomenon. One famous example is the Weberian construction of a bureaucracy. Ideal types are not supposed to be a description of the typical or average features of the phenomenon; instead they are constructed through the collection of those aspects of a phenomenon which can be unified in a logical way. The ideal type is thus not intended to correspond to anything empirical or be tested empirically, but is rather used to understand the phenomenon through the comparison between instances of the phenomenon and the ideal type, and to aid the formulation of hypotheses. An ideal type is thus neither descriptive in the sense of telling us how something is nor normative in the sense of telling us how something ought to beScott and Marshall (2009). Instead, the ideal type is intended to represent how the phenomenon would have been in ideal circumstances, and it is precisely the idealization involved in construction of the ideal type that ensures their fruitfulness:

> "The more sharply and precisely the ideal type has been constructed, thus the more abstract and unrealistic in this sense it is, the better it is able to perform its functions in formulating terminology, classifications, and hypotheses"(Weber, 1978, p.21).

Being unrealistic is thus a virtue.

Goldmann (2005) does not refer to Weber (1978) with regard to what an ideal type is. Instead, (Goldmann, 2005, p.38) maintains that treating them as ideal types amounts to having "two simplified models to be used as tools for empirical enquiry" where the focus is "less on the similarity between ideal and reality than on the difference"(Goldmann, 2005, p.45). To further illustrate what it means to understand the logic of appropriateness as an ideal type, Goldmann (2005) uses as example game theory and how it has been used to articulate in precise terms claims about strategic interaction without assuming that social life consists "of two-by-two games with deterministic outcomes"(Goldmann, 2005, p.45). Indeed, the simplifications involved in game theory are such that game theory can be described as providing idealized models of various situations involving strategic interaction.

The notion as presented by Goldmann (2005) is thus also close to the notion of an idealized model in science. Following Frigg and Hartmann (2017), an idealized model is "a deliberate simplification of something complicated with the objective of making it more tractable" with examples being "frictionless planes, point masses, infinite velocities, isolated systems, omniscient agents, and markets in perfect equilibrium".

Frigg and Hartmann (2017) distinguishes between Aristotelian and Galilean idealizations: the former involves abstracting away properties that are not relevant such as the classical mechanics model of the planetary system whereas the latter amounts to introducing deliberate distortions such as omniscient agents in economics. Game theory is arguably an idealization of the latter type. With deliberate distortions such as the assumption that agents are utility-maximizing and have full knowledge of the game and other players' actions, one makes the phenomenon in question more tractable through the formulation of exact definitions and a mathematical theory about which various theorems can be proved.

Turning our attention back to the logics of consequences and appropriateness, we can make the following observations. Firstly, treating them as idealizations or ideal types means that their main purpose is to get an understanding of the phenomenon rather than to provide realistic causal connections intended to explain the phenomenon in an empirically adequate way: in virtue of being heuristics, they are more of a guide than the truth.

Secondly, because they are more of a guide than the truth, we can account for the discrepancies between the ideal types and the observed phenomenon; actions can involve elements of both ideal types without that implying an overlap between the ideal types. We obtain conceptual clarity without rejecting empirical entanglement.

Finally, it offers an analogy with regard to Weberian ideal types of social action. (Weber, 1978, p.25) distinguishes four ideal types of social action:

- instrumentally rational action: action as means to achieve some end
- value-rational action: action based on valuing something for itself such as ethical or aesthetic values independently of the success of the action

- affectual action: based on feelings or emotional states
- habitual action: based on ingrained habits

About these, (Weber, 1978, p.26) notes that "it would be very unusual to find concrete cases of action, especially of social action, which were oriented *only* in one or another of these ways", just as maintained by March and Olsen (1996) regarding logics of appropriateness and consequences and established in one case by Persson (2007).

Furthermore, we observe that the logic of consequences understood as ideal type and instrumentally rational action amounts to more or less the same claim. A perspective involving the logic of consequence sees "action as driven by a logic of anticipated consequences and prior preferences"(March and Olsen, 1996, p.949) which seems just to be more or less a reformulation of instrumental rationality. One could even argue that an action based on the logic of consequences as ideal type *just is* an instrumentally rational action since an action is instrumentally rational "when the end, the means, and the secondary results are all rationally taken into account and weighed"(Weber, 1978, p.26). For example, we are told by (March and Olsen, 1998, p949) that logic of consequences is about how agents "choose among alternatives alternatives by evaluating their likely consequences for personal or collective objectives". This seems also to be merely a reformulation of the Weberian notion of instrumental rationality.

With regard to actions based on logic of appropriateness, one might at first think that they should be understood as value-rational actions insofar as the actions in question are reasoned rather than habitual. However, it seems wrong to say that a bodyguard's actions are based on the bodyguard valuing something for itself. It is fair to say that the action is pursued independent of its success in the sense that the bodyguard acts on the basis of their commitment to being a bodyguard, but not that the bodyguard's actions are due to the bodyguard valuing the norms associated with the role of being a bodyguard. They might also just take their job serious. Instead, it seems more in line with the logic of appropriateness to say that we are dealing with an action based on following a norm independently of the success of the action, and thus that we should talk of norm- or rule-rationality rather than value-rationality, or even juridical rationality to follow Eckhoff and Jacobsen (1960). In that case, one could even suggest that instrumental rationality is a special case of rule-rationality, and thus that rule-rationality is the basic form of rationality along the lines of Schnädelbach (1998). The latter is in tune with March and Olsen (1996) preferred understanding of the relationship between the logics of appropriateness and consequences.

While Goldmann (2005)'s proposal to treat the logics of consequences and appropriateness as ideal types allowed us to account for the empirical entanglement of the logics without thereby giving up on conceptual distinctness, there is a noteworthy substantial difference between the logics of consequences and appropriateness as ideal types on the one hand and Goldmann (2005)'s use of game theory as example on the other hand. The logics as presented in this chapter are informal descriptions of some sort, the same goes for the Weberian notions of rationality. This is not the case with game theory. Game theory is a mathematical theory which can be considered as spelling out how to be instrumentally rational, that is, choosing the option with highest pay-off, in a situation involving multiple agents making instrumentally rational decisions, and thus as explaining how to act according to logic of consequences in such a situation, assuming that logic of consequences and instrumental rationality are, as argued above, more or less the same thing. Game theory then, is, is arguably an *explication* of aspects of instrumental rationality and logic of consequences through the application of *formal methods*. Before we proceed to developing something similar for the logic of appropriateness, it is appropriate to elaborate on the italicized words.

Chapter 3

Theory and method

3.1 Formal methods in the philosophy of science

The aim of this section is to elaborate on the meta-scientific perspective underlying the contention that the development of a logical-mathematical model can advance our understanding of the logic of appropriateness. The presented perspective is along the lines of that by Horsten and Douven (2008).

The use of formal logic in philosophy of science is traditionally associated with the Vienna Circle, logical positivism and their view on science and philosophy, a view which according (Carnap, 1930, p.12) consists in that the role of philosophy is to perform logical analysis of the sentences and concepts of the empirical sciences to clarify its sentences. With the use of formal logic, they aimed at providing formally adequate definitions of various concepts in philosophy of science, for example that of a scientific theory, explanation, confirmation, reduction, causation and scientific law. The questions of philosophy of science were to be answered through the definition of these concepts with the aid of the formal methods available at the time, namely the first-order predicate logic developed for example by Frege (1879) and Russell and Whitehead (1910). As explained by Creath (2017), the formal methods were for them thus a tool to reconstruct the empirical sciences in a fashion that supported a unification of natural and social sciences. Central to this endeavour was a distinction between empirically significant sentences on the one hand, and metaphysical sentences on the other hand, where a resistance to a formalization which reduced seemingly non-empirical concepts to empirically significant concepts was considered as showing that the non-empirical concepts were of a metaphysical nature and thus, from a scientific perspective, meaningless. Indeed, formalization was intended as a tool to distinguish the scientific from the unscientific.

However, as stressed by for example Müller (2010) and Horsten and Douven (2008), this research program fell short on delivering adequate definitions of various concepts based on their own strict requirements due to a limited toolbox. In addition, the negative view on scientific concepts that didn't lend themselves to a formalization was too radical. The use of formal methods in the philosophy of science need not be accompanied with such a negative view on pre-analysed concepts, nor should it restrict itself to first-order predicate logic as tool. Instead, as argued by Horsten and Douven (2008), it is precisely through the development of other formal methods such as (Bayesian) probability theory, model theory, recursion theory and intensional logic that the use of formal methods becomes relevant to philosophy of science in a fashion related to, but distinctively different from that associated with logical positivism.

