Bachelor's Thesis

Trust Logics and Their Horn Fragments: Formalizing Socio-Cognitive Aspects of Trust

Karl Nygren

LiTH - MAT - EX - - 2015/01 - - SE

Trust Logics and Their Horn Fragments: Formalizing Socio-Cognitive Aspects of Trust

Department of Mathematics, Linköpings universitet

Karl Nygren

LiTH - MAT - EX - - 2015/01 - - SE

Examensarbete: **16 hp**

Level: **G2**

Supervisor: **Andrzej Szałas**,
Department of Computer and Information Science, Linköpings universitet

Examiner: **Anders Björn**,
Department of Mathematics, Linköpings universitet

Linköping: **March 2015**

Abstract

This thesis investigates logical formalizations of Castelfranchi and Falcone's (C&F) theory of trust [9, 10, 11, 12]. The C&F theory of trust defines trust as an essentially mental notion, making the theory particularly well suited for formalizations in multi-modal logics of beliefs, goals, intentions, actions, and time.

Three different multi-modal logical formalisms intended for multi-agent systems are compared and evaluated along two lines of inquiry. First, I propose formal definitions of key concepts of the C&F theory of trust and prove some important properties of these definitions. The proven properties are then compared to the informal characterisation of the C&F theory. Second, the logics are used to formalize a case study involving an Internet forum, and their performances in the case study constitute grounds for a comparison. The comparison indicates that an accurate modelling of time, and the interaction of time and goals in particular, is integral for formal reasoning about trust.

Finally, I propose a Horn fragment of the logic of Herzig, Lorini, Hübner, and Vercouter [25]. The Horn fragment is shown to be too restrictive to accurately express the considered case study.

Keywords: trust, modal logic, multi-agent systems, Horn fragment

URL for electronic version:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-115251

Abstract in Swedish: Sammanfattning

I den här uppsatsen undersöker jag logiska formaliseringar av Castelfranchi och Falcones (C&F) teori om tillit [9, 10, 11, 12]. C&F definierar tillit som en form av mental attityd, vilket gör teorin väl lämpad för att formaliseras i multimodala logiska system som tar trosföreställningar, mål, intentioner, handlingar och tid i beaktning.

Tre sådana logiska system presenteras, jämförs och utvärderas. Jag definierar viktiga begrepp ur C&Fs teori, och bevisar egenskaper hos dessa begrepp. Egenskaperna jämförs sedan med de informellt definierade egenskaperna hos C&Fs tillitsteori. De logiska systemen används därefter för att formalisera ett testscenario, och systemen jämförs med testscenariot som utgångspunkt. Jämförelsen visar att en noggrann modellering av interaktionen mellan tid och agenters mål är viktig för formella tillitsmodeller.

Slutligen definierar jag ett Horn-fragment av Herzig, Lorini, Hübner och Vercouters [25] logik. Horn-fragmentet visar sig vara för restriktivt för att formalisera alla delar av testscenariot.

Nyckelord: tillit, modallogik, multiagentsystem, Horn-fragment

Acknowledgements

I would like to thank my supervisor Andrzej Szałas and my examiner Anders Björn for their continuous support during the writing of this thesis. Thanks also to Mikael Skagenholt for proofreading, and to Linda Rydström for her useful comments.

Contents

Chapter 1

Introduction

Trust is important in all kinds of social situations where any sorts of agents—human or artificial—interact. Such situations include, for instance, human-computer interaction, multi-agent systems (MAS), and Internet communications like Internet forums, chat rooms, and e-commerce.

There are many approaches to trust found in the literature. One of the most widely used is the approach proposed by Gambetta, where trust is defined as the particular level of subjective probability with which a trusting agent assesses that another agent will perform a certain task [22, p. 217]. According to this approach, trust is no more than the result of repeated interaction between a trusting agent and a trusted agent; iterated experiences of the trusted agent's success strengthen the trusting agent's trust, while iterated failures decrease the amount of trust.

In this thesis, I will adopt another approach, proposed by Castelfranchi and Falcone (hereafter referred to as C&F). They have developed a cognitive theory of trust (referred to as the C&F theory), where trust is defined in terms of beliefs and desires (see [9, 10, 11, 12]). Thus, trust is essentially a mental notion. In order to trust someone or something Y, one has to believe that Y is capable of doing what one needs and wants to be done, and one has to believe that Y actually will do it; Y's behaviour has to be predictable.

This position lays focus on trust as an *attitude* directed at possible trustees, rather than mere subjective probability or risk assessment, which makes the theory easily embedded in a belief-desire-intention (BDI) framework [6, 8, 21]. BDI frameworks, in turn, are particularly well suited for modal logical formalizations.

The topic of this thesis is such formalizations of the C&F theory using modal logic.

1.1 Goals and structure of thesis

This thesis is written with a two-fold purpose. The first is to compare and evaluate three proposed logical formalisms, all intended to formalize the C&F theory of trust. This is done both by direct comparison of how well the three logics capture key aspects of the C&F theory, and by comparison of how well the logics perform in a proposed scenario.

The second purpose is to define a Horn fragment of one of the logics, and investigate whether the Horn fragment is expressive enough to formalize said scenario. Horn fragments of logics are important in practical applications, for example in logic programming and deductive databases.

Chapter 2 will introduce the reader to the C&F theory of trust. Chapter 3 presents the three considered logics. The chapter highlights important properties of the three logics, which are then compared and evaluated in relation to the C&F theory of trust. There is also a short introduction to modal logic. In Chapter 4, a case study is developed with the purpose of highlighting specific properties of the logics in relation to an experimental scenario. In Chapter 5, the general Horn fragment of one of the logics is defined. The Horn fragment is then used to formalize the scenario presented in Chapter 4, in order to indicate that the logic could be used as an implementable theoretical framework for reasoning about trust. The final chapter contains conclusions and ideas for further research. An appendix containing a short survey of propositional logic is also included.

Chapter 2

The C&F theory of trust

In this chapter, I will pursue a detailed description of the C&F theory, but first, the concepts of *agents* and *multi-agent systems* will be presented.

2.1 Agents, mental attitudes, and interactions

The C&F theory is developed, not only as a BDI theory, but also within a multi-agent system framework (see [44] and [45] for introductions to MAS).

An *agent* is an "individual entity capable of independent action" [16, p. 51]. In a broad sense, this includes humans as well as software systems.

I will consider agents with *mental attitudes*; in particular, agents that are capable of *beliefs*, *goals*, and *intentions*.[1] These are integral concepts in the theory of practical reasoning (see e.g. [6, 21]), and the starting point for the C&F theory.

The *beliefs* of an agent correspond to the information the agent has about its environment, including information regarding other agents. The *goals* of an agent are the circumstances that the agent would choose (or, the circumstances that the agent prefers to bring about). Intentions are a "special consistent subset of an agent's goals, that it chooses to focus on for the time being" [21, p. 4].

2.1.1 Multi-agent systems

Multi-agent systems (MASs) are

> computational systems in which a collection of loosely coupled autonomous agents interact in order to solve a given problem. As this problem is usually beyond the agents' individual capabilities, agents exploit their ability to *communicate, cooperate, coordinate*, and *negotiate* with one another [21, p. 1].

[1] I will use the term 'goal' instead of 'desire'; however, I will use the abbreviation BDI (belief-desire-intention), as it is the conventional term. It should also be noted that there are many theories of practical reasoning; for instance, Bratman [6] takes intentions to be first-class citizens in agency, while the logical formalizations considered in the later chapters of this thesis reduce intentions to preferred actions. I will not dwell on this distinction here.

Note that this definition is (deliberately) inexact. This is because there are many competing definitions of MASs found in the literature. For my purposes here, it suffices to note that MASs are systems of interactive agents. C&F argue that most kinds of social interactions requires some kind of *delegation*[2]—letting other agents do things for you and acting on behalf of other agents—and that delegation more or less require trust in many cases [9]. Thus, trust is an absolutely integral part of MASs.

It should be noted that, even though MASs concern computer science applications and the C&F theory is developed within a MAS framework, much of the theory is abstract and interdisciplinary, and applies to, for instance, inquiries about human trust reasoning.

2.2 Trust as a five-argument relation

According to the C&F theory, trust is a five-argument relation. The five arguments are [12, p. 36]:

- A trusting agent X. X is necessarily a cognitive, intentional agent (see Section 2.2.1). I will often refer to the trusting agent as the *truster*.

- An entity Y (object or agent) which is the object of the truster's trust. Y will often be referred to as the *trustee*.

- A goal g_X of the truster. It will be useful to think of the goal g_X as a logical formula representing a certain state of affairs.

- An action α by which Y possibly can bring about a certain state of affairs, represented by formulas in a set P of which the goal g_X is an element.

- A context C under which X considers trusting Y.

The context component will usually be omitted in order to simplify the presentation. The context is still important for the theory, so alternative ways of incorporating it in the trust relation will be considered.

Following C&F, I will use the predicate

$$\text{TRUST}(X, Y, g_X, \alpha, C)$$

or, when the context is omitted,

$$\text{TRUST}(X, Y, g_X, \alpha)$$

to denote the mental state of trust.

The five arguments in the trust relation will be analysed in more detail below.

[2] See e.g. [17, 18].

2.2.1 The truster and the trustee

The trusting agent is necessarily a cognitive, intentional entity. This means that the truster is capable of having mental attitudes: beliefs, intentions, and goals [21]. Since trust by definition consists of beliefs and goals, an entity incapable of beliefs and goals is also incapable of trust.

The trustee is not, unlike the truster, necessarily an intentional entity. This becomes clear when one considers uses of 'trust' like in the sentence "I trust my alarm clock to wake me up at 6 p.m." or when one trusts a seemingly rickety chair to hold one's weight. The trustee can thus be either another agent, or an object like an alarm clock (see [15] and Section 2.3.3 below on trust in intentional versus trust in non-intentional entities.)

2.2.2 The goal component

The C&F theory stresses the importance of the goal-component in the notion of trust. Basically, a truster cannot trust a trustee without the presence of a goal: when X trusts Y, X trusts Y *for doing that and that*, or trusts Y *with that and that*, etc.

It will be useful to think of the goal g_X as a *logical formula* representing a certain preferred state of affairs.

The goal component is necessary, since it distinguishes trust from mere foreseeing or thinking. Following C&F [11], the combination of a goal and a belief about the future is called a *positive expectation.* This means that X both wants g_X to be true (has the goal g_X) and believes that g_X will be true. Thus, trust is a positive expectation rather than a neutral belief about the future, i.e. a forecast.

A goal need not be an explicit goal of the truster; it need not be a goal which X has incorporated in her active plan. Thus, the notion 'goal' as used in this context differs from the common language use, where 'goal' often refers to a "*pursued external objective to be actively reached*" [12, p. 46]. C&F [12, p. 46] use the following example of a case where the goal of a truster X is not explicit. An agent X can trust another agent Y to pay her taxes, but X might not have the explicit goal of Y paying her taxes. However, if X learns that Y is in fact not paying her taxes, X might be upset and angry with Y. According to C&F, that X in fact has the goal of Y paying her taxes is the reason why X would be upset with Y. This example also highlights the fact that goals need not be pursued goals. An agent X can have goals which she does not pursue or wish to pursue.