According to this more moderate view on the use of formal methods in philosophy of science, "formal methods can shed light on problems in the philosophy of science, but it would be unreasonable to expect that formal methods can, on their own, solve problems in the philosophy of science"(Horsten and Douven, 2008, p.158). Formal methods can be used to spell out the details of a concept and uncover hidden assumptions in theories. One example is the concept of confirmation and how Bayesian probability theory has arguably advanced our understanding of the problem of induction but without thereby solving the problem. In that way, one can, as Horsten and Douven (2008) do, argue that they function like spectacles through which we can investigate concepts:

> "Ironically perhaps, Kuhnian ideas can be used to explain how formal methods can yield increased insight in a domain. Formal methods function as paradigms in the Kuhnian sense of the word in that they are used for modelling concepts and problems in the philosophy of science. As such, they function as spectacles through which we can look at these concepts and problems and in this way give us insight into them. In this respect, philosophy of science does not differ from the sciences themselves. The mathematical theory of analysis, for instance, functions as a paradigm in classical mechanics in the same sense in which the formalism of Bayesian networks functions as a paradigm in the recent study of causality. Even the logical framework in the narrow sense functions as a paradigm in this sense. It allows us to view a scientific theory as a finite object: a finite set of basic principles closed under logical deduction"(Horsten and Douven, 2008, p.159).

It follows, importantly, that there is no such thing as the definite formal analysis of a scientific concept, or one particular universal formal method. Instead, different methods can yield different results and thus be used to illuminate different aspects of the same concept. The use of formal methods is guided both by the concepts being formalized and what we are interested in exploring with regard to that concept.

In addition to concepts employed in most sciences and thus those dealt with in general philosophy of science, formal methods can also be employed in specific sciences such as social sciences and organizational theory (if the latter counts as a scientific discipline).

In the case of the logic of appropriateness, one can argue that the concept of being an appropriate action is precisely such a concept which is fundamental to institutionalism as paradigm within organizational theory, more or less in the same way as causation is fundamental to most sciences.

3.2 Explication

While our theoretical perspective on the use of formal methods is more moderate than that associated with logical positivism, the method applied in the formalization of a concept is still the one associated with logical positivism, namely explication.

An explication consists in making ”more exact a vague or not quite exact concept used in everyday life or in an earlier stage of scientific or logical development, or rather of replacing it by a newly constructed, more exact concept”(Carnap, 1956, p.7). The method was first introduced under the label of *rational reconstruction* in Carnap's *Der logische Aufbau der Welt* originally published in 1928. In the preface to the second edition the aim of the method is described as follows:

> ”Unter rationaler Nachkonstruktion ist hier das Aufsuchen neuer Bestimmungen für alte Begriffe verstanden. Die alten Begriffe sind gewöhnlich nicht durch überlegte Formung, sondern durch spontane Entwicklung mehr oder weniger unbewußt entstanden. Die neuen Bestimmungen sollen den alten in Klarheit und Exaktheit überlegen sein und sich vor allem besser in ein systematisches Begriffsgebäude einfügen.”(Carnap, 1966, p.X)

The aim is thus to improve our conceptual framework by replacing a vague or imprecise concept, the *explicandum*, with a more exact concept, the *explicatum*, that serves more or less the same purpose, and through that aid the advance of science. Of course, based on the meta-scientific view presented in the above section, we should think of an explication not as providing a replacement, but as making precise aspects of a vague or imprecise concept. The method itself, *to explicate* a concept, is still the same, as we shall see when considering the criteria for a good explicatum.

As criteria for a good explicatum, Carnap (1950) lists similarity to the explicandum, exactness, fruitfulness and simplicity as criteria. The explicatum should be similar to the explicandum in the sense that we can use it in most of the cases where the explicandum is used, but ”considerable difference are permitted”(Carnap, 1950, p.7) in certain circumstances. The characterization of the explicatum should be as exact as required to permit the introduction of the explicatum into a conceptual framework containing other scientific concepts, and it should be useful for the formulation of empirical laws and/or logical principles. Finally, it should be as simple as the previous requirements permit and if there are multiple explications satisfying the first three criteria, we should choose the simplest explication.

An explication of a concept involves, to the extent which it is fruitful to obtain the required exactness, the use of logical-mathematical methods. It also encourages the use of quantitative over qualitative concepts to achieve the required exactness. However, it is important to stress that the method does not, as such, stand in contrast to the hermenautic tradition we can associate with institutionalism since the aim of the method is merely to improve concepts, not to enforce a positivistic methodology. The explication should involve an interplay or dialectic between the explicatum developed with formal tools and the current usage of the explicandum in scientific discourses. With the aim being to improve our understanding the explicandum through such an interplay, one could even argue that the method itself amounts to engaging in a hermeneutic circle.

Reconnecting the method with the meta-scientific view presented in the previous section, we first observe that if the aim of an explication through the application of formal methods is to shed new light on a concept or a theory, for example by uncovering hidden assumptions or by making precise the relationship between concepts involved in the theory, then the fruitfulness criterion can be understood directly in relation to that aim. It is thus crucial, before beginning an explication, to set out the aim for the explication; if not, how can we evaluate it? Similarly, the aim of the explication has an impact on exactness and simplicity through the choice of formal methods, and finally also to the requirement of similarity. For example, the use of first-order predicate logic can provide a simple framework and will certainly be exact, but might fail to shed any light on the concepts simply because we lack the similarity between the explicatum and the explicandum. It would thus not be useful. This was as mentioned above and as argued by Horsten and Douven (2008), one of the main problems with the toolbox of the logical positivists. Our toolbox will hopefully be more suitable.

3.3 Expected utility theory as explication of logic of consequences

To illustrate the perspective and the method described in this section, we shall in this section present expected utility theory as an explication of the logic of consequences.

Expected utility theory refers to a variety of formal theories of choice within decision theory. The use of "theory" here might seem objectionable, but it is sometimes common to refer to a field of research as "theory". Consider for example "proof theory" and "model theory" of which neither refer to specific theories in the usual sense but are rather fields of research. In any case, the conceptual starting-point in decision theory is that agents have preferences over prospects or options. The agent's preferences is a comparative attitude and an agent prefers some option over another option just in case the former option represents a more desirable outcome than the latter option. We distinguish between a weak and a strict preference-relation over options where the weak preference-relation corresponds to ”less than or equal to” in arithmetic, i.e. an option A is less or equally preferred to an option B, i.e. A is not preferred to B, and the strict preference-relation corresponds to ”less than” in arithmetic, i.e. an option A is less preferred to an option B. The weak preference-relation is normally assumed to satisfy the following constraints over a domain of options:

- Completeness: for every options A and B, either A is not preferred to B or B is not preferred to A
- Transitivity: for every options A,B and C: if A is not preferred to B and B is not preferred to C then A is not preferred to C

An agent’s preferences over options is thus, in order-theoretic terminology, a weak ordering. With completeness, all options in a domain are comparable to each other, and with transitivity we obtain something very much like consistency.

To every complete and transitive ordering of an agent’s preferences, there is a utility function based on ordinals representing the ordering in the sense that for every options A and B, A is not preferred to B if and only if the ordinal utility of A is less than or equal to the utility of B. In addition to the utility of an option relative to other options, one can also define the *expected utility* of an option relative to other options by associating with each option a subjective probability in addition to the utility which then represents the chances of ”winning” that option. This is expected utility theory and the important results within expected utility theory concerns which conditions the preferences over options are sufficient for there to be a utility function such that an agent can be said to maximize expected utility(Steele and Stefánsson, 2016). Varieties are then obtained for example by requiring various constraints on the preferences or also use different ways to calculate the utility. For example, if we add two additional two constraints on the preference-relation, continuity and independence, we can prove that there there is an expected utility function such that any preference-ordering satisfying the constraints can be said to maximize expected utility.

Expected utility theory can be understood as explication of instrumental rationality, but as discussed by Briggs (2017), there is significant disagreement about which constraints that should hold on the preferences for the agent to act instrumentally rational. Nonetheless, we can still maintain that we obtain with expected utility theory various suggestions for precise conditions for acting in a way that ensure that, on average, one prefers the means to ones end, and we thus told something about the idea of acting in an instrumentally rational way. This relates to fruitfulness as criterion for an explication. We can for example with the help of expected utility theory prove that for every preference-ordering satisfying the relevant constraints there is a way to assign either ordinals or intervals which in turn shows that even if the agents themselves don't have to order them numerically, such an ordering exists. Without expected utility theory as a formal framework such a claim would have remained an unverified assumption about instrumental rationality which nonetheless for example various versions of utilitarianism can be said to rely on. Expected utility theory shows us what is required for an agent to be instrumentally rational with regard to expected utility: its preferences must have such and such an ordering because if it doesn't, then the agent cannot select the utility-maximizing option. With expected utility theory we have learned more about what being instrumentally rational amounts to. We also note that the concepts defined in expected utility theory are more exact, but that we can still see the similarity to instrumental rationality. If we understand logic of consequences as ideal type and thus as more or less the same as instrumental rationality, expected utility theory is thus also an explication of the logic of consequences. After all, the logic of consequences is clearly presented in a way that align it with expected utility theory since a "decision maker would be expected to choose the one combination that maximizes expected return"(March, 2009, p.5) and "political actors are imagined to be endowed with preferences or interests that are consistent, stable, and exogenous to the political system"(March and Olsen, 1996, p.251).

Chapter 4

Explicating the logic of appropriateness

4.1 Appropriateness in game theory

We now turn our attention to explicating the logic of appropriateness. As a starting-point and to illustrate what we should be interested in when it comes to explicating the logic of appropriateness, we will have a look at the extent to which the logic of appropriateness can be articulated within game theory.