Thus, a goal exists even before it is made explicit and active. In summary, the following are the important properties of the goal component [12, p. 47].

- Every instance of trust is relative to a goal. Before deciding to trust and delegate to a trustee, possible trustees are evaluated in relation to a goal, that is not yet active or pursued.
- When a truster decides to trust and delegate to a trustee, the goal becomes pursued: an active objective in the truster's plan.

Even though the goal could be the only consequence of a certain action, in many cases the action in question results in several effects. Thus, the goal g_X is considered to be an element of the set P of results of an action: $g_X \in P$.

The fact that every action potentially results in several side effects, apart from the desired goal state, complicates the analyses of particular cases. Therefore, in order to simplify the analyses of formalizations, this detail will often be ignored.

2.2.3 The action component

The action α is the action which X believes can bring about the desired state of affairs P with $g_X \in P$. The action is a causal interaction with the world which results in a certain state of affairs; after α has been performed, the formulas in a set P holds.

2.2.4 The context

The context in the above trust relation is the context or scenario where Y is a candidate for X's trust, and where the external factors allow Y to perform the required action. The context is important, since different contexts can produce different trust-relations. For example, under normal circumstances, X might trust Y with an action α and a goal g_X to a high degree, but under extreme circumstances, for example if Y has to act in a war zone, X might trust Y to a smaller degree.

The analysis of the context component can be further refined by distinguishing between two kinds of contexts [12, p. 83]: The context of X's evaluation, and the context in which Y performs α. The first kind of context, the evaluation context, involves such considerations as the mood of X, her social position, her evaluation of Y's social position, her beliefs, etc. The second kind of context, the execution context, involves things such as the physical environment in which Y performs α, the social environment; including form of government, norms, social values, etc.

When formalizing the C&F theory logically, the context component will often be omitted as an argument in the trust relation. Instead, other ways to (partially) incorporate the context can be used, or one can assume a specific context.

Castelfranchi *et al.* [13] use internal and external preconditions, in the sense that X can only trust Y if the internal preconditions and external preconditions for Y to perform the required action are fulfilled. For example, if Bob trusts Mary to shoot Bill, then Bob believes that, for example, Mary's arm is not paralysed, Mary's eyes are apt for aiming etc. (these are examples of internal preconditions, that is, things that are internal to the trustee), and Bob believes that no one will block Mary's shooting by knocking the gun out of her hand, no obstacle will be in the way, etc. (these are examples of preconditions that are external to the trustee).

2.3 Trust as a layered notion

Trust in C&F theory is a layered notion, which means that the different notions associated with trust are embedded in each other. As mentioned earlier, trust is essentially a mental notion, but the concept of 'trust' can also be used in contexts

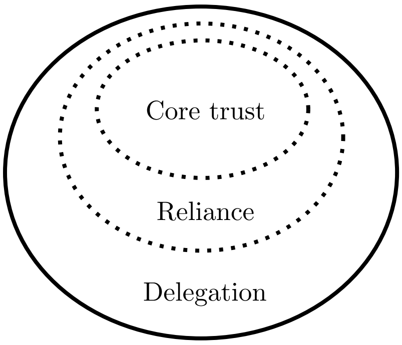

Figure 2.1: The layered stages of trust.

like "intending to trust" or "decide to trust". According to C&F theory, there are three different stages of trust [12, pp. 36, 64–65]:

- Trust is essentially mental, in the sense that it is an *attitude* towards a trustee; it is a certain *belief* or *evaluation* about the trustee's capability and willingness to perform a certain task in relation to a goal. In short, trust in this most basic sense is a *disposition* of a truster towards a trustee. I will follow C&F and refer to this part of the trust concept as *core trust*.
- Trust can also be used to describe the *decision* or *intention* to delegate an action to a trustee. This is a mental attitude, just like the above *core trust*. This dimension actually involves two very similar but distinct notions: *reliance* and *decision to trust*.
- It can also be the actual delegation of an action, the *act* of trusting a trustee with an action. This is a result of the above decision; the decision has here been carried out. This part of the trust relation is called *delegation*.

The embedded relation between these three parts of the notion of trust is illustrated in Figure 2.1.

2.3.1 Trust as act and trust as mental attitude

As seen above, in common language 'trust' is used to denote both the mental attitude of trust and the act of trusting, the delegation. According to C&F theory, core trust and reliance trust are the mental counterparts to delegation [9, 11]. This means that core trust and reliance trust are strictly mental attitudes, preceding delegation.

Trust as positive evaluations and as positive expectations

According to C&F [12, p. 43], every case of trust involves some attribution of internal skills to the trustee. When these attributions, or evaluations, of the trustee are used as a basis for the decision of trusting and/or delegating, the positive evaluations form the basis of core trust.[3]

The positive evaluations in trust involves two important dimensions:

[3] Note that these evaluations are relative to a goal of the truster: "Y is good for...".

- Competence (capability). The competence dimension involves attribution of skills, know-how, expertise, knowledge, etc.; that is, the attribution of internal powers relevant to a certain goal and a certain action to a trustee Y.

- Predictability and willingness (disposition). This dimension consists of beliefs about the trustee's actual behaviour, rather than beliefs about her potential capability of performing an action. It is an evaluation about how the trustee will behave: not only is Y capable of performing α, but Y is actually *going to do* α.

For X to trust Y, X should have a positive evaluation of Y's capability in relation to g_X and α, as well as a positive evaluation of Y's actual behaviour in relation to g_X and α. However, trust should not be understood as completely reducible to positive evaluations; such a move would completely ignore the motivational aspect, i.e. the goal component. So, more or less explicit and intentional positive evaluations are necessary for trust, but not sufficient. One can, for example, have a positive evaluation of an agent Y, without necessarily having trust in Y.

C&F state that trust is a *positive expectation* about the future. The positive evaluations are used as a base for making predictions about the behaviour of the trustee, and therefore also about the future. However, predictions and expectations are not synonyms; an expectation is a prediction that is relevant for the agent making the prediction, and the predicting agent wants to verify the prediction: she is "*waiting* in order to know whether the prediction is true or not." [12, p. 54] Thus, a *positive expectation* is a prediction about the future in combination with a goal: when X trusts Y with α and g_X, X both wants g_X to be true, and believes that it will be true thanks to Y's performance of α.

A positive evaluation of Y's willingness in relation to a certain task α ("Y is going to perform α") is merely a prediction of Y's behaviour if the goal-component is missing; if the performance of α (or preventing the execution of α) is not a goal for the agent doing the evaluation, then X has a neutral expectation about the future world-state.

It is also important to note the quantitative aspect when talking about trust as positive expectations (see also Section 2.4.) This becomes clear when one considers sentences like "I hope that Y does α" and "I trust that Y does α." Both sentences can be said to be positive expectations about the future. However, there is, according to C&F [12, p. 60], a difference in degree. When I *trust* that Y will perform α, I am quite certain that α actually will be performed (and thus realizing my goal g), while when I *hope* that Y will perform α, my positive evaluations about Y are uncertain; I am not actively counting on Y to perform α.

Core trust

As we have seen, the most basic notion in the trust relation is the core trust. I will now further develop the analysis of this notion. Core trust is, as said, a mental attitude, a belief or evaluation of a trustee's capability and intention of performing a certain task. Thus, core trust can be defined in the following way: A truster X has *core trust* towards (*trusts*) a trustee Y if and only if

1. X has the goal g_X,
2. X believes that
   (a) Y, by performing an action α, can bring about the state of affairs P with $g_X \in P$,
   (b) Y has the capability to perform α,
   (c) Y will perform α.

Core trust is “here-and-now” trust, in the sense that the truster trusts the trustee in relation to an active goal—a goal that is had by the truster “here-and-now”—and an action which the trustee can and will perform in the near—or in any case definitive—future. As Herzig *et al* [25, pp. 12–13] point out, weakening the definition of trust by only requiring that Y performs α eventually raises problems with for example procrastination.

This also means that core trust is a *realization* of en evaluation of a trustee in relation to a goal that is (has become) active. An evaluation can, however, be done in relation to a not yet active goal. Such an evaluation is called a *trust disposition.*

Trust disposition

As seen, core trust is trust “here-and-now”: if a truster has core trust in a trustee in relation to an action α and a goal g, the truster expects the trustee to perform α in the near future. However, this notion does not capture the nature of all trust relations. C&F claim that one also has to consider *trust dispositions* [12, p. 68]. The notion of core trust is an *actualization* (realization) of a trust disposition.

Consider the following example. A manager X wants to recruit a new employee. The decision to appoint a particular employee Y is based on trust; if X appoints Y, X trusts that Y will perform all required tasks. Recruiting an employee is (most typically) a long term investment, and as such it should be based on a broad evaluation of the trustworthiness of the employee in relation to several tasks and goals. In addition, the manager probably wants to recruit someone who is flexible and capable of performing several tasks to accomplish goals that are not relevant at present time, but might become relevant in the future.

When considering the above example, it becomes clear that the trust relations underlying the manager’s decision to employ a certain agent Y cannot only be of core trust type; the manager also considers her evaluation in relation to potential goals.

C&F does not precisely define the notion of trust disposition. However, they state that it is a property of trust dispositions that they can underlie core trust [12, p. 68]. Using the pseudo logical notation from Figure 2.2, the evaluation underlying a trust disposition can be expressed as the two beliefs

$$\mathrm{Bel}_X(\mathrm{CanDo}_Y(\alpha))$$

and

$$\mathrm{Bel}_X(k \rightarrow \mathrm{WillDo}_Y(\alpha) \wedge \mathrm{After}_\alpha(P))$$

where k is the circumstance that activates Y's performance of α to ensure P. k is something like "X asks Y to..." or "if such-and-such happens...", etc.

From the above two beliefs, the belief that k holds, and the goal that $g_X \in P$, the truster X moves from a trust disposition to actual core trust.

Reliance trust

First of all, for X to rely on Y, X has to believe that she is dependent on Y (i.e. X holds a *dependence belief*) to realize her goal g_X [11].

The dependence belief can take several forms: it could be a strong dependence belief, which means that X believes that g_X cannot be realized without the help of Y, or it could be a weak dependence belief, where X believes that g_X can be realized without Y, for example if X herself performs the required action α, but the delegation of the α to Y fits better into X's overall plan. That g_X is to be realized by Y's action, instead of X performing the action herself, is not exclusive; there is at least a third possibility: that the action α could be performed by another agent Z. Thus, C&F states that in order to decide to delegate an action α to any other agent, X must form the goal that she does not wish to perform α herself. Furthermore, X wishes Y to perform α, and not any other possible trustee. In summary, a truster X relies on a trustee Y if X decides to pursue the goal g_X through Y, rather than bringing it about herself, and does not search for alternative trustees [9]. Thus, reliance trust can be defined in the following way: A truster X has *reliance trust* towards (*relies* on) a trustee Y, if and only if

1. X has the goal not to perform action α,
2. X has the goal to let Y perform action α,
3. X believes X is dependent on Y.

In Figure 2.2, the ingredients of core trust and reliance trust is represented and simplified with some pseudo logical notation.