Game theory is multi-agent decision theory that aims to ”understand situations in which decision makers interact.”(Bicchieri, 2004). Each player in a standard game aims at maximizing utility. As an example, let us consider the classical game known as the Prisoner's dilemma. The traditional background story for the game is that two partners in crime have been arrested by the police and are placed in two different interrogation rooms where they have the choice between putting the blame on their partner, that is, defecting from the partnership, or remain silent, that is, cooperating with their partner as agreed upon. If one of them defects, that person goes free and the other gets four years in prison, if both do it, they get two years in prison. There is however sufficient evidence to convict both of them to one year in prison if both remain silent.

The one-shot game can be portrayed with the following matrix:

	D	C
D	2,2	4,0
C	0,4	3,3

This is a two-player game, where each player has two moves, D(efect) and C(ooperate). If both cooperate, then the payoff is 3 for each, and if both defect, the payoff is 2 for each. If one defects and the other cooperates, then the defecting one receives 4 and the cooperating one receives 0. The best strategy for this game if it is only played once is to defect for both players as the payoff is better: $2 + 4/2 > 3 + 0/2$. Similarly, if the game is played exactly n times where each player will remember the previous move, it is also better to defect every game. The proof is by backwards induction from n to 0. Assume that we are at game n. Since this is the last game, it counts as a one-shot game and each player is better of defecting. If that is the case, then both players should also defect at the penultimate game as this will also provide a higher pay-off. This reasoning is then repeated till we reach the first game.

The option to cooperate with the partner by remaining silent can easily be portrayed as the appropriate thing to do considering the partnership. A criminal enterprise is an institution with its social norms, one of them being not to talk to the police. The option to defect from the partnership, on the other hand, ends up representing an action based on the logic of consequence if we assume that the logic of consequence is about utility-maximizing. We can thus easily replace the maximizing-utility strategy with a strategy according to which one picks the option that is predefined as the appropriate one. This may, but need to be, a strategy according to which one picks the option that gives the most to everyone, i.e. for common good, since the appropriate action could have been an action that gives the least utility to every player. Being norm-rational amounts to picking the appropriate option because it is dictated by some norm regardless of its consequences.

However, while we can in that way introduce an agent that acts appropriately by stipulating one option to represent the appropriate action, and thus illustrate the difference between the logics of consequences and appropriateness from a game-theoretic perspective, such an explication of the logic of appropriateness cannot be said to illuminate any aspects that we couldn't also have achieved without the use of game theory. It helps us to illustrate conceptual distinctness, but it fails to say anything interesting about the logic of appropriateness itself. For example, it doesn't touch upon the aspects that goes into the process of choosing the appropriate action in the cases where there is more than one role with multiple norms of which some are inconsistent and some incomplete. It also doesn't say anything about the relationship between an agent's beliefs about the situation, the roles played by the agent and the appropriate action. An explication should make a contribution to our understanding of the logic of appropriateness, not just be an illustration.

4.2 Roles, situations and appropriateness

Based on our conclusion in section 4.1, it seems reasonable to maintain that our explication should provide a formal model that captures how the "matching of a situation to the demands of a position"(March and Olsen, 1998, p.23) delivers the appropriate action. The formal model should thus allow us to reason about the concepts involved in the logic of appropriateness, namely role, situation and appropriate action. Before we can start developing a formal model in which these concepts are represented, we should clarify what we expect from these concepts with regard to the logic of appropriateness.

Let us begin with the concept of a role. The basic idea in logic of appropriateness is that agents take up roles within institutions and each such role come with expectations, permissions and obligations, that is, the norms. Some norms will be unconditional, while other norms will be conditional on some event. We could include other normative concepts such as supererogatory actions, that is, actions that go beyond "the call of duty" but can nonetheless be included as practices. We can for example imagine an organization in which some actions are actively encouraged but nonetheless not obligatory. Working late as trainee in an accountancy practice, for example, could be one such norm: while it is not obligatory to work late, it is praiseworthy to do so within the firm, and thus at least portrayed within the organization as a supererogatory action. On the other hand, such a behaviour from trainees could also be simply expected.

An important aspect of such norms is that they may only be "codified to some extent, but the codification is often incomplete. Inconsistencies are common. As a result, compliance with any specific rule is not automatic"(March and Olsen, 1989, p.22), especially since "rules and their applicability is often ambiguous"(March and Olsen, 1989, p.24). In other words, we should require from a formal model that it includes the possibility of inconsistent and incomplete norms.

In addition to the normative dimension, a role should also include a doxastic dimension since March and Olsen (1989) suggest that taking up a role amounts to accepting a conceptual scheme which in turn influences how one understands situations. This can be spelled out in different ways. On the one hand, we can talk about evidence gathering as influenced by a role: the role guides to some extent what one looks for and considers to be relevant for determining the appropriate action. On the other hand, we could think of the agent's beliefs as being relative to a role. The latter seems to be the extreme option since it would mean that an agent would change their belief when changing role, as opposed to merely changing focus with regard to evidence gathering which results in different evidence being gathered. The former is also more in tune with (March and Olsen, 1989, p.41)'s underlining of how what a person "sees" is influenced by the institution. Relativizing evidence-gathering to a role and allowing each role to gather possibly conflicting evidence seems sufficient to capture how "situations can be defined in different ways that call forth different rules"(March and Olsen, 1989, p.24).

A role contains thus norms for action and evidence gathering, and we are left with connecting the roles with appropriate action. Firstly, we observe that the concept of an appropriate action should be subjective rather than than objective, more or less in the same way as expected utility theory involves subjective rather than objective probabilities. In other words, an action is appropriate for an agent when the agent's understanding of the situation is such that a norm requiring that action applies. Now, the agent's understanding of the situation will depend on their beliefs in addition to the available evidence which in turn depends on the role(s). It means that we are, when establishing whether something is appropriate, not interested in what is true, but what the agent believes and has evidence for. Indeed, what the agent is matching with the norms is not the situation itself but rather the agent's understanding of the situation, because how could the agent be in position to match something which it might not have access to?

While the norms associated with a role may be such that conflicting actions are obligatory, or it may also be the case that one's understanding of a situation is such that conflicting actions are obligatory, it seems still to be the case that conflicting actions cannot both be appropriate. Instead, the agent needs to dissolve the conflict in order to determine what what is appropriate. For example, (March and Olsen, 1989, p.25) suggest that "when more than one potentially relevant rule is evoked, the problem is to apply criteria of similarity in order to use the most appropriate rule. In some cases, higher-order rules may be used to make the choice." The most appropriate rule, we can assume, is the rule that prescribes the action which is the most appropriate action to do, and we are supposed to use some criteria of similarity (between perhaps our understanding of the situation and something else) to make that choice. Now, the use of "most" in this connection seems also to suggest that there should be an ordering of norms of according to their relevance to the situation which is established through similarity.

In addition to conflicts within a role about what should be the most appropriate action, it may also be the case, since "individuals have multiple identities"(March and Olsen, 1989, p.24), that there is a conflict with regard to what's appropriate among roles taken up by the agent. One seemingly natural way to dissolve such conflicts, which however hasn't as far as I know been suggested by March and Olsen (1989), is to introduce an priority-ordering on the roles. To illustrate why such an ordering is plausible, let's consider a father travelling with his daughter on an airport and the boarding to the plane has just started. Unfortunately, the father has lost sight over his daughter and is now faced with a dilemma. From the perspective of his role as flight passenger, he is required to board the plane. From the perspective of his role as father, he is required to find his daughter as he cannot leave her on the airport. Each action is appropriate with regard to some role, but they are conflicting. However, it seems fair to assume that the role as father has priority over the role as flight passenger, at least in this case. With a priority-ordering on the roles, only one of the actions is the appropriate thing to do even if both are obligatory based on one's understanding of the situation. Of course, unless a priority-ordering satisfies certain conditions, it might still leave things undecided, for example if two roles are equally prioritized in a certain situation. That is however to be expected; it might not be clear what is the appropriate thing to do, either within a role or between multiple roles. On the other hand, it might also be interesting to clarify which conditions such a priority-ordering on roles must satisfy to dissolve conflicts with regard to what is appropriate.

This concludes, for now, our discussion of the concepts and their relationships which we should aim to represent in a logical-mathematical model for the logic of appropriateness. This discussion together with the observation in section 4.1 can be considered as a preliminary answer to our first research question. We now proceed to show one way in which they can be modelled as an answer to our second research question.

4.3 Appropriateness in neighbourhood semantics

4.3.1 Semantics based on frames

As discussed in section 3, we do not have any reason to think that there is a single formalism suitable for explicating every concept, and the formalism one chooses depends both on what one is explicating and what one wishes to achieve with that explication. We have chosen neighbourhood semantics as formalism for an explication of the logic of appropriateness with regard to the aspects sketched above in section 4.2 because neighbourhood semantics has been applied to provide interesting explications of concepts such as that of being a conditional obligation as discussed by Decew (1981), and of an agent's evidence and beliefs about some situation by for example van Benthem et al. (2014). Since these are central to the logic of appropriateness, neighbourhood semantics seems to be a good starting-point for an explication.

Semantics is concerned with the meaning of linguistic expressions, and neighbourhood semantics is an approach to defining the meaning of (descriptive) sentences of a (predominantly formal) language, where the meaning of a sentence is understood as the conditions under which a sentence is true. The truth-conditions for each sentence of the language in question is specified by defining models, where a model is an idealized representation of whatever the sentences of the language are intended to describe.

The models themselves are defined using set theory. The description of the models will thus involve set-theoretic concepts and notation that the reader might not be familiar with, and we provide a quick glossary of the most crucial terms.

- $\{\ldots,\ldots\}$ describes a *set*, where a set is basically a collection of things such that two sets are identical if and only if they contain the same elements.
- A set can be specified either by listing its elements, or by presenting the condition for being an element of that set using the following notation: $X = \{a \mid \text{some condition } a \text{ must satisfy }\}$
- If a is an object and X is a set, then $a \in X$ means that a is an element of X and $a \notin X$ means that a is not an element of X
- If X and Y are sets, then $X \subseteq Y$ means that X *is a subset of* Y; that all elements of X are also elements of Y.

- $\emptyset$ refers to *the empty set*, i.e. the set containing no object. It is a subset of every set.

- $\cap$ and $\cup$ are operations on sets such that:

  - $X \cap Y$ defines the set containing the elements that X and Y has in common, their *intersection*.
  - $X \cup Y$ defines the set containing the elements of both X and Y, their *union*.

- The power set of a set X, written as X^2, is the set containing every subset of X, so if $Y \subseteq X$ then $Y \in X^2$.

- $\langle \ldots, \ldots \rangle$ describes a list or sequence of objects. The smallest such sequence of objects is an ordered pair.

- $X \times Y$ refers to the Cartesian product of X and Y, the set of ordered pairs obtained by relating each element of X with each element of Y: $X \times Y = \{\langle a, b\rangle \mid a \in X$ and $b \in Y\}$.

- A set of ordered pairs is a (binary) *relation*. If $Z \subseteq X \times Y$ then Z is a binary relation on $X \times Y$.

- If a binary relation Z on $X \times Y$ is such that every element of X is the first component of *one and only one* ordered pair of Z, then Z is a binary function. The notation $Z(a) = b$ means that the function Z maps $a \in X$ to $b \in Y$.

For a more elaborate introduction, see Bagaria (2017).

Models in neighbourhood semantics is a generalization of what is known as ”possible world” or frame semantics which is for example used to explicate the concept of necessity and has its origin in the works of Carnap (1956) and Kripke (1959). The underlying idea is that, to evaluate the truth of a sentence such as ”it is necessary that 1+1=2”, one should ”check” that there is no possible scenario, alternative situation or ”world” in which ”1+1=2” is false, where a situation or a ”world” is something like an imaginable variation of our actual world (where ”world” does not refer to our planet, but rather something like our universe, assuming our universe marks the relevant borders). As a first step towards developing a neighbourhood semantics for the logic of appropriateness, we will provide a frame semantics for the concept of necessity.

The frame semantics will be defined for a language consisting of the set of formulas obtained with a countable set of so-called atomic formulas which we will refer to using lower-case Latin letters starting from p, the connectives $\vee$, $\wedge$, $\neg$, $\Box$ and parentheses (and). We refer to any formula of the language using lower-case Greek letters. The set of formulas is defined as follows:

- all atomic formulas are formulas
- if ϕ is a formula then $(\neg\phi)$ and $(\Box\phi)$ is a formula
- if ϕ and ψ are formulas then $(\phi \vee \psi)$ and $(\phi \wedge \psi)$ are formulas
- Nothing else is a formula

Let us call this language $\mathcal{L}_\Box$. We will omit parentheses when they are not required for disambiguation.

The atomic formulas can be understood as representing simple descriptive claims such as "snow is white", "potatoes are tasty" or "1+1=2". If p represents "snow is white" and q represents "1+1=2" then $\neg p$ corresponds to "it is not the case that snow is white", $p \vee q$ corresponds to "snow is white or 1+1=2", $p \wedge q$ corresponds to "snow is white and 1+1=2" and $\Box p$ corresponds to "it is necessary that snow is white".

A formal language such as $\mathcal{L}_\Box$ is thus both a Galilean and an Aristotelian idealization of natural or ordinary languages such as English or Norwegian. It is an Aristotelian idealization in the sense that we have excluded anything seemingly unessential in order to put the focus on the logical aspects of claims involving the concepts of negation represented by $\neg$ (a logical idealization of "not"/"it is not the case"), inclusive disjunction represented by $\vee$ (a logical idealization of "or"), conjunction represented by $\wedge$ (a logical idealization of "and") and necessity. Notice also that all of our sentences as descriptive, there are no questions or imperatives. It is a Galilean idealization because we have thereby distorted English: these are not normal sentences we could expect the typical English speaker to use; they easily get very awkward.

In frame semantics, we let each model be a ordered pair $M = \langle W, V\rangle$, the frame. The frame consists of two components, W and V. W is a set of objects; the "worlds" at which formulas are either true or false. We shall refer to such "worlds" as "points of evaluation" to have a somewhat neutral terminology. V is a function from the atomic formulas to the power set of W, assigning to each atomic formula of $\mathcal{L}_\Box$ a subset of W. We then define a relation $\Vdash_M$ for each model M on $W \times \mathcal{L}_\Box$ such that for each $w \in W$:

- For all atomic formulas p: $w \Vdash_M p$ if and only if $w \in V(p)$
- $w \Vdash_M \neg\phi$ if and only if $w \nVdash_M \phi$
- $w \Vdash_M \phi \vee \psi$ if and only if either $w \Vdash_M \phi$ or $w \Vdash_M \psi$
- $w \Vdash_M \phi \wedge \psi$ if and only if either $w \Vdash_M \phi$ and $w \Vdash_M \psi$
- $w \Vdash_M \Box\phi$ if and only if for each $v \in W$, $v \Vdash_M \phi$

With $\Vdash_M$ being a relation, $w \Vdash_M \phi$ is an abbreviation of $\langle w, \phi\rangle \in\Vdash_M$ and $w \nVdash_M \phi$ corresponds to $\langle w, \phi\rangle \notin\Vdash_M$, but our notation increases readability and is quite standard.

The intended understanding of $w \Vdash_M \phi$ is that ϕ is true at w and $w \nVdash_M \phi$ as ϕ is false at w. Going forward, we shall omit the subscript M unless the context requires it.

The language we use to describe the models we define for the language $\mathcal{L}_\Box$ is itself a mix between English and a formal language for set theory. This language can also be formalized and we can present a proof system for deriving statements about the models in this language from other statements about the models in this language along the lines of Negri (2005). That would however be too tedious and is not the usual procedure.

With the help of the above models we can now define a set of truth-preserving inferences for the language $\mathcal{L}_\Box$ as a relation on $\mathcal{L}_\Box^2 \times \mathcal{L}_\Box$, that is, the Cartesian product of the power set of $\mathcal{L}_\Box$ and $\mathcal{L}_\Box$. Let Γ be a set of formulas of $\mathcal{L}_\Box$ and $\vDash$ represent ”truth-preservingly entail” so that $\Gamma \vDash \phi$ says that the formulas in Γ truth-preservingly entail ϕ. Then $\Gamma \vDash \phi$ if and only if, for every model $M = \langle W, V \rangle$ and for every $w \in W$, if $w \Vdash \psi$ for every formula in $\psi \in \Gamma$, then $w \Vdash \phi$. That an inference is truth-preserving means thus that whenever the premises are true, the conclusion is also true.

To illustrate this and how to reason about the models, consider now whether, according to our model, it is the case that the sentence ”it is necessary that snow is white” truth-preservingly entails ”snow is white”. Let M' be a model based on our frames, let w' be an element of W of M'. Assume that $w' \Vdash \Box p$. It follows by the definition of $\Vdash$ that for all $v \in W$ of M', $v \Vdash p$. Since w' is one of the v's in W of M', it follows that $w' \Vdash p$. Since w' was an arbitrary element of W, the previous holds for every element of W. Since M' itself was also arbitrary, the previous holds for every such model. It follows that $\Box p \vDash p$. Whenever it is necessary that snow is white is true, it is also true that snow is white. Quite rightly so. Our set of inferences is *a logic*, and we can call our logic a *logic of necessity*. It contains other concepts too, but those belong in this case to the background against which we define necessity. Indeed, the aim of the above semantics is to define a logic in which the logical aspects of the concept of necessity are represented in such a way that we see its connections with other concepts. Let us try to achieve the same for appropriateness, that is, define something which could count as a *logic of appropriateness* according to these conventions.

4.3.2 Neighbourhoods in frames for representing evidence

Neighbourhood semantics is obtained by extending a frame with one or more functions from W to W^{2^2}, that is, from the set of points of evaluation W to the power set of the power set of W. These functions are called neighbourhood functions, and they are employed to provide clauses for the definition of $\Vdash$ for one or more connectives. Each such subset of W is called a neighbourhood, so each point of evaluation is thus assigned a set of neighbourhoods through a neighbourhood function.