Delegation

Delegation is necessarily an *action*, something which causally interacts with the world to produce a certain state of affairs. According to C&F theory, core trust and reliance are the mental counterparts to delegation, which underlie and explain the act of trusting [11, 9].

The relationship between core trust, reliance and delegation

If a truster X actually delegates an action to a trustee Y, then (usually) X has the mental attitudes of core trust and reliance trust, and if X has the core and reliance trust attitudes towards Y in relation to a certain goal and action, then (usually) X delegates to Y.

Under ideal conditions, i.e. conditions where a truster X freely decides to trust without the interference of external constraints, the following holds [12, p. 38]: X relying on Y implies X having core trust in Y and X delegating to Y implies X relying on Y.

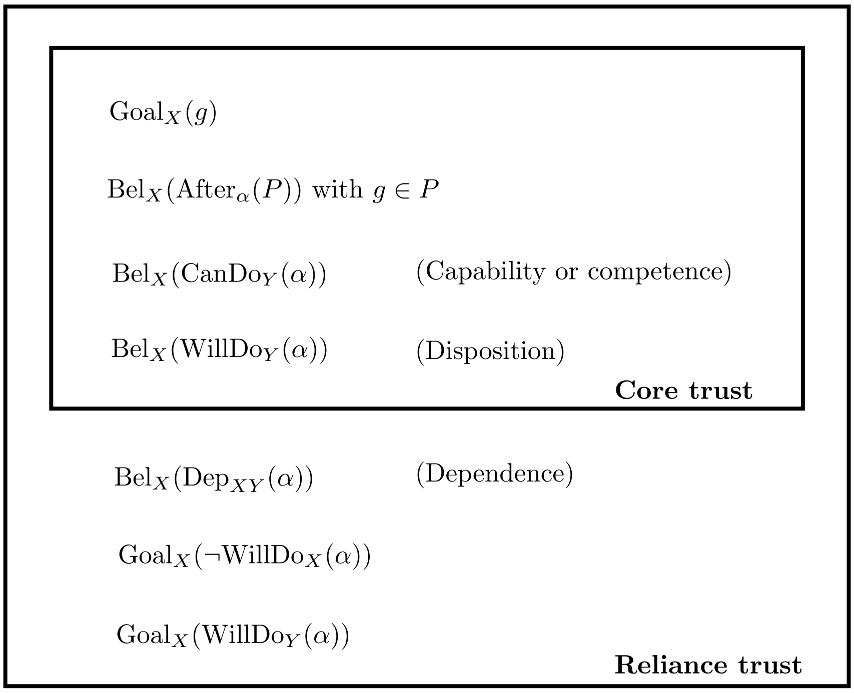
$\text{Goal}_X(g)$
$\text{Bel}_X(\text{After}_\alpha(P))$ with $g \in P$
$\text{Bel}_X(\text{CanDo}_Y(\alpha))$
(Capability or competence)
$\text{Bel}_X(\text{WillDo}_Y(\alpha))$
(Disposition)
Core trust
$\text{Bel}_X(\text{Dep}_{XY}(\alpha))$
(Dependence)
$\text{Goal}_X(\neg\text{WillDo}_X(\alpha))$
$\text{Goal}_X(\text{WillDo}_Y(\alpha))$
Reliance trust

Figure 2.2: The ingredients of core trust and reliance trust.

In most cases, however, agents do not act under ideal conditions. For example, under normal circumstances, every truster has the external constraint of not being able to evaluate all possible trustees. There can also be extreme circumstances, where actual delegation is prohibited, or X is forced to delegate. Thus, X could delegate to Y without trusting (as core trust and decision to trust) Y, and decide to trust Y without delegating to Y.

Extreme circumstances can be handled by the model by acknowledging that the *decision* to delegate implies the *decision* to rely on, which implies having core trust. So, under "normal" circumstances, core trust is necessary but not sufficient for reliance, and reliance is necessary but not sufficient for delegation. This also means that *reliance* must carefully be distinguished from *decision to trust.*

Decision to trust

In the above section, it was shown that reliance is a distinct notion from the decision to trust. In C&F theory, the decision to trust involves the mental attitudes of core trust, together with the decision to make use of a trustee as a part in an overall plan. Decision to trust can be defined as [13, p. 64]: *X decides to trust Y* if and only if

1. X has core trust in Y,

2. X relies on Y.

2.3.2 The difference between trust, distrust, mistrust and lack of trust

There are important differences between the concepts of distrust, mistrust and lack of trust [12, 34, 42, 43]. First, there is a difference between $\neg\text{TRUST}(X, Y, g_X, \alpha)$, which is the negated trust predicate, and *lack of trust*. Lack of trust is, according to C&F [12, p. 119] when a truster has no positive or negative evaluation of a trustee. The truster simply does not know if she trusts or distrusts the trustee. Thus, lack of trust must be defined as a lack of belief about the trustee's capability and predictability.

The concept of *distrust* is grounded in evaluations that show the trustee's incapability or unwillingness in relation to a certain goal and action; which means that the definition of distrust must include a belief about Y's incapability or unwillingness to perform α.

Further, the concept of *mistrust* is grounded in a negative evaluation of the trustee; the truster believes that the trustee is capable and willing to do the opposite of the truster's goal. In broader terms [12, p. 118], the trustee Y is capable and willing, is good and powerful, but for the wrong things (i.e. the opposite of X's goals).

It is important to stress that, even though neither distrust, mistrust or lack of trust equates with $\neg\text{TRUST}(X, Y, g_X, \alpha)$, trust and distrust, trust and mistrust, and trust and lack of trust are mutually exclusive, i.e. one cannot both trust and distrust another agent, etc. [43].

2.3.3 Trust in intentional and non-intentional entities

Recall that the trusting agent must be an entity capable of mental attitudes, while the object of trust need not be capable of mental attitudes. It could, however, be argued that even a simple device such as an alarm clock is an entity capable of (basic) mental attitudes. At least, one often ascribes such attitudes to certain entities. For example, it seems perfectly natural to say "my alarm clock believes that the time is 7 p.m.".

Thus, there might not be a meaningful distinction between, for example, trust in humans and trust in alarm clocks, at least not on the most basic level of core trust. Indeed, according to the C&F theory, all instances of core trust are grounded in evaluations about willingness and capability, independently of the object of trust. There is no principal difference between trust in agents and trust in entities that are not agents.[4]

In the following formalizations of the C&F theory, I will focus on trust in *agents*, which is why the willingness dimension will be expressed in terms of *intentions*.

2.4 Quantitative aspects

In previous sections, trust has been analysed from a qualitative point of view. However, it is a basic fact that trust can be graded. For example, X might

[4] It is important to stress that this is a property of *core trust*; more complex forms of social trust might put further requirements on trusting agents, for example the capacity for higher order mental attitudes.

trust Y to a certain degree in relation to g_X, and trust Z to a certain degree. Then, when deciding which agent she should rely on, X can compare the degree of trust in Y and Z, and then choose who she should rely on based on that comparison.

C&F [9] claim that there is a strong coherence between their cognitive definition of trust and the degree or strength of trust. The definition of trust as a conjunction of the truster's goal and beliefs about the trustee's capability and willingness allows a quite natural analysis of strength of trust; in trust relations, different levels of uncertainty of X's attribution of properties to Y leads to different levels of trust in Y. This can be illustrated by considering some common language uses of the components in the trust definition: "I am certain that Y can perform α", "I have reasonable doubt about Y's willingness to perform α, but Y is the only one that is competent enough." If one is certain about Y capability, as in the first example, one is more inclined to trust Y; the risk of Y failing to perform α is considered to be very small. In the second example, the trust in Y is probably quite low; there is a significant risk that Y will never perform α.

According to C&F [9], the degree of trust is a function of the beliefs about the trustee's capability and willingness: The stronger the belief, the greater the trust. This can also be put as the degree of trustworthiness of Y in relation to α: Stronger beliefs of X about Y's capability and willingness to do α, makes X consider Y more trustworthy in relation to α [12].

The notation $DoT_{XY\alpha}$ is introduced as the degree of X's trust in Y about α, and DoC_X denotes the degree of credibility of X's beliefs, where "credibility" means the strength of X's beliefs [9, 10]. It is now possible to express the degree of X's trust in Y about α as a function of the degree of credibility of X's beliefs about Y (using the pseudo logical notation from Figure 2.2):

$$DoT_{XY\alpha} = DoC_X(\mathrm{After}_\alpha(g)) * DoC_X(\mathrm{CanDo}_Y(\alpha)) * DoC_X(\mathrm{WillDo}_Y(\alpha)) \tag{2.1}$$

By using a utility function, measuring the utility of not delegating versus delegating an action to Y, the *trust threshold*—the threshold for when core trust in Y is strong enough to motivate a decision to trust and/or delegating to Y—can be decided.

An agent X has three choices in every situation involving a goal g_X and an action α [12, p. 102]: to try to achieve g_X by performing α herself, or to delegate the achievement of g_X by delegating to another agent the task α, or to do nothing relative to g_X. For the sake of simplicity, I will ignore the third choice; to do nothing at all relative to g_X.

The following notation is introduced [12, p. 102]:

- $U(X)_{p^+}$, the utility of X's successful performance of α and the following achievement of g_X,
- $U(X)_{p^-}$, the utility of X failing to perform α and the following achievement of g_X,
- $U(X)_{d^+}$, the utility of a successful delegation of the performance of α to achieve g_X to another agent Y,

- $U(X)_{d^-}$, the utility of failure in delegating α to an agent Y in order to achieve g_X

Following Expected Utility Theory (see, for example, [7]), in order to delegate, the expected utility of doing so must be greater than the expected utility of not delegating. The following inequality captures this; in order to delegate, it must hold that (where $0 < DoT_{XY\alpha} < 1$, $0 < DoT_{XX\alpha} < 1$ and $DoT_{XX\alpha}$ denotes the degree of X's trust in herself relative to α) [12, p. 103]:

$$DoT_{XY\alpha} * U(X)_{d^+} + (1 - DoT_{XY\alpha}) * U(X)_{d^-} > \\ DoT_{XX\alpha} * U(X)_{p^+} + (1 - DoT_{XX\alpha}) * U(X)_{p^-} \quad (2.2)$$

From (2.2), the threshold for delegating can be written as

$$DoT_{XY\alpha} > DoT_{XX\alpha} * A + B \quad (2.3)$$

where

$$A = \frac{U(X)_{p^+} - U(X)_{p^-}}{U(X)_{d^+} - U(X)_{d^-}} \quad (2.4)$$

and

$$B = \frac{U(X)_{p^-} - U(X)_{d^-}}{U(X)_{d^+} - U(X)_{d^-}} \quad (2.5)$$

The analysis presented here allows for the comparison of trust in different trustees. It also allows for a trust threshold to be calculated, which shows if a particular case of trust is sufficient for reliance and delegation.

2.4.1 Quantitative aspects and logical formalizations

In some logical formalizations of the concept of trust, trust is seen as binary: either X trusts Y or X does not trust Y (see for example [5, 25].) However, some authors incorporate a quantitative aspect in their logics. As seen above, degree of trust is a function of the credibility of the beliefs making up the positive evaluation of the trustee. This basic principle can be incorporated in a modal logic by introducing graded belief operators. Hübner and Demolombe [30] have extended a modal logic formalizing binary trust from containing only ungraded belief operators $\mathtt{Bel}_i$, to containing graded operators $\mathtt{Bel}_i^k$ (other logics formalising the concept of graded mental attitudes are presented in [13, 20]). In the original logic, $\mathtt{Bel}_i\varphi$ reads "agent i believes φ to be true", while $\mathtt{Bel}_i^k\varphi$ reads "agent i believes φ with strength k."

Chapter 3

Trust logics: Formalizing the C&F theory of trust

This chapter introduces modal logic as a way to formalize the C&F theory of trust. Modal logic offers a natural way to reason about mental states of agents, which makes it a useful tool for reasoning about cognitive aspects of trust. Three different trust logics—the logic of Herzig, Lorini, Hübner, and Vercouter [25], the logic of Demolombe and Lorini [14, 29], and the logic of Bonnefon, Longin, and Nguyen [5], all developed to capture aspects of the C&F theory, are reviewed.

Since the C&F theory is vast and complex, only certain key concepts are considered in the formalizations. The concepts considered are the basic notion of core trust, the concepts of distrust, mistrust, and lack of trust. The concept of trust disposition is also considered. The notions of reliance and decision to trust are not considered; the reason for this is primarily the complexity of the dependence belief, which could take a large number of forms, involved in reliance trust. I have also decided not to address the quantitative aspects. As mentioned in Chapter 2, the context argument in the trust relation will be omitted.

Also, the goal state is simplified, in that the formalizations will not consider the possible side effects of an action: I will treat the goal g as the only result of an action, instead of considering the world-state set P, of which the goal g is an element. That is, with the pseudo logical notation from Figure 2.2, I will consider

$$\text{After}_\alpha(g),$$

instead of

$$\text{After}_\alpha(P) \text{ with } g \in P.$$

The first section of this chapter contains a short introduction to modal logic, and is intended to provide the reader with the necessary tools to understand the later formalisms.

3.1 A short introduction to modal logic

Modal logic stems from the work of philosophers who needed a tool for reasoning about philosophical (and linguistic) concepts like belief and knowledge (doxastic and epistemic logics), time (temporal and tense logics), actions (dynamic logics), and moral concepts like permission and obligation (deontic logics). In research on intelligent agents and multi-agent systems (abbreviated MAS), one often assumes that intelligent agents have mental attitudes, like beliefs, desires, and intentions. In addition, one has to be able to reason about agents' mental attitudes and actions over time. Modal logics are particularly well fit to reason about these things. In particular, modal logic both enables reasoning about *properties* of intelligent agents and the environment in which they act. If these properties are expressed as logical formulas in some inference system, then modal logics can be part of intelligent agents' own reasoning capabilities [33, p. 761].

In this section, I will provide the reader with the necessary understanding of modal logic, especially in relation to multi-agent systems (MAS) (the reader unfamiliar with basic propositional logic can find a short survey in Appendix A).

3.1.1 Mono-modal logics

A mono-modal logic contains only one modality. Typically, mono-modal logics revolve around the operator $\Box$, which on a neutral reading means "it is necessary that...". Depending on the kind of concept one would like to formalize, the modal operator $\Box$ is subject to differences in intuitive meaning, axioms and inference rules, and governing semantic constraints.

The following definition gives the language of a basic propositional mono-modal logic [36, p. 3].

Definition 3.1.1. With a nonempty set of atomic propositions $ATM = \{p, q, r...\}$ and the connectives $\neg$, $\vee$, and $\Box$, the following syntax rules recursively give the language of the logic:

$$\varphi ::= p \,|\, \neg\varphi \,|\, \varphi \vee \varphi \,|\, \Box\varphi.$$

The above expression govern what kind of sentences can be considered as *well-formed formulas* (hereafter referred to simply as formulas): if φ is a formula, then φ is an atomic proposition, or a negated formula, or a disjunction of two formulas, or a formula under the modal operator $\Box$.

The intuitive, neutral meanings of the connectives $\neg$, $\vee$, and $\Box$ are:

- $\neg\varphi$: it is not the case that φ;
- $\varphi \vee \psi$: φ or ψ;
- $\Box\varphi$: φ is necessarily true.[1]

[1] As mentioned before, $\Box$ can have many other meanings.

The following abbreviations are introduced:

$$\begin{aligned}
\top &\overset{\text{def}}{=} p \lor \neg p; \\
\bot &\overset{\text{def}}{=} \neg\top; \\
\varphi \land \psi &\overset{\text{def}}{=} \neg(\neg\varphi \lor \neg\psi); \\
\varphi \to \psi &\overset{\text{def}}{=} \neg\varphi \lor \psi; \\
\varphi \leftrightarrow \psi &\overset{\text{def}}{=} (\varphi \to \psi) \land (\psi \to \varphi); \\
\Diamond\varphi &\overset{\text{def}}{=} \neg\Box\neg\varphi;
\end{aligned}$$

with the intuitive meanings:

- $\top$: true;
- $\bot$: false;
- $\varphi \land \psi$: φ and ψ;
- $\varphi \to \psi$: if φ, then ψ;
- $\varphi \leftrightarrow \psi$: φ, if and only if ψ;
- $\Diamond\varphi$: it is possible that φ is true.[2]

With these definitions in place, several logics can be constructed. For example, if the $\Box$ operator is interpreted as meaning “it is known that ...”, the following axiom should intuitively hold:

(Tax) $\Box\varphi \to \varphi$

meaning that what is known is true.

A variety of different systems can be obtained by combining different axioms. However, there are several axioms that apply to a wide range of different modal systems. First, *modus ponens*,

(MP) from $\vdash \varphi$ and $\vdash \varphi \to \psi$, infer $\vdash \psi$[3]

is a rule of derivation in every modal logic. Second, many modal systems share the axiom

(K) $\Box\varphi \land \Box(\varphi \to \psi) \to \Box\psi$

and the *necessitation rule*

(Nec) from $\vdash \varphi$, infer $\vdash \Box\varphi$.

The necessitation rule is a derivation rule; if φ is a theorem of the system in question, then one can infer $\Box\varphi$. This is intuitively right; every tautology is necessarily true.

In the following sections, I will talk about *normal* modal systems. The definition of a normal modal logic runs as follows [36, p. 32]:

[2] Note that the meaning of $\Diamond$ is related to the meaning of $\Box$.

[3] $\vdash \varphi$ expresses that φ is a theorem. When expressing that a formula φ is a theorem of a specific logic (which most often is the case), the expression $\vdash_S \varphi$, where S is the name of the logic in question, is used.

Definition 3.1.2. A modal logic S is *normal* if

- each instance of **K** is in S;
- S is closed under **Nec**.

The following is a useful theorem that holds in any normal modal logic. The theorem says that $\Box$-type operators can be distributed over conjunction, and that $\Diamond$-type operators can be distributed over disjunction. The expression $\vdash_S \varphi$ is used to denote that a formula φ is a theorem of the modal system S.

Theorem 3.1.1. *If S is a normal modal logic, then:*

(a) $\vdash_S \Box(\varphi \wedge \psi) \leftrightarrow \Box\varphi \wedge \Box\psi$;

(b) $\vdash_S \Diamond(\varphi \vee \psi) \leftrightarrow \Diamond\varphi \vee \Diamond\psi$;

(c) $\vdash_S \Box\varphi \vee \Box\psi \rightarrow \Box(\varphi \vee \psi)$;

(d) $\vdash_S \Diamond(\varphi \wedge \psi) \rightarrow \Diamond\varphi \wedge \Diamond\psi$.

This theorem is proven in e.g. [36, pp. 7–8, 31]. The converses of Theorems 3.1.1(c) and 3.1.1(d), i.e. $\Box(\varphi \vee \psi) \rightarrow \Box\varphi \vee \Box\psi$ and $\Diamond\varphi \wedge \Diamond\psi \rightarrow \Diamond(\varphi \wedge \psi)$, are not valid.

Note that the basic monomodal logic would have been essentially the same if the $\Diamond$-operator had been treated as basic, in the sense that it had been given in the syntax rules instead of the $\Box$-operator.[4] The $\Box$-operator could then have been defined as $\Box\varphi \overset{\text{def}}{=} \neg\Diamond\neg\varphi$. An operator which is defined in terms of another operator in the same way in which the $\Diamond$-operator is defined (i.e. $\Diamond\varphi \overset{\text{def}}{=} \neg\Box\neg\varphi$) is called a *dual* to the operator occurring on the right hand side of the definition. The following theorem states a useful property of dual modal operators in any normal modal logic.

Theorem 3.1.2. *If S is a normal modal logic, then:*

(a) $\vdash_S \neg\Box\varphi \leftrightarrow \Diamond\neg\varphi$;

(b) $\vdash_S \neg\Diamond\varphi \leftrightarrow \Box\neg\varphi$.

A proof of this theorem can be found in [36, pp. 9–10].

Substitution of logically equivalent formulas is valid in any normal modal logic.

Theorem 3.1.3. *For any normal modal logic S, if φ' is exactly like φ except that an occurrence of ψ in φ is replaced by ψ', then if $\vdash_S \psi \leftrightarrow \psi'$, then $\vdash_S \varphi \leftrightarrow \varphi'$.*

The proof can be found in [36, pp. 8–9].

[4] In fact, both the $\Box$-operator and the $\Diamond$-operator could have been given in the syntax rules, again resulting in essentially the same logic.

Semantics

All logics that will be presented in this thesis have possible world semantics. The idea behind the concept of possible worlds stems from the works of logician (and philosopher) Saul Kripke—therefore, possible world semantics is also known as Kripke semantics. Previously, I gave the $\Box$-operator and the $\Diamond$-operator the intuitive meanings "It is necessary that ..." and "it is possible that ...", respectively. Philosophically, necessity is often interpreted as *truth in all possible worlds*, where a possible world is a world differing in some respects from the actual world. For example, one can maintain, it is possible that Barack Obama could have been a logician, instead of a politician, even though Barack Obama *actually* is a politician. Using the concept of a possible world, this possibility is expressed as "There is a possible world where Barack Obama is a logician". Possible world semantics is a way to implement the idea of necessity as truth in all possible world in a formal semantics.

As seen, modal logics can take various forms depending on which axioms is assumed to govern the modal operators. Even when the $\Box$-operator is interpreted as meaning "It is necessary that ...", different axioms can describe different aspects of necessity. Therefore, the idea of truth in all possible worlds must be relativised in some way. In possible world semantics, this is done by restating the claim of truth in all possible worlds to truth in all *accessible* worlds. The idea is that some worlds is accessible from the world where a formula is evaluated, while other worlds might be inaccessible.

Formally, possible world semantics revolve around a non-empty set of possible worlds W (the elements in W are also called *points*, *states*, *situations*, *times* etc., depending on the interpretation of the logic) on which a binary *accessibility* relation R is defined. W and R make up a *Kripke frame* $\langle W, R\rangle$.

Definition 3.1.3. A (Kripke) *model* is a triple $M = \langle W, R, V\rangle$, where $\langle W, R\rangle$ is a Kripke frame and V is a function $V : ATM \to 2^W$ assigning to each atomic proposition $p \in ATM$ a subset $V(p)$ of W, consisting of worlds where p is true.