To turn our frame semantics into a neighbourhood semantics, we start with representing the concept of evidence, more or less along the lines of van Benthem et al. (2014) and Pacuit (2017). To that purpose we augment the language $\mathcal{L}_\Box$ by adding a new connective $[E_i]$. This connective can be understood as for example "agent i has evidence for" or "source i provides evidence for", but we shall, due to our goal which is to provide a neighbourhood semantics for the logic of appropriateness, use indices such as i to represent particular roles. The connective should thus be understood as something like "the agent has, through the role i, obtained evidence for". The new language $\mathcal{L}_{\Box[E_i]}$ is defined by including that $([E_i]\phi)$ counts as a formula in the recursive definition of $\mathcal{L}_\Box$.

We now let our models be $M = \langle W, N_{E_i}, V\rangle$ which are like before with regard to W and V, but we add a neighbourhood function N_{E_i} and we add the following clause to the definition of $\Vdash$:

- $w \Vdash [E_i]\phi$ if and only if there is a $X \in N_{E_i}(w)$ such that $v \Vdash \phi$ for each $v \in X$

The truth-condition for a formula of the form $[E_i]\phi$ thus reads: ϕ is available evidence at w if and only if there is a evidence-neighbourhood of w such that ϕ is true at each point of evaluation in that neighbourhood.

Finally, van Benthem et al. (2014) require that the neighbourhood function N_{E_i} satisfy the following constraint:

- For each $w \in W : \emptyset \notin N_{E_i}(w)$ and $W \in N_{E_i}(w)$

The idea behind the neighbourhood function in the case of a logic of evidence is that each subset of W in $N_{E_i}(w)$ represents a piece of evidence somehow available to i at w. A piece of evidence can be good or bad. It can be good in the sense that it supports some formula because that formula holds at each element of X. It can be bad when it supports no formula in that way. The constraint ensures that there cannot be evidence for something which is necessarily false (even if two pieces of evidence can contradict each other), and that there is always some piece of evidence for everything which is necessary, i.e. which is true at each point of evaluation in the frame. Again, these are idealizations. Firstly, evidence doesn't come in this shape, but we are merely representing evidence in this way to capture logical aspects of the concept of evidence. Secondly, maybe we can have misleading evidence supporting a contradiction, and maybe we do not have evidence for everything which is necessary.

The constraints ensure that the set of truth-preserving inferences will for these models be such that for example $\Box\phi \vDash [E_i]\phi$ for any formula ϕ of $\mathcal{L}_{\Box[E_i]}$, but $[E_i]\phi \nvDash \phi$; having evidence for ϕ doesn't imply that ϕ is true.

We will however reject that we always have evidence for everything which is necessary. Since the evidence we are after will evidence which is relevant for norms, and conditional obligations are simply obligations if they are in force whenever something necessary hold; it is not evidence that would be gathered for conditional obligations.

4.3.3 Obligations and conditional obligations

We proceed to introduce the other concepts required for the logic of appropriateness. In addition to evidence, the logic of appropriateness also involves norms, the idea being that agents obtain evidence for norms applying. To represent the norms, we shall rely on a few additional idealizations. Firstly, we shall assume that all norms are somehow of the same category and that they can be represented linguistically as "it ought to be the case that ϕ", or that "the agent ought to see that ϕ" where ϕ then is for example "the copy machine is filled with paper" or "the desk is clean". We shall represent norms of this form with a new connective $[O_i]$, where i is again an index representing a role to which the various norms are associated. $[O_i]\phi$ reads thus that "it is obligatory according to role i that one ought to act so that ϕ". In addition to adding the appropriate clause to the definition of the formulas, we also add a neighbourhood function N_{O_i} to the frame and the following clause for $\Vdash$:

- $w \Vdash [O_i]\phi$ if and only if there is a $X \in N_{O_i}(w)$ such that for each $v \in W$, if $v \Vdash \phi$ then $v \in X$

Note the difference between the clauses for $[E_i]\phi$ and $[O_i]\phi$. In the former case, it suffices that there is a neighbourhood of w such that ϕ holds at each point of evaluation in that neighbourhood, whereas in the latter case, there must be a neighbourhood of w such that each point of evaluation at which ϕ holds is in that neighbourhood. The immediate consequence of this is that, while if ϕ entails ψ then $[E_i]\phi$ entails $[E_i]\psi$, this fails to hold in general for $[O_i]$. It is thus the case that if some evidence entails ψ, then there is evidence for ψ. This is not the case for obligations. Furthermore, it is the case that if the agent has evidence for $\phi \wedge \psi$ then the agent has evidence for ϕ and evidence for ψ. This is also not the case for obligations.

We also note that, in the same way as we do not have evidence for necessities, necessities are not obligatory for the obvious reason that if something is anyway the case it shouldn't be an obligation. In addition, inconsistent obligations are possible. The latter is clearly in line with the understanding of norms in March and Olsen (1989).

As in the case of evidence, we can follow Anglberger et al. (2015) and understand the neighbourhood function N_{O_i} as providing a set of obligations at each point of evaluation, of which some obligations are clear and others unclear. An obligation is clear if the set contains all the points of evaluation at which the formula is true, and unclear if not. A neighbourhood of N_{O_i} need not *eindeutig* represent some obligation since it might not be the case that the neighbourhood corresponds to the set of points of evaluation at which some sentence is true. This can be considered as a way of expressing the lack of codification stressed by March and Olsen (1989).

We have now the tools to represent unconditional but not conditional obligations, and it seems reasonable, as argued in section 4.2, to assume that the latter are central to the logic of appropriateness; most institutional norms are such that they kick in if certain conditions hold. What is missing is a concept of conditionality.

While we can define the material conditional in terms of negation and disjunction through letting "if ϕ then ψ" be an abbreviation of "either it is not the case that ϕ or (it is the case that) ψ", this conditional is unsuitable for expressing conditional obligations because the material conditional is monotonic: "if ϕ then ψ" entails "if ξ and ϕ then ψ", but this goes against the thought that conditional obligations are non-monotonic: the conditional obligation that if the copy machine is out of paper then one ought to see that it is filled with paper shouldn't entail the conditional obligation that if the copy machine is out of paper and the copy machine is broken then one ought to see that it is filled with paper. The addition of more antecedents can cancel out an obligation.

We can define a more suitable conditional with neighbourhood semantics. The idea behind this conditional is something along the following lines: To evaluate a statement of the form "If...then.." in a situation, we consider the most similar variations of the current situation in which the antecedent is true, that is, a counterfactual situation in which the antecedent is true. If the consequent is true in those situations, then we can conclude the antecedent counterfactually implies the consequent. This idea originates in the work of Lewis (1973) and can be made formally precise as follows, however with a slight variation from the standard definition because we do not want counterfactuals with necessarily false antecedent to be necessarily true since it would be as useless to let anything be obligatory in contradictory situations as letting everything necessary be obligatory. Instead, it seems better to let such counterfactuals be false.

We add thus a new neighbourhood function N_C and define the connective $\to$ as follows:

- $w \Vdash \phi \to \psi$ if and only if there is $X \in N_C(w)$ such that $v \Vdash \phi$ for some $v \in X$ and for every $v' \in X$, if $v' \Vdash \phi$ then $v' \Vdash \psi$

In addition, we would like the counterfactual to be subjective in the sense that we are not interested in variations on the current situation as such, but rather our understanding of the current situation, that is, what the agent believes about the current situation. Our reason for this will become clear when we use the counterfactual to define the most appropriate action since we require to that purpose a notion of similarity between what the agent believes to be the case and the situation at which an obligation holds.

To ensure that we capture the intuition that the $N_C(w)$ represent similar situations to the agent's understanding of the current situation, we follow Pacuit (2017) and add the following conditions on $N_C(w)$ for every $w \in W$:

- For all $X, Y \in N_C(w)$, either $X \subseteq Y$ or $Y \subseteq X$

- The union of any number of neighbourhoods in $N_C(w)$ is also a neighbourhood of $N_C(w)$

- The intersection of any number of neighbourhoods in is also a neighbourhood of $N_C(w)$

The result is that the neighbourhoods of a point of evaluation take the shape of spheres, where the smallest sphere is intended to represent the agent's beliefs at w along the lines of Berto (2018). Moreover, each sphere is an extension of that sphere such that the smallest sphere is contained in every sphere and for every two spheres, one of them is contained in the other, and both the sphere obtained by joining them together and the sphere obtained by combining what they have in common are among the spheres. This can be turned into an objective counterfactual by requiring that the smallest sphere is the sphere containing w only as opposed to some set of points of evaluation representing the beliefs of the agent.

While the neighbourhoods in the case of evidence and obligations represented either a piece of evidence or a norm, the neighbourhoods in the case of the counterfactual represent thus the degree of similarity with the agent's current beliefs, so the further away, i.e. the larger the sphere, the less similar its points of evaluation are to how the agent believes things are.