Now, it can be defined what it means for a formula φ to be *satisfied* in a model M at a world $w \in W$, abbreviated $M, w \vDash \varphi$.

Definition 3.1.4. For any world $w \in W$ in the model M, the following holds (here, the abbreviation "iff" is used to express "if and only if"):

- $M, w \vDash \top$;
- $M, w \nvDash \bot$ ($\nvDash$ is an abbreviation for "not $\vDash$");
- $M, w \vDash p$ iff $w \in V(p)$;
- $M, w \vDash \neg\varphi$ iff $M, w \nvDash \varphi$;
- $M, w \vDash \varphi \vee \psi$ iff $M, w \vDash \varphi$ or $M, w \vDash \psi$;
- $M, w \vDash \varphi \wedge \psi$ iff $M, w \vDash \varphi$ and $M, w \vDash \psi$;
- $M, w \vDash \varphi \to \psi$ iff $M, w \nvDash \varphi$ or $M, w \vDash \psi$;
- $M, w \vDash \varphi \leftrightarrow \psi$ iff ($M, w \vDash \varphi$ iff $M, w \vDash \psi$);
- $M, w \vDash \Box\varphi$ iff $M, v \vDash \varphi$ for every v such that $(w, v) \in R$;

- $M, w \vDash \Diamond\varphi$ iff there is a v such that $M, v \vDash \varphi$ and $(w, v) \in R$.

If a formula φ is satisfied at every world w in a model M, φ is said to be *globally satisfied* in the model M (abbreviated $M \vDash \varphi$), and a formula φ is said to be *valid* (abbreviated $\vDash \varphi$) if it is globally satisfied in every model M. Finally, φ is said to be *satisfiable* if there exists a model M and a world w in M such that $M, w \vDash \varphi$ [3, p. 4].

Note that a formula preceded by the $\Box$-operator, i.e. a formula $\Box\varphi$, is true when the formula φ is true in *every* accessible world. Because of this condition, modal operators like the $\Box$-operator, i.e., operators whose semantics "quantify" over all accessible worlds, are called *universal* modal operators. A modal operator with semantics like the $\Diamond$-operator, are called an *existential* modal operator. This is because if a formula $\Diamond\varphi$ is true, then there *exists* an accessible world where φ is true.[5]

In order to "match" the semantics to the relevant axiomatics, different *semantic constraints* are placed on the accessibility relation R. It can be proven that Axioms **K** and **Nec** are valid without any special restrictions on the accessibility relation R (see [36, p. 46] for a proof). For other Axioms to be valid, further constraints on R might be needed. For example, for the Axiom **Tax** to be valid, it has to be the case that, for every model M and every world w in M, if $\Box\varphi$ is satisfied at w, φ must be satisfied at w. The constraint on R in this case is *reflexivity*:

Definition 3.1.5. The relation R on W is *reflexive* if and only if for all $w \in W$, $(w, w) \in R$.

If R is reflexive, then every world w is accessible from itself; by Definition 3.1.4, $\Box\varphi$ is true in a world w if φ is true at every world accessible from w. Since w is accessible from itself, $\Box\varphi$ cannot be true at w unless φ is true at w as well.[6]

The modal system *KD45*

In this section, the modal system *KD45*, which is a standard logic of belief, will be presented and discussed. The syntactic primitive of the logic is a nonempty set of atomic propositions $ATM = \{p, q, ...\}$. The following syntax rules recursively give the language of the logic *KD45*:

$$\varphi ::= p \,|\, \neg\varphi \,|\, \varphi \vee \varphi \,|\, \mathtt{B}\varphi.$$

The above expression means that a formula φ is one of the following: an atomic proposition, or a negated formula, or two formulas connected by the connective $\vee$, or a formula under the operator $\mathtt{B}$. The operator $\mathtt{B}$ is a $\Box$ type modal operator, with the intuitive meaning "it is believed that...". The usual connectives and the constants $\top$ and $\bot$ are defined in the same way as above.

The logic *KD45* is governed by a set of axioms and corresponding semantic constraints. First, let me state the axioms [33, p: 995]:

[5] The reader familiar with first-order predicate logic can note a parallel to universal and existential quantifiers ($\forall$ and $\exists$).

[6] Note that this is *not* a proof of the correspondence between Axiom **Tax** and the reflexivity of the accessibility relation R. In most cases, the correspondence is not straightforward, and complicated proofs are used to formally prove that there is a correspondence between syntactical axioms and semantic constraints.

(PC) all theorems of propositional logic;

(K) $(\mathtt{B}\varphi \wedge \mathtt{B}(\varphi \rightarrow \psi)) \rightarrow \mathtt{B}\psi$;

(D) $\neg(\mathtt{B}\varphi \wedge \mathtt{B}\neg\varphi)$;

(4) $\mathtt{B}\varphi \rightarrow \mathtt{BB}\varphi$;

(5) $\neg\mathtt{B}\varphi \rightarrow \mathtt{B}\neg\mathtt{B}\varphi$;

(MP) from $\vdash_{KD45} \varphi$ and $\vdash_{KD45} \varphi \rightarrow \psi$, infer $\vdash_{KD45} \psi$;

(Nec) from $\vdash_{KD45} \varphi$, infer $\vdash_{KD45} \mathtt{B}\varphi$.

Axiom **K** allows for deduction, and captures the intuitive principle that if it is believed that if φ then ψ, then if φ is believed, ψ is believed as well. Axiom **D** guarantees that beliefs cannot be inconsistent; a rational reasoner cannot both believe φ and $\neg\varphi$. Axioms **4** and **5** are principles of introspection; what is believed is believed to be believed, and what is not believed is believed to not be believed. Axioms **MP** and **Nec** are the inference rules of the logic. The Axiom **MP** (*modus ponens*) is straightforward. Axiom **Nec** says that if φ is a theorem, then φ is believed; in other words, all tautologies are believed to be true.

Thus, *KD45* is a normal modal logic. If the Axioms **4** and **5** are left out, the resulting system is called *KD*.

The semantics of *KD45* is a possible world semantics as defined in Defintions 3.1.3 and 3.1.4. The above axioms result in semantic constraints that have to be placed on the accessibility relation R; for the axioms to become valid, R must be serial, transitive and euclidean [33, p. 994]. For Axiom **4** to become valid, R must be *transitive*:

Definition 3.1.6. The relation R on W is *transitive* if, for $w, v, u \in W$, $(w, v) \in R$ and $(v, u) \in R$, then $(w, u) \in R$.

For Axiom **5** to become valid, R must be *euclidean*:

Definition 3.1.7. The relation R on W is *euclidean* if, for every $w, v, u \in W$, $(w, v) \in R$ and $(w, u) \in R$, then $(v, u) \in R$.

For Axiom **D** to become valid, R must be *serial*:

Definition 3.1.8. The relation R on W is *serial* if, for every $w \in W$, there is a $v \in W$ such that $(w, v) \in R$.

3.1.2 Multi-modal logics for multi-agent systems

Mono-modal logics can be combined to form *multi-modal* systems (for an advanced treatment of combining modal logic, see [28]). Often, intelligent agents are thought to have *mental* attitudes; their behaviour is modelled after their beliefs, goals, intentions, etc., and how those mental attitudes change over time [33, p. 992]. As seen in the previous section, the modal logic *KD45* enables reasoning about an agent's beliefs. There are also a wide range of modal logics dealing with, for instance, knowledge, goals, obligations, etc. Meyer and Veltman argue that, when formalizing the mental attitudes underlying the behaviour of intelligent agents, mono-modal logics are not that interesting *per se*; it is the *dynamics*—how different logics interact—over time that are interesting [33, p. 992]. Modal logics for intelligent agents can also be extended to several agents, resulting in logics suitable for multi-agent systems.

Multi-agent systems (MAS) studies intelligent agents acting in relation to each other. A logic for MAS thus needs to allow reasoning about mental attitudes and actions of several agents in parallel. As mentioned, things get interesting when several modal logics are combined in different ways. In order to simplify the presentation, I have chosen to focus the following discussion on a *KD45* type modal logic, extended to a set of agents.

Extension of a modal logic to a set of agents is most often accomplished by introducing indexed modal operators in the language of the logic [26, pp. 764–765]. The syntactic primitives for such a *KD45* type modal logic for several agents consist not only of a nonempty set $ATM = \{p, g, ...\}$ of atomic propositions, but also of a nonempty, finite set of agents, $AGT = \{i_1, i_2, ..., i_n\}$. I will use variables $i, j, k, ...$ to denote agents. The language of the logic is the same as for *KD45*, but instead of one single operator $\mathtt{B}$, there are operators $\mathtt{B}_i$, where $i \in AGT$. $\mathtt{B}_i\varphi$ intuitively means "agent i believes that φ is true". Thus, the logic includes one belief operator for each agent in AGT. All operators $\mathtt{B}_i$ are governed by the axiom schema for *KD45* (see Section 3.1.1).

The semantics of the logic is a possible world semantics, closely resembling the usual semantics for *KD45*. A model $\langle W, R, V \rangle$ is a couple, with W being a set of possible worlds and V a valuation function, as usual, and R is a function such that $R : AGT \to W \times W$ maps every agent i to a binary serial transitive euclidean relation R_i between possible worlds in W. Thus, R can be seen as a collection of relations R_i, one relation for each agent in AGT. One could also say that a model F is a tuple $\langle W, R_1, R_2, ..., R_n, V \rangle$, where n is the number of agents in AGT. The set of worlds w' such that $(w, w') \in R_i$ are the set of all worlds compatible with agent i's beliefs at world w.

Truth of $\mathtt{B}_i\varphi$ in a model M at a world w (abbreviated $M, w \vDash \mathtt{B}_i\varphi$) resembles that of *KD45*; $M, w \vDash \mathtt{B}_i\varphi$ if and only if $M, w' \vDash \varphi$ for all w' such that $(w, w') \in R_i$.

I will now proceed to explain how one can combine a logic of belief like the one discussed above, with a logic of choices. The presentation here is loosely based on that of Demolombe and Lorini [14], and should not be seen as a complete logic; rather it is an example of how axioms and semantic constraints could be defined to capture interactions between beliefs and choices.

The basic operator of the logic of choice is $\mathtt{Choice}_i$, with the intuitive meaning "agent i has the chosen goal ..." or "agent i wants ... to be true". The choice logic is a *KD* type logic, which means that it is governed by the following two axioms:

($\mathbf{K_{Choice}}$) $(\mathtt{Choice}_i\varphi \wedge \mathtt{Choice}_i(\varphi \to \psi)) \to \mathtt{Choice}_i\psi$;

($\mathbf{D_{Choice}}$) $\neg(\mathtt{Choice}_i\varphi \wedge \mathtt{Choice}_i\neg\varphi)$

and closed under **MP** and

($\mathbf{Nec_{Choice}}$) from φ, infer $\mathtt{Choice}_i\varphi$.

Axiom $\mathbf{D_{Choice}}$ says that an agent's choices cannot be inconsistent; an agent cannot both have the chosen goal that φ and the chosen goal that $\neg\varphi$.