Given that the smallest sphere at a point of evaluation represents the agent's beliefs at that point of evaluation, and we can assume that the agent has some kind of evidence for each of their beliefs, it seems fair to connect the neighbourhood functions for evidence and the counterfactual as follows:

- If $X \in N_C(w)$ is such that for all $Y \in N_C(w), X \subseteq Y$, then $X \in E_i(w)$.

The available evidence for the agent relative to a role i at a point of evaluation w includes evidence for the agent's current beliefs at w.

4.3.4 Appropriateness relative to a role

We are now in position to articulate one aspect of the logic of appropriateness. Given the connection between the understanding of the situation and what a role requires on the one hand, and what is appropriate to do, as expressed in list of questions in section 2.1, we can say that if it is appropriate to act such that ϕ, then there is evidence for the condition that makes it obligatory to act such that ϕ. Let us represent this.

In the same way as we have connectives for evidence and obligation, we also add a connective $[A_i]$ for appropriateness relative to a role, and for which we also provide a neighbourhood function, N_{A_i}. Based on the previous paragraph, we obtain the following constraint on N_{A_i}:

- if $Z \in N_{A_i}(w)$ then there is $X \in N_{E_i}(w)$ such that there is a $Y \in N_C(w)$ such that $X \cap Y \neq \emptyset$ and for each $v \in X \cap Y, Z \in N_{O_i}(v)$

With this we have a condition for some action being appropriate: there is evidence for the conditions making the action obligatory, where calling them obligatory involves an idealization, since it might also be for example an expectation or permission. If we had connectives for expectation and permission, the constraint should be modified to accommodate the new neighbourhood functions through the addition of alternatives to "for each $v \in X \cap Y, Z \in O_i(v)$", i.e. "for each... or for each... or for each...". In any case, if there is no evidence supporting the antecedent of the conditional ought, why should one think that the obligation is in force? Then the norm wouldn't be the reason for the agent's action.

However, the above constraint shouldn't be replaced by an equivalence since there is (various) evidence for multiple or even conflicting obligations. They cannot all be appropriate in the sense of the most appropriate act to do, that is, what someone in that role should do in that situation. The challenge is thus to turn potentially multiple obligations into one appropriate action.

The idea in the logic of appropriateness is that one of the various actions for which there evidence that can be expected or obligatory or suitable, i.e. fitting the normative context, is more suitable than the others, and that is the appropriate action. March and Olsen (1989) describes the reasoning leading up to concluding which action is appropriate as analogous to legal reasoning which proceeds by similarity. Now, our counterfactual is based on a similarity-ordering of spheres: the closer the sphere is to the agent's beliefs, the more similar its points of evaluation are to the agent's beliefs.

This opens for the proposal that the more similar the situation is at which something is obligatory is to what the agent's believes, the more appropriate the action is. If two actions are equally close, then both, i.e. their conjunction, are appropriate. To illustrate why both should be appropriate together, consider a university lecturer in a teaching and research position is trying to decide what activities to do the next term. Being hired to do both excellent teaching and high-impact research while also understanding that this is the job, both obligations will be among the current beliefs, that is, in the centre of the sphere, possibly together with other obligations. It follows that both obligations are appropriate, and the appropriate thing to do is thus the conjunction of providing excellent teaching and producing high-impact research, not each of them separately.

This suggests the following perhaps *prima facie* overly complicated definition of N_{A_i} in which we have used $\exists$ for "there is" and $\forall$ for "for each":

- $N_{A_i}(w) = \{\bigcap\{v' \in Z \mid \exists X \in N_{E_i}(w)$ and $\exists Y \in N_C(w)$ such that $X \cap Y \neq \emptyset$ and $\forall v \in X \cap Y, Z \in N_{O_i}(v)$ and it is not the case that $\exists Z'$ such that $\exists X \in N_{E_i}(w)$ and $\exists Y \in N_C(w)$ such that $X \cap Y \neq \emptyset$ and $\forall v \in X \cap Y, Z' \in N_{O_i}(v)$ for which $\exists Y' \in N_C(w)$ such that $\exists v' \in Y'$ such that $Z' \in N_{O_i}(v')$ and it is not the case that $\exists w' \in Y'$ such that $Z \in N_{O_i}(w')\}\}$

Let us go through its various components:

- Only one action is the appropriate action, which means that there is, at every points of evaluation, one only neighbourhood for the neighbourhood function for appropriateness.

- This neighbourhood is defined as the intersection of the neighbourhoods representing the most relevant obligations, which means that the neighbourhood will be such that the conjunction of the relevant actions prescribed by the obligation hold.
- The most relevant obligations are those for which there is no other obligations supported by evidence such that the latter are closer to the centre of the spheres.

This is one way of articulating the idea that similarity is the keyword for determining what is appropriate, and which is rather natural within the framework we are using. However, since March and Olsen (1989) do not go into details with regard to this, it is difficult to adequately assess it and compare it with alternatives. It does capture the idea that the most appropriate action is that which is obligatory in the situation most similar to how we believe the situation is, but there are other features that can be compared with regard to similarity. One alternative proposal could be to define either the canonical or the last situation in which a rule was applied, and compare the agent's current beliefs with that situation. However, the canonical situation is sort of already contained in the antecedent of the conditional ought, and we have thus already included this thought in our proposal since we are comparing how similar the situation in which the evidence for the antecedent holds is to what the agent believes. Similarly, one can argue that the last situation in which the norm applied is also included in the antecedent of the conditional ought.

We also note that the definition does not currently exclude contradictory actions being appropriate if the appropriate action turns out to be complex. This follows from how there can be different pieces of evidence supporting equally near obligations which in turn are conflicting. This shows arguably a limit with similarity as criterion for appropriateness, and dissolving that would require us to improve the definition with an additional criterion. We could for example add the requirement that the neighbourhood cannot contain complimentary sets of points of evaluation. This modification is left for future work.

If we had defined not only a concept of obligation but also a separate concept of expectation, it could be reasonable to amend the definition of appropriateness by letting obligations have priority over expectations. This could be achieved by ensuring that, when both an obligation and an expectation hold in situations equally similar to the agent's current understanding of the situation, only the obligation is appropriate, not the expectation.

Finally, we note that, as in the case of obligations, what is appropriate might be unclear because the neighbourhood in question might not be such that it corresponds to the set of points of evaluation at which some formula is true. This follows from the fact that we simply used neighbourhoods of obligations and evidence to define it without introducing some mechanism to ensure that these neighbourhoods themselves are clear. Ambiguity in obligation leads to ambiguity in what's appropriate, but that seems to be in line with March and Olsen (1989).

4.3.5 Generalizing to multiple roles

With the neighbourhood function N_{A_i} we have made a first approximation towards capturing appropriateness relative to a particular role. It cannot be stressed enough that it is intended as a Galilean idealization, and that is why we can allow ourselves such a simple similarity-mechanism to pick out the most appropriate (possibly complex) action. However, as discussed above in section 4.2, we are not interested merely in capturing what is appropriate to do relative to a particular role, but relative to multiple role since an agent can at a given time play multiple roles and may, in a given situation, face a dilemma with regard to conflicting actions being appropriate according to different roles.

In the literature on frame and neighbourhood semantics involving for example the beliefs of multiple agents one introduces a set of indices where each index represents one agent. It is natural to apply the same strategy in our case, but to think of each index as representing a role. Our sets of roles shall be finite, and we shall use numerals to refer to indices but continue to use i and j as variables for them. We can furthermore also define multiple agents as subsets of the set of indices, or simply treat our models as concerning one agent currently playing each of the roles represented by the set of indices. We shall in this dissertation restrict our attention to one agent playing multiple roles to avoid some complications regarding the counterfactual conditional. We use upper-case Latin letters starting from G for subsets of the set of indices, and simply I for referring to the set of indices.

Following these considerations, we augment our language with new connectives as follows where I is a set of indices:

- For each non-empty subset G of I, $[E_G]$, $[O_G]$, $[A_G]$ are connectives (thus replacing $[E_i]$ and $[O_i]$ in addition to actually adding a connective for appropriateness as we only defined the neighbourhood function above)

In addition, we add neighbourhood functions for each index (but not subsets of I) as follows:

- For each $i \in I$, N_{E_i}, N_{O_i}, and N_{A_i} are neighbourhood functions defined as explained above.

These are then extended to neighbourhood functions for evidence and obligation for each non-empty subset G of I as follows:

- $N_{E_G} = \bigcup N_{E_i}$ for each $i \in G \in I$

- $N_{O_G} = \bigcup N_{O_i}$ for each $i \in G \in I$

This ensures that the evidence available to an agent with the roles G is the totality of evidence gathered through the lenses of each role, and their obligations is the totality of obligations from each role.

As expected, the neighbourhood function for appropriateness involving multiple roles will be slightly more complicated. While we used similarity to the agent's current beliefs to extract the appropriate action from the set of obligations belonging to one role, the same strategy doesn't seem fit in this case due to situations such as that described in section 4.2 regarding boarding a plane or finding one's missing daughter at the airport. Instead, we seem to have a priority-ordering of the roles independent thereof. This priority-ordering can be introduced either as fixed for one model or as relative to each point of evaluation (but in both cases relative to each collection G of I. If we assume that a model is only supposed to model one decision problem, then it seems sufficient to have the ordering for each G fixed for one model. We leave the exploration of the ordering relativized to each point of evaluation to future work.