Now, beliefs and choices interact in certain ways. Typical principles governing the interactions of choices and beliefs are the following principles of introspection with respect to choices:

($\mathbf{4_{Choice,\ Believe}}$) $\texttt{Choice}_i\varphi \rightarrow \texttt{B}_i\texttt{Choice}_i\varphi$;

($\mathbf{5_{Choice,\ Believe}}$) $\neg\texttt{Choice}_i\varphi \rightarrow \texttt{B}_i\neg\texttt{Choice}_i\varphi$.

These two axioms say that what is chosen by an agent i is believed by i to be chosen, and what is not chosen is believed not to be chosen.

A model in the combined logic is a tuple $M = \langle W, R, C, V\rangle$, where W is a set of possible worlds, V is a valuation function, and R is a collection of binary serial transitive euclidean relations R_i on W for each $i \in AGT$, just like above, while C is a collection of binary serial relations C_i on W for each $i \in AGT$. The set of worlds w' such that $(w, w') \in C_i$ are the set of worlds which are compatible with agent i's choices at world w. All relations C_i need to be serial for Axiom $\mathbf{D_{Choice}}$ to become valid.

For Axioms $\mathbf{4_{Choice,\ Belief}}$ and $\mathbf{5_{Choice,\ Belief}}$ to become valid, the following semantic constraints are defined [14]. For every $i \in AGT$ and $w \in W$:

S1 if $(w, w') \in R_i$, then, for all v, if $(w, v) \in C_i$ then $(w', v) \in C_i$;

S2 if $(w, w') \in R_i$, then, for all v, if $(w', v) \in C_i$ then $(w, v) \in C_i$.

3.2 Herzig, Lorini, Hübner, and Vercouter's logic

In this section, the logic $\mathcal{HHVL}$ developed by Herzig, Lorini, Hübner, and Vercouter [25] will be presented. In the same paper, they also extend the logic $\mathcal{HHVL}$ in order to enable reasoning about reputation. I will only focus on the aspects related to the C&F theory as presented in Chapter 2.

3.2.1 Syntax

The syntactic primitives of the logic are: a nonempty finite set of agents $AGT = \{i_1, i_2, ..., i_n\}$; a nonempty finite set of actions $ACT = \{e_1, e_2, ..., e_k\}$; and a nonempty set of atomic propositions $ATM = \{p, q, ...\}$. Variables $i, j, ...$ denote agents, and variables $\alpha, \beta, ...$ denote actions.

The language of the logic is given by the following syntax rules:

$$\varphi ::= p \,|\, \neg\varphi \,|\, \varphi \vee \varphi \,|\, \texttt{G}\varphi \,|\, \texttt{After}_{i:\alpha}\varphi \,|\, \texttt{Does}_{i:\alpha}\varphi \,|\, \texttt{Bel}_i\varphi \,|\, \texttt{Choice}_i\varphi$$

where $p \in ATM$, $i \in AGT$, and $\alpha \in ACT$. The usual connectives $(\wedge, \rightarrow, \leftrightarrow)$ and *true* and *false* ($\top$ and $\bot$) are defined as in Section 3.1.1.

The intuitive meanings of the operators are as follows:

- $\texttt{G}\varphi$: φ will always be true;
- $\texttt{After}_{i:\alpha}\varphi$: immediately after agent i does α, φ will be true;[7]

[7] Note that the logic models time as a discreet sequence of time points. Thus, that φ is true immediately after some performance of an action by an agent means that φ is true at the next time point (see [2] for a further discussion of the discreetness of time).

- $\texttt{Does}_{i:\alpha}\varphi$: agent i is going to do α and φ will be true afterwards;
- $\texttt{Bel}_i\varphi$: agent i believes φ;
- $\texttt{Choice}_i\varphi$: agent i has the chosen goal φ.

The following abbreviations are introduced:

$$\begin{aligned}
\texttt{G}^*\varphi &\overset{\text{def}}{=} \varphi \wedge \texttt{G}\varphi;\\
\texttt{Capable}_i(\alpha) &\overset{\text{def}}{=} \neg\texttt{After}_{i:\alpha}\bot;\\
\texttt{Intends}_i(\alpha) &\overset{\text{def}}{=} \texttt{Choice}_i\texttt{Does}_{i:\alpha}\top;\\
\texttt{F}\varphi &\overset{\text{def}}{=} \neg\texttt{G}\neg\varphi;\\
\texttt{F}^*\varphi &\overset{\text{def}}{=} \neg\texttt{G}^*\neg\varphi;\\
\texttt{Poss}_i\varphi &\overset{\text{def}}{=} \neg\texttt{Bel}_i\neg\varphi.
\end{aligned}$$

The intended meanings of the abbreviations are as follows:

- $\texttt{G}^*\varphi$: φ is true now and will always be true;
- $\texttt{Capable}_i(\alpha)$: i is capable of doing α;
- $\texttt{Intends}_i(\alpha)$: i intends to do α;
- $\texttt{F}\varphi$: φ will eventually be true;
- $\texttt{F}^*\varphi$: φ is true now or will be true sometimes in the future;
- $\texttt{Poss}_i\varphi$: according to i, φ is possible.

Axiomatics

The following are the axioms of the logic $\mathcal{HHVL}$ [25]:

(PC) all theorems of propositional logic;

($\mathbf{K_{Bel}}$) $(\texttt{Bel}_i\varphi \wedge \texttt{Bel}_i(\varphi \to \psi)) \to \texttt{Bel}_i\psi$;

($\mathbf{D_{Bel}}$) $\neg(\texttt{Bel}_i\varphi \wedge \texttt{Bel}_i\neg\varphi)$;

($\mathbf{4_{Bel}}$) $\texttt{Bel}_i\varphi \to \texttt{Bel}_i\texttt{Bel}_i\varphi$;

($\mathbf{5_{Bel}}$) $\neg\texttt{Bel}_i\varphi \to \texttt{Bel}_i\neg\texttt{Bel}_i\varphi$;

($\mathbf{K_{Choice}}$) $(\texttt{Choice}_i\varphi \wedge \texttt{Choice}_i(\varphi \to \psi)) \to \texttt{Choice}_i\psi$;

($\mathbf{D_{Choice}}$) $\neg(\texttt{Choice}_i\varphi \wedge \texttt{Choice}_i\neg\varphi)$;

($\mathbf{4_{Choice}}$) $\texttt{Choice}_i\varphi \to \texttt{Bel}_i\texttt{Choice}_i\varphi$;

($\mathbf{5_{Choice}}$) $\neg\texttt{Choice}_i\varphi \to \texttt{Bel}_i\neg\texttt{Choice}_i\varphi$;

($\mathbf{K_{After}}$) $(\texttt{After}_{i:\alpha}\varphi \wedge \texttt{After}_{i:\alpha}(\varphi \to \psi)); \to \texttt{After}_{i:\alpha}\psi$;

($\mathbf{K_{Does}}$) $(\texttt{Does}_{i:\alpha}\varphi \wedge \neg\texttt{Does}_{i:\alpha}\neg\psi) \to \texttt{Does}_{i:\alpha}(\varphi \wedge \psi)$;

$(\mathbf{Alt}_{\mathbf{Does}})$ $\mathtt{Does}_{i:\alpha}\varphi \to \neg\mathtt{Does}_{j:\beta}\neg\varphi;$

$(\mathbf{K}_{\mathbf{G}})$ $(\mathtt{G}\varphi \wedge \mathtt{G}(\varphi \to \psi)) \to \mathtt{G}\psi;$

$(\mathbf{4}_{\mathbf{G}})$ $\mathtt{G}\varphi \to \mathtt{GG}\varphi;$

$(\mathbf{H}_{\mathbf{G}})$ $(\mathtt{F}\varphi \wedge \mathtt{F}\psi) \to (\mathtt{F}(\varphi \wedge \mathtt{F}\psi) \vee \mathtt{F}(\psi \wedge \mathtt{F}\varphi) \vee \mathtt{F}(\varphi \wedge \psi));$

$(\mathbf{Inc}_{\mathbf{After,\ Does}})$ $\mathtt{Does}_{i:\alpha}\varphi \to \neg\mathtt{After}_{i:\alpha}\neg\varphi;$

$(\mathbf{IntAct1})$ $(\mathtt{Choice}_i\mathtt{Does}_{i:\alpha}\top \wedge \neg\mathtt{After}_{i:\alpha}\bot) \to \mathtt{Does}_{i:\alpha}\top;$

$(\mathbf{IntAct2})$ $\mathtt{Does}_{i:\alpha}\top \to \mathtt{Choice}_i\mathtt{Does}_{i:\alpha}\top;$

$(\mathbf{WR})$ $\mathtt{Bel}_i\varphi \to \neg\mathtt{Choice}_i\neg\varphi;$

$(\mathbf{Inc}_{\mathbf{G,\ Does}})$ $\mathtt{Does}_{i:\alpha}\varphi \to \mathtt{F}\varphi;$

$(\mathbf{OneStepAct})$ $\mathtt{Does}_{i:\alpha}\mathtt{G}^*\varphi \to \mathtt{G}\varphi;$

$(\mathbf{MP})$ from $\vdash_{\mathcal{HHVL}} \varphi$ and $\vdash_{\mathcal{HHVL}} \varphi \to \psi$, infer $\vdash_{\mathcal{HHVL}} \psi$;[8]

$(\mathbf{Nec}_{\mathbf{Bel}})$ from $\vdash_{\mathcal{HHVL}} \varphi$, infer $\vdash_{\mathcal{HHVL}} \mathtt{Bel}_i\varphi$;

$(\mathbf{Nec}_{\mathbf{Choice}})$ from $\vdash_{\mathcal{HHVL}} \varphi$, infer $\vdash_{\mathcal{HHVL}} \mathtt{Choice}_i\varphi$;

$(\mathbf{Nec}_{\mathbf{G}})$ from $\vdash_{\mathcal{HHVL}} \varphi$, infer $\vdash_{\mathcal{HHVL}} \mathtt{G}\varphi$;

$(\mathbf{Nec}_{\mathbf{After}})$ from $\vdash_{\mathcal{HHVL}} \varphi$, infer $\vdash_{\mathcal{HHVL}} \mathtt{After}_{i:\alpha}\varphi$;

$(\mathbf{Nec}_{\mathbf{Does}})$ from $\vdash_{\mathcal{HHVL}} \varphi$, infer $\vdash_{\mathcal{HHVL}} \neg\mathtt{Does}_{i:\alpha}\neg\varphi$.[9]

[8] $\vdash_{\mathcal{HHVL}} \varphi$ denotes that φ is a theorem of $\mathcal{HHVL}$.

[9] Note that $\mathtt{Does}_{i:\alpha}$ is an *existential* modal operator. That is why the Axioms $\mathbf{K}_{\mathbf{Does}}$ and $\mathbf{Nec}_{\mathbf{Does}}$ look different than usual; however, these axioms mean that $\mathtt{Does}_{i:\alpha}$ is a K type logic, with an existential modal operator as basic instead of its universal dual.