In line with such considerations we associate with each G of I an ordering $\preceq$ on G, $\preceq_G$. We can articulate the idea that what is appropriate in the case of multiple roles is what is appropriate according to the role(s) with highest priority as follows:

- $N_{A_G} = \{\bigcap_{i \in G}\{w \in X \mid X \in N_{A_i}$ and it is not the case that $\exists j \in G$ such that $i \preceq j$ and $N_{A_j} \neq \emptyset\}\}$

The neighbourhood function of a collection of roles G is defined as the conjunction of the appropriate actions of the highest prioritized roles.

Since it's now the totality of evidence that is being considered, one could consider to amend the definition of appropriateness relative to a role to include not only the evidence related to a role, but any evidence. Consider for example someone working undercover in some organization. Information obtained through the role one plays within that organization can surely influence the choices one makes as undercover agent. Similarly, one can use evidence obtained through one's job in choices outside that job since the information obtained can be useful also for other roles. This is of course sometimes illegal; consider for example insider-trading. Still, this is something one *can* do, and also something which is easily fixed in our neighbourhood semantics. We shall however not explore that amendment here.

4.3.6 Summary and illustration

We now present the class of neighbourhood models intended to model a decision problem for one agent using the logic of appropriateness.

Let I be a finite set $\{1,2,3,\ldots,n\}$ and let $\mathbb{G}$ be a set of non-empty subsets of I. Let $\mathcal{L}_A$ be the set of formulas defined with a countable set of atomic formulas, the unary connectives $\{[E_G] \mid G \in \mathbb{G}\}$, $\{[O_G] \mid G \in \mathbb{G}\}$, $\{[A_G] \mid G \in \mathbb{G}\}$ and $\neg$ and the binary connectives $\rightarrow$, $\wedge$ and $\vee$.

A model M is $\langle W, \{N_{A_i} \mid i \in I\}, \{N_{O_i} \mid i \in I\}, \{N_{E_i} \mid i \in I\}, N_C, V\rangle$ such that W is a non-empty set of objects, V is a function assigning each atomic formula a subset of W, $\{N_{A_i} \mid i \in I\}, \{N_{O_i} \mid i \in I\}, \{N_{E_i} \mid i \in I\}$ and N_C are neighbourhood functions such that the following constraints hold:

- $N_{A_i}(w) = \{\bigcap\{v' \in Z \mid \exists X \in N_{E_i}(w)$ and $\exists Y \in N_C(w)$ such that $X \cap Y \neq \emptyset$ and $\forall v \in X \cap Y, Z \in N_{O_i}(v)$ and it is not the case that $\exists Z'$ such that $\exists X \in N_{E_i}(w)$ and $\exists Y \in N_C(w)$ such that $X \cap Y \neq \emptyset$ and $\forall v \in X \cap Y, Z' \in N_{O_i}(v)$ for which $\exists Y' \in N_C(w)$ such that $\exists v' \in Y'$ such that $Z' \in N_{O_i}(v')$ and it is not the case that $\exists w' \in Y'$ such that $Z \in N_{O_i}(w')\}\}$.
- For each $w \in W, \emptyset \notin N_{E_i}(w)$
- If $X \in N_C(w)$ is such that for all $Y \in N_C(w), X \subseteq Y$, then $X \in E_i(w)$
- For all $X, Y \in N_C(w)$, either $X \subseteq Y$ or $Y \subseteq X$
- If $X_1, X_2, \ldots \in N_C(w)$ then $X_1 \cup X_2 \cup \ldots \in N_C(w)$
- If $X_1, X_2, \ldots \in N_C(w)$ then $X_1 \cap X_2 \cap \ldots \in N_C(w)$

For each $G \in \mathbb{G}$, we define the neighbourhood functions $N_{A_G}, N_{O_G}, N_{E_G}$ satisfying the following constraints:

- $N_{E_G} = \bigcup_{i \in G} N_{E_i}$
- $N_{O_G} = \bigcup_{i \in G} N_{O_i}$
- $N_{A_G} = \{\bigcap_{i \in G}\{w \in X \mid X \in N_{A_i}$ and it is not the case that $\exists j \in G$ such that $i \preceq j$ and $N_{A_j} \neq \emptyset\}\}$

$\Vdash$ is defined as follows:

- For all atomic formulas p: $w \Vdash_M p$ if and only if $w \in V(p)$
- $w \Vdash_M \neg\phi$ if and only if $w \nVdash_M \phi$
- $w \Vdash_M \phi \vee \psi$ if and only if either $w \Vdash_M \phi$ or $w \Vdash_M \psi$
- $w \Vdash_M \phi \wedge \psi$ if and only if either $w \Vdash_M \phi$ and $w \Vdash_M \psi$
- $w \Vdash_M \Box\phi$ if and only if for each $v \in W$, $v \Vdash_M \phi$

- $w \Vdash [O_G]\phi$ if and only if there is $X \in N_{O_G}(w)$ such that $\forall v \in W$, if $v \Vdash \phi$ then $v \in X$

- $w \Vdash [A_G]\phi$ if and only if there is $X \in N_{A_G}(w)$ such that $\forall v \in W$, if $v \Vdash \phi$ then $v \in X$

- $w \Vdash [E_G]\phi$ if and only if there is $X \in N_{E_G}(w)$ such that if $v \in X$ then $v \Vdash \phi$

- $w \Vdash \phi \rightarrow \psi$ if and only if there is $X \in N_C(w)$ such that $v \Vdash \phi$ for some $v \in X$ and for every $v' \in X$, if $v' \Vdash \phi$ then $v' \Vdash \psi$

Truth-preserving entailment is defined as usual: for every model, for every point of evaluation at that model, if the premises are true then the conclusion is true.

We proceed to illustrate how the models can be used to model decision problems. As example we shall use the case of the father&child at the airport discussed above in section 4.2.

First, we assign some of the propositional letters a suitable meaning:

- p means "boarding has started"

- p' means "one is boarding the plane"

- q means "one's child is missing"

- q' means "one is looking for one's child"

The sentences could probably have been formulated with more care and precision, but they'll do for our purposes.

In our case we have two roles, so $I = \{1, 2\}$ where 2 is assigned the father-role and 1 is assigned the passenger-role. We let now $\mathbb{G}$ contain the sets $\{1\}$, $\{2\}$ and I which we call G, H and $G \cup H$ respectively. We also stipulate that $1 \preceq 2$. It is more important to be a father than an airline passenger.

In our frame we add only six objects in W, so $W = \{w_1, \ldots, w_6\}$. V is defined as follows for our four interpreted sentences:

- $V(p) = \{w_2, w_6\}$

- $V(p') = \{w_3\}$

- $V(q) = \{w_4, w_6\}$

- $V(q') = \{w_5\}$

Other sentences are assigned appropriate sets.

Finally, we need to stipulate the neighbourhoods for evidence, obligation and the conditional. To keep things simple, we shall for the conditional only define three spheres which we associate with each point of evaluation: For all $w \in W, N_C(w) = \{\{w_1\}, \{w_1, w_2, w_4\}, W\}$. The spheres satisfy the relevant conditions: closed under intersections and unions, and for each pair of them, one is contained in the other. We do a similar trick for evidence by keeping things uniform: $\forall w \in W$, $N_{E_1}(w) = \{W, \{w_2\}, \{w_1\}\}$ and $N_{E_2}(w) = \{W, \{w_4\}, \{w_1\}\}$. The inner-most sphere counts as evidence, together with W, both in tune with the requirements on the models. In addition, there's evidence for p within role 1 and evidence for q within role 2. The neighbourhood functions for obligations are defined as follows:

- $N_{O_1}(w_2) = \{V(p')\}$, otherwise empty for every $w \in W$
- $N_{O_2}(w_4) = \{V(q')\}$, otherwise empty for every $w \in W$

Let us see what this model tells us about the situation.

First, we observe that $w_2 \Vdash [O_G]p'$ and $w_4 \Vdash [O_H]q'$ since, in both cases, there is a deontic neighbourhood for G or H of the point of evaluation in question which contains all the points at which either p' or q' is true, respectively. Secondly, we note that at every $w \in W$, $w \Vdash p \to [O_G]p'$ and $w \Vdash q \to [O_H]q'$. In the first case, we reason as follows: there is a point of evaluation in a sphere in $N_C(w)$ such that p is true, namely w_2 in $\{w_1, w_2, w_4\}$, and furthermore so that whenever p is true at a point of evaluation in that sphere, which is only w_2, then $[O_G]p'$ is also true. The reasoning for the second case is analogous.

Turning our attention to what is appropriate in the situation, we shall first establish that p' is the appropriate action relative to $\{1\} = G$ at each $w \in W$. First, we find a neighbourhood of $N_{E_1}(w)$ and one of $N_C(w)$ which overlaps and $V(p')$ is in $N_{O_1}(w')$ for every $w' \in N_{E_1}(w) \cap N_C(w)$. While they have multiple overlapping neighbourhoods, only $\{w_1, w_2, w_4\} \in N_C(w)$ and $\{w_2\} \in N_{E_1}(w)$ satisfy this requirement since $N_{O_1}(w_2) = \{V(p')\}$, and there is no need to invoke the rest of the definition. Since N_{A_G} is defined as the intersection of the points of evaluation in the appropriateness neighbourhood of each highest prioritized $i \in G$, it follows that $N_{A_G} = N_{A_1}$, and thus that p' is appropriate relative to G. The analogous reasoning regarding q' will establish that q' is the appropriate action relative to $\{2\} = H$.

To establish what is appropriate relative to every role, that is $G \cup H$, we only need to take into consideration the priority-ordering according to which $1 \preceq 2$. It follows that q' is the appropriate action at each $w \in W$.

This must admittedly seem overly complicated way to conclude the same as we did in our discussion of the example we modelled, but it is worth pointing out that we have here spelled out the reasoning to some extent, and that the corresponding reasoning, if we had applied expected utility theory as logic of consequences to model the situation, it would have appeared equally complicated. Moreover, the models are not intended as a tool for decision-makers to use. Instead, it represents the logical relationship between the components in the logic of appropriateness.

This concludes the presentation of our first attempt at explicating the logic of appropriateness in neighbourhood semantics.

Chapter 5

Conclusions

Our aim has been to spell out the details of logic of appropriateness as a Weberian ideal type and furthermore provide a first attempt at a formalization thereof. However, being intended an explication thereof, the pressing issue regarding the formalization and the models is whether this exercise has been useful with regard to our understanding of the logic of appropriateness.

One lesson from this exercise could be the following: the analogous problem with the practical impossibility of determining the best action from a utilitarian perspective shows up also in the case of the logic of appropriateness. Even with our relatively simple measure of similarity which we used to determine which action was the most appropriate with regard to a role is such that it is, expect for in simple cases, practically impossible to follow up on. Ordering situations described by our evidence in spheres to determine which obligation is closer to how we actually understand a situation is surely as demanding as assigning probabilities to events and ensure that our preferences satisfy certain conditions. Spelling out the details of the logic of appropriateness illustrates what is required to determine the appropriate action, and it is doesn't seem to be an *easier* task than determining the best action from a consequentialist perspective. The logic of appropriateness has analogous idealization problems as the logic of consequences.

Continuing on this train of thoughts, one can say that the attempt to formalize the logic of appropriateness goes to show the difference between sketching an idea and working out the details of a proposal. March and Olsen (1989) do not actually work out the details of the logic of appropriateness to the extent that it can turned into a *logic*. We were able to do it by proposing one way to make precise their similarity-talk and adding a priority-ordering of roles, but without these tricks we wouldn't have obtained something worth calling a logic which would've been in the spirit of March and Olsen (1989).

We can thus conclude, at least preliminary, that one way in which formalizing the logic of appropriateness is useful lies in how it highlight these issues. This conclusion is only preliminary because the formalization presented in section 4.3 is not intended as a completed project. Here is a list of modifications that could be explored:

- One can modify the definitions of appropriateness by requiring that the neighbourhood cannot contain complimentary sets to rule out contradictions being appropriate.
- The priority-ordering can be relativized to each point of evaluation.
- We can modify the definition of appropriateness relative to a role by using the totality of evidence rather than evidence gathered by a role
- The counterfactual conditional can be relativized to multiple agents.
- The models as such can also be extended to deal with multiple agents.

This list is not exhaustive. In addition, it wouldn't surprise me that the various definitions we have presented come with what we can hopefully describe as "childhood diseases", various issues that must be fixed but which I haven't discovered yet.

I think that there is more to the usefulness of a formalization of the logic of appropriateness than what we have concluded here, but exploring that would take us beyond the scope of a master's thesis. For example, to which extent can we use the models or modifications of the models presented in section 4.3 to evaluate and possibly criticize applications of the logic of appropriateness within organizational theory? What are its connections with research on practical reasoning in philosophy? With research on deontic and epistemic logic?

Bibliography

Anglberger, A. J. J., Gratzl, N. and Roy, O. (2015), 'Obligation, free choice, and the logic of weakest permissions', *Review of Symbolic Logic* **8**(4), 807–827.

Bagaria, J. (2017), Basic set theory, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2017 edn, Metaphysics Research Lab, Stanford University.

Berto, F. (2018), 'Simple hyperintensional belief revision', *Erkenntnis* .

Bicchieri, C. (2004), Rationality and game theory, *in* P. Rawling and A. R. Mele, eds, 'The Oxford Handbook of Rationality', Oxford University Press, pp. 182–205.

Briggs, R. (2017), Normative theories of rational choice: Expected utility, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2017 edn, Metaphysics Research Lab, Stanford University.

Carnap, R. (1930), 'Die alte und die neue logik', *Erkenntnis* **1**(1), 12–26.

Carnap, R. (1950), *Logical Foundations of Probability*, Chicago]University of Chicago Press.

Carnap, R. (1956), *Meaning and Necessity. Second Edition*, University of Chicago Press.

Carnap, R. (1966), *Der Logische Aufbau der Welt. Dritte Auflage*, Meiner Verlag.

Christensen, T., Lægreid, P., Roness, P. G. and Røvik, K. A. (2015), *Organisationsteori for offentlig sektor*, Universitetsforlaget.

Creath, R. (2017), Logical empiricism, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', fall 2017 edn, Metaphysics Research Lab, Stanford University.

Decew, J. W. (1981), 'Conditional obligation and counterfactuals', *Journal of Philosophical Logic* **10**, 55–72.

Eckhoff, T. and Jacobsen, K. D. (1960), *Rationality and Responsibility in Administrative and Judicial Decision-making*, Munksgaard.

Frege, G. (1879), *Begriffsschrift: Eine Der Arithmetische Nachgebildete Formelsprache des Reinen Denkens*, L. Nebert.

Frigg, R. and Hartmann, S. (2017), Models in science, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2017 edn, Metaphysics Research Lab, Stanford University.

Goldmann, K. (2005), 'Appropriateness and consequences: The logic of neo-institutionalism', *Governance: An International Journal of Policy, Administration, and*

Institutions **18**(1), 35–52.

Horsten, L. and Douven, I. (2008), ‘Formal methods in the philosophy of science’, *Studia Logica* **89**(2), 151–162.

Kripke, S. A. (1959), ‘A completeness theorem in modal logic’, *Journal of Symbolic Logic* **24**(1), 1–14.

Lewis, D. K. (1973), *Counterfactuals*, Blackwell.

March, J. G. (2009), *Primer on Decision Making: How Decisions Happen*, The Free Press.

March, J. G. and Olsen, J. P. (1989), *Rediscovering Institutions: The Organizational Basis of Politics*, The Free Press.

March, J. G. and Olsen, J. P. (1996), ‘Institutional perspectives on political institutions’, *Governance: An International Journal of Policy and Administration* **9**(3), 247–264.

March, J. G. and Olsen, J. P. (1998), ‘The institutional dynamics of international political orders’, *International Organization* **52**(4), 943–969.

March, J. G. and Olsen, J. P. (2011), The logic of appropriateness, *in* R. E. Goodin, ed., ‘The Oxford Handbook of Political Science’, Oxford University Press, pp. 478–495.

Müller, T. (2010), Formal methods in the philosophy of natural science, *in* F. Stadler, D. Dieks, W. Gonzales, S. Hartmann, T. Uebel and M. Weber, eds, ‘The Present Situation in the Philosophy of Science’, Springer, pp. 111–123.

Negri, S. (2005), ‘Proof analysis in modal logic’, *Journal of Philosophical Logic* **34**(5-6), 507–544.

Pacuit, E. (2017), *Neighborhood Semantics for Modal Logic*, Springer.

Persson, T. (2007), ‘Explaining european union adjustments in sweden’s central administration’, *Scandinavian Political Studies* **30**(2), 204–228.

Russell, B. and Whitehead, A. N. (1910), *Principia Mathematica Vol. I*, Cambridge University Press.

Røvik, K. A. (2007), *Trender og Translasjoner: Ideer som former det 21. århundrets organisasjon*, Universitetsforlaget.

Schnädelbach, H. (1998), ‘Rationalitätstypen’, *EuS* **1**, 79–89.

Scott, J. and Marshall, G. (2009), *A Dictionary of Sociology*, Oxford Dictionary of Sociology, Oxford University Press.

Steele, K. and Stefánsson, H. O. (2016), Decision theory, *in* E. N. Zalta, ed., ‘The Stanford Encyclopedia of Philosophy’, winter 2016 edn, Metaphysics Research Lab, Stanford University.

van Benthem, J., Fernández-Duque, D. and Pacuit, E. (2014), ‘Evidence and plausibility in neighborhood structures’, *Annals of Pure and Applied Logic* **165**(1), 106–133.

Weber, M. (1978), *Economy and Society: an outline of interpretative sociology*, University of California Press.