To see this, assume that $\Diamond$ is a normal existential modal operator, with its dual defined as usual as $\Box\varphi \overset{\mathrm{def}}{=} \neg\Diamond\neg\varphi$. Assume that the schema $\Diamond\varphi \wedge \neg\Diamond\neg\psi \to \Diamond(\varphi \wedge \psi)$ is an axiom. Then:

$$\begin{aligned}
\Diamond\neg\varphi \wedge \neg\Diamond\neg\psi \to \Diamond(\neg\varphi \wedge \psi) &\equiv \Diamond\neg\varphi \wedge \Box\psi \to \Diamond(\neg\varphi \wedge \psi)\\
&\equiv \neg\Diamond(\neg\varphi \wedge \psi) \to \neg(\Diamond\neg\varphi \wedge \Box\psi)\\
&\equiv \neg\Diamond\neg(\varphi \vee \neg\psi) \to \neg\Diamond\neg\varphi \vee \neg\Box\psi\\
&\equiv \Box(\psi \to \varphi) \to (\Box\psi \to \Box\varphi)\\
&\equiv \Box\varphi \wedge \Box(\varphi \to \psi) \to \Box\psi.
\end{aligned}$$

Note also that the schema $\Diamond\varphi \wedge \neg\Diamond\neg\psi \to \Diamond(\varphi \wedge \psi)$ is a theorem in normal modal logics:

$$\begin{aligned}
\Box\varphi \wedge \Box(\varphi \to \neg\psi) \to \Box\neg\psi &\equiv \Box(\varphi \to \neg\psi) \to (\Box\varphi \to \Box\neg\psi)\\
&\equiv \neg\Diamond\neg(\neg\varphi \vee \neg\psi) \to (\neg\Box\varphi \vee \Box\neg\psi)\\
&\equiv \neg\Diamond(\varphi \wedge \psi) \to (\neg\Box\varphi \vee \neg\Diamond\psi)\\
&\equiv \neg\Diamond(\varphi \wedge \psi) \to \neg(\Box\varphi \wedge \Diamond\psi)\\
&\equiv \Diamond\psi \wedge \Box\varphi \to \Diamond(\psi \wedge \varphi)\\
&\equiv \Diamond\psi \wedge \neg\Diamond\neg\varphi \to \Diamond(\psi \wedge \varphi).
\end{aligned}$$

3.2.2 Semantics

A frame F in the logic is a tuple $\langle W, A, B, C, D, G \rangle$, where A, B, C, D, G are defined below. For the sake of convenience, all relations on W are seen as functions from W to 2^W. Thus, for every relation in the collections of relations A, B, C, D, and G, the expression $A_{i:\alpha}(w)$ denotes the set $\{w' : (w, w') \in A_{i:\alpha}\}$, etc.

- W is a nonempty set of possible worlds.
- A is a collection of binary relations $A_{i:\alpha}$ on W for every agent $i \in AGT$ and action $\alpha \in ACT$. $A_{i:\alpha}(w)$ is the set of worlds w' that can be reached from w through i's performance of α.
- B is a collection of binary serial transitive euclidean relations B_i on W for every $i \in AGT$. $B_i(w)$ is the set of worlds w' that are compatible with i's beliefs at w.
- C is a collection of binary, serial relations C_i on W for every $i \in AGT$. $C_i(w)$ is the set of worlds w' that are compatible with i's choices at w.
- D is a collection of binary deterministic[10] relations $D_{i:\alpha}$ on W for every $i \in AGT$ and $\alpha \in ACT$. If $(w, w') \in D_{i:\alpha}$, then w' is the unique *actual successor* world of w, that is actually reached from w by the performance of α by i.
- G is a transitive connected[11] relation on W. $G(w)$ is the set of worlds w' which are future to w.

If $D_{i:\alpha}(w) \neq \emptyset$, then $D_{i:\alpha}$ is *defined* at w. Similarly for every $A_{i:\alpha}$: if $A_{i:\alpha} \neq \emptyset$, then $A_{i:\alpha}$ is *defined* at w. When $D_{i:\alpha}(w) = \{w'\}$, then i performs α at w, resulting in the next state w'. When $w' \in A_{i:\alpha}(w)$, then, if i performs α at w, this might result in w'. Hence, if $w' \in A_{i:\alpha}(w)$, but $D_{i:\alpha}(w) = \emptyset$, then i does not perform α at w, but if i had performed α at w, this might have resulted in the outcome w'.

The following are the semantic constraints on frames in the logic. For every $i, j \in AGT$, $\alpha, \beta \in ACT$, and $w \in W$:

S1 if $D_{i:\alpha}$ and $D_{j:\beta}$ are defined at w, then $D_{i:\alpha}(w) = D_{j:\beta}(w)$;

S2 $D_{i:\alpha} \subseteq A_{i:\alpha}$;

S3 if $A_{i:\alpha}$ is defined at w and $D_{i:\alpha}$ is defined at w' for all $w' \in C_i(w)$, then $D_{i:\alpha}$ is defined at w;

S4 if $w' \in C_i(w)$ and $D_{i:\alpha}$ is defined at w, then $D_{i:\alpha}$ is defined at w';

S5 $D_{i:\alpha} \subseteq G$;

S6 if $w' \in D_{i:\alpha}(w)$ and $v \in G(w)$, then $w' = v$ or $v \in G(w')$;

[10] A relation $D_{i:\alpha}$ is *deterministic* iff for every $w \in W$, if $(w, w') \in D_{i:\alpha}$ and $(w, w'') \in D_{i:\alpha}$, then $w' = w''$ [25, p. 219].

[11] The relation G is *connected* iff, for every $w \in W$: if $(w, w') \in G$ and $(w, w'') \in G$, then $(w', w'') \in G$ or $(w'', w') \in G$ or $w' = w''$ [25, p. 219].

S7 $C_i(w) \cap B_i(w) \neq \emptyset$;

S8 if $w' \in B_i(w)$, then $C_i(w) = C_i(w')$.

Constraint **S1** says that every world can only have one successor world; i.e. there can only be one next world for every world w. Constraint **S2** says that if a world w' is the successor world to w, then w' must be *reachable* from w. Constraints **S3** and **S4** guarantee that there is a relationship between agents' choices and their actions. Constraint **S5** guarantees that every world w' resulting from an action α performed by i at w is in the future of w. Constraint **S6** captures that there is no *third* world between a world w and the outcome w' of an action performed at w. Thus, $D_{i:\alpha}(w) = \{w'\}$ can be considered to be the next state of w. Constraint **S7** captures the principle that an agent i's choices must be compatible with her beliefs. Finally, Constraint **S8** captures the principle of introspection with respect to choices: agents are aware of their choices.

Models M of the logic are couples $\langle F, V \rangle$, where F is a frame, and V is a function which associates each world $w \in W$ with a set $V(w)$ of atomic propositions that are true at world $w \in W$.

For every model M, world $w \in W$ and formula φ, the expression $M, w \vDash \varphi$ means that φ is true at world w in model M. The truth conditions for atomic propositions and the usual connectives are defined as in Section 3.1.1. The remaining truth conditions for the logic are the following:

- $M, w \vDash \mathtt{G}\varphi$ iff $M, w' \vDash \varphi$ for all w' such that $(w, w') \in G$.
- $M, w \vDash \mathtt{After}_{i:\alpha}\varphi$ iff $M, w' \vDash \varphi$ for all w' such that $(w, w') \in A_{i:\alpha}$.
- $M, w \vDash \mathtt{Does}_{i:\alpha}\varphi$ iff there is a $w' \in D_{i:\alpha}(w)$ such that $M, w' \vDash \varphi$.
- $M, w \vDash \mathtt{Bel}_i\varphi$ iff $M, w' \vDash \varphi$ for all w' such that $(w, w') \in B_i$.
- $M, w \vDash \mathtt{Choice}_i\varphi$ iff $M, w' \vDash \varphi$ for all w' such that $(w, w') \in C_i$.

A formula φ is said to be *$\mathcal{HHVL}$-satisfiable* if there exists a model M and a world w in $\mathcal{HHVL}$ such that $M, w \vDash \varphi$.

3.2.3 Formalizing the C&F theory

Herzig *et al.* [25] make a distinction between *occurrent trust* and *dispositional trust*, somewhat corresponding to the distinction between core trust and trust disposition in the C&F theory (see Chapter 2, Section 2.3.1). Occurrent trust corresponds to the concept of core trust, and is formally defined in $\mathcal{HHVL}$ as:

Definition 3.2.1.

$$\begin{aligned}\mathtt{OccTrust}(i, j, \alpha, \varphi) \overset{\text{def}}{=} &\mathtt{Choice}_i\mathtt{F}\varphi \\ &\wedge \mathtt{Bel}_i(\mathtt{Intends}_j(\alpha) \wedge \mathtt{Capable}_j(\alpha) \wedge \mathtt{After}_{j:\alpha}\varphi).\end{aligned}$$

Occurrent trust is the "here-and-now" trust in a trustee in relation to an active and pursued goal of the truster. However, as pointed out by C&F, it is possible to evaluate a trustee in relation to a *potential goal* (see Chapter 2, Section 2.3.1). It might be useful to recall an example of dispositional trust. I might trust a local bookstore with providing me a particular book in the future, if it ever were to become my goal of owning that particular book. That is, at present time, it is not a goal of mine of owning a particular book, but given certain circumstances (I might for example have a vague idea about possibly writing an essay on a topic in the future, and if I want to write that essay, I believe that I need a particular book covering the relevant topic) the potential goal of owning the book becomes an active and pursued goal of mine. In such a case, a concept of dispositional trust is needed; I can trust that if I want to write the essay, the local bookstore will provide me with the needed book.

Herzig *et al.* provides precise definitions of the concepts of potential goal and dispositional trust, which captures the main points of C&F's discussion of the matter.

Formally, a *potential goal* is defined in $\mathcal{HHVL}$ as:

Definition 3.2.2.

$$\mathtt{PotGoal}_i(\varphi, k) \overset{\mathrm{def}}{=} \mathtt{Poss}_i(\mathtt{F}^*(k \wedge \mathtt{Choice}_i\mathtt{F}\varphi)).$$

So an agent i has the potential goal φ under circumstances k if and only if i believes it to be possible that she wants φ to be true at some point in the future when circumstances k hold.

The informal definition of dispositional trust runs as follows [25, p. 227]: agent i is disposed to trust agent j with the performance of α in relation to the goal φ under circumstances k, if and only if:

1. i has the potential goal φ under circumstances k;
2. i believes that always, if i has the goal φ and k holds, then
   (a) j, by performing α, will bring about φ;
   (b) j will be capable to perform α;
   (c) j will intend to do α.

Formally, dispositional trust is defined as:

Definition 3.2.3.

$$\begin{aligned}\mathtt{DispTrust}(i, j, \alpha, \varphi, k) &\overset{\mathrm{def}}{=} \mathtt{PotGoal}_i(\varphi, k)\\ &\wedge \mathtt{Bel}_i\mathtt{G}^*((k \wedge \mathtt{Choice}_i\mathtt{F}\varphi) \rightarrow (\mathtt{Capable}_j(\alpha) \wedge \mathtt{Intends}_j(\alpha) \wedge \mathtt{After}_{j:\alpha}\varphi)).\end{aligned}$$

Herzig *et al.* do not provide formal definitions of distrust, mistrust and lack of trust. These concepts are, as seen in Chapter 2, important for the C&F theory of trust. I propose the following definitions, and then prove some properties of them.

First, the concept of distrust is informally defined in the C&F theory as the conjunction of a goal of the truster, and the truster's beliefs about the trustees lack of intention, or lack of capability, or lack of opportunity to perform the action relevant for the accomplishment of the truster's goal. Formally, distrust is defined as:

Definition 3.2.4.

$$\begin{aligned}\mathtt{DisTrust}(i, j, \alpha, \varphi) \overset{\mathrm{def}}{=} &\mathtt{Choice}_i\mathtt{F}\varphi\\ &\wedge \mathtt{Bel}_i(\neg\mathtt{Intends}_j(\alpha) \vee \neg\mathtt{After}_{j:\alpha}\varphi \vee \neg\mathtt{Capable}_j(\alpha)).\end{aligned}$$

The concept of mistrust is informally defined in the C&F theory as the conjunction of a goal of the truster, and the truster's belief that the trustee is capable and willing to accomplish the opposite of the truster's goal. Formally, this is defined in the logic in the following way:

Definition 3.2.5.

$$\begin{aligned}\texttt{MisTrust}(i,j,\alpha,\varphi) \overset{\text{def}}{=} & \texttt{Choice}_i\texttt{F}\varphi \\ & \wedge \texttt{Bel}_i(\texttt{Capable}_j(\alpha) \wedge \texttt{Intends}_j(\alpha) \wedge \texttt{After}_{j:\alpha}\neg\varphi).\end{aligned}$$

Lack of trust is, as seen in Chapter 2, the lack of belief about a trustee's capability and willingness. Thus, lack of trust can formally be defined as:

Definition 3.2.6.

$$\begin{aligned}\texttt{LackTrust}(i,j,\alpha,\varphi) \overset{\text{def}}{=} & \texttt{Choice}_i\texttt{F}\varphi \\ & \wedge \neg\texttt{Bel}_i(\texttt{Intends}_j(\alpha) \wedge \texttt{Capable}_j(\alpha) \wedge \texttt{After}_{j:\alpha}\varphi).\end{aligned}$$

The concept of dispositional trust as presented in Chapter 2 and formalized in this section can be extended to a capture a concept of dispositional *mistrust*. This is in accordance with the C&F theory, where dispositional trust is analysed as an evaluation of a trustee relative to a potential goal; dispositional mistrust in a trustee is a negative evaluation of the trustee's intentions. I propose the following definition of dispositional mistrust:

Definition 3.2.7.

$$\begin{aligned}&\texttt{DispMisTrust}(i,j,\alpha,\varphi) \overset{\text{def}}{=} \texttt{PotGoal}_i(\varphi,k) \\ &\wedge \texttt{Bel}_i\texttt{G}^*((k \wedge \texttt{Choice}_i\texttt{F}\varphi) \rightarrow (\texttt{Capable}_j(\alpha) \wedge \texttt{Intend}_j(\alpha) \wedge \texttt{After}_{j:\alpha}\neg\varphi)).\end{aligned}$$

The remaining part of this section contains a number of theorems stating formal properties of the defined concepts.

By the axioms $\mathbf{Inc}_{\mathbf{After,\ Does}}$, **IntAct1**, and **IntAct2**, the following theorem can be proven [25, p. 222]:

Theorem 3.2.1. *Let $i \in AGT$ and $\alpha \in ACT$. Then*

$$\vdash_{\mathcal{HHVL}} \texttt{Capable}_i(\alpha) \wedge \texttt{Intends}_i(\alpha) \leftrightarrow \texttt{Does}_{i:\alpha}\top.$$

This theorem highlights an important property of the relation between the operators $\texttt{After}_{i:\alpha}$ and $\texttt{Does}_{i:\alpha}$.

Theorem 3.2.2. *Let $i \in AGT$ and $\alpha \in ACT$. Then*

$$\vdash_{\mathcal{HHVL}} \texttt{After}_{i:\alpha}\varphi \wedge \texttt{Does}_{i:\alpha}\top \rightarrow \texttt{Does}_{i:\alpha}\varphi.$$

Proof. Assume the opposite, i.e. $\texttt{After}_{i:\alpha}\varphi \wedge \texttt{Does}_{i:\alpha}\top \wedge \neg\texttt{Does}_{i:\alpha}\varphi$.

By Axiom $\mathbf{Inc}_{\mathbf{After,\ Does}}$, it holds that $\vdash_{\mathcal{HHVL}} \texttt{After}_{i:\alpha}\varphi \rightarrow \neg\texttt{Does}_{i:\alpha}\neg\varphi$.

Thus, the initial assumption implies $\neg\texttt{Does}_{i:\alpha}\varphi \wedge \neg\texttt{Does}_{i:\alpha}\neg\varphi$, which in turn, by standard principles of propositional logic and distribution of the normal, existential operator $\texttt{Does}_{i:\alpha}$ over disjunction, is equivalent to $\neg\texttt{Does}_{i:\alpha}(\varphi \vee \neg\varphi)$. But the last formula is equivalent to $\neg\texttt{Does}_{i:\alpha}\top$.

Hence, the assumption implies a contradiction (i.e. $\texttt{Does}_{i:\alpha}\top \wedge \neg\texttt{Does}_{i:\alpha}\top$), which proves the theorem. □

The following theorem is proven by Herzig *et al.* [25, p. 224], and highlights some interesting properties of occurrent trust.

Theorem 3.2.3. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

(a) $\vdash_{\mathcal{HHVL}} \texttt{OccTrust}(i,j,\alpha,\varphi) \to \texttt{Bel}_i\texttt{Does}_{j:\alpha}\varphi$;

(b) $\vdash_{\mathcal{HHVL}} \texttt{OccTrust}(i,j,\alpha,\varphi) \to \texttt{Bel}_i\texttt{F}\varphi$;

(c) $\vdash_{\mathcal{HHVL}} \texttt{OccTrust}(i,j,\alpha,\varphi) \leftrightarrow \texttt{Bel}_i\texttt{OccTrust}(i,j,\alpha,\varphi)$;

(d) $\vdash_{\mathcal{HHVL}} \texttt{Bel}_i\neg\texttt{Capable}_j(\alpha) \to \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$;

(e) $\vdash_{\mathcal{HHVL}} \texttt{Bel}_i\neg\texttt{Intends}_j(\alpha) \to \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$;

(f) $\vdash_{\mathcal{HHVL}} \texttt{Bel}_i\neg\texttt{After}_{j:\alpha}\varphi \to \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$;

(g) $\vdash_{\mathcal{HHVL}} \texttt{Bel}_i\texttt{After}_{j:\alpha}\neg\varphi \to \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$.

The following theorem states some interesting properties of the notions of mistrust and distrust.

Theorem 3.2.4. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

(a) $\vdash_{\mathcal{HHVL}} \texttt{MisTrust}(i,j,\alpha,\varphi) \to \texttt{Bel}_i\texttt{Does}_{j:\alpha}\neg\varphi$;

(b) $\vdash_{\mathcal{HHVL}} \texttt{MisTrust}(i,j,\alpha,\varphi) \to \texttt{Bel}_i\texttt{F}\neg\varphi$;

(c) $\vdash_{\mathcal{HHVL}} \texttt{DisTrust}(i,j,\alpha,\varphi) \to \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$;

(d) $\vdash_{\mathcal{HHVL}} \texttt{MisTrust}(i,j,\alpha,\varphi) \to \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$.

Proof.
(a) First, by Definition 3.2.5,

$$\vdash_{\mathcal{HHVL}} \texttt{MisTrust}(i,j,\alpha,\varphi) \to \texttt{Bel}_i(\texttt{Capable}_j(\alpha) \wedge \texttt{Intend}_j(\alpha) \wedge \texttt{After}_{j:\alpha}\neg\varphi)$$

is a theorem. The right hand side of the implication is, by Theorem 3.2.1, Axiom $\mathbf{K_{Bel}}$, and Axiom $\mathbf{Nec_{Bel}}$ equivalent to $\texttt{Bel}_i(\texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{Does}_{j:\alpha}\top)$, which yields the theorem

$$\vdash_{\mathcal{HHVL}} \texttt{MisTrust}(i,j,\alpha,\varphi) \to \texttt{Bel}_i(\texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{Does}_{j:\alpha}\top).$$

Second, by Theorem 3.2.2, it holds that:

$$\vdash_{\mathcal{HHVL}} \texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{Does}_{j:\alpha}\top \to \texttt{Does}_{j:\alpha}\neg\varphi.$$

By Axioms $\mathbf{Nec_{Bel}}$ and $\mathbf{K_{Bel}}$, it is concluded that

$$\vdash_{\mathcal{HHVL}} \texttt{Bel}_i(\texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{Does}_{j:\alpha}\top) \to \texttt{Bel}_i\texttt{Does}_{j:\alpha}\neg\varphi,$$

from which the theorem follows.

(b) By Axiom $\mathbf{Inc_{G,\,Does}}$, $\vdash_{\mathcal{HHVL}} \texttt{Does}_{j:\alpha}\neg\varphi \to \texttt{F}\neg\varphi$ is a theorem. With $\mathbf{Nec_{Bel}}$ and $\mathbf{K_{Bel}}$, it follows that $\vdash_{\mathcal{HHVL}} \texttt{Bel}_i\texttt{Does}_{j:\alpha}\neg\varphi \to \texttt{Bel}_i\texttt{F}\neg\varphi$. The theorem follows from this and Theorem 3.2.4(a).

(c) Assume the opposite, i.e. $\mathtt{DisTrust}(i, j, \alpha, \varphi) \wedge \mathtt{OccTrust}(i, j, \alpha, \varphi)$. First,

$$\vdash_{\mathcal{HHVL}} \mathtt{DisTrust}(i, j, \alpha, \varphi) \rightarrow \mathtt{Bel}_i(\neg\mathtt{Intends}_j(\alpha) \vee \neg\mathtt{Capable}_j(\alpha) \vee \neg\mathtt{After}_{j:\alpha}\varphi)$$

by Definition 3.2.4.

Second,

$$\vdash_{\mathcal{HHVL}} \mathtt{Bel}_i(\neg\mathtt{Intends}_j(\alpha) \vee \neg\mathtt{Capable}_j(\alpha) \vee \neg\mathtt{After}_{j:\alpha}\varphi) \\ \rightarrow \neg\mathtt{Bel}_i(\mathtt{Intends}_j(\alpha) \wedge \mathtt{Capable}_j(\alpha) \wedge \mathtt{After}_{j:\alpha}\varphi)$$

by standard principles of propositional logic and Axiom $\mathbf{D_{Bel}}$. Now,

$$\vdash_{\mathcal{HHVL}} \mathtt{OccTrust}(i, j, \alpha, \varphi) \rightarrow \mathtt{Bel}_i(\mathtt{Intends}_j(\alpha) \wedge \mathtt{Capable}_j(\alpha) \wedge \mathtt{After}_{j:\alpha}\varphi)$$

by Definition 3.2.1. Hence, the initial assumption is contradicted, and the theorem holds.

(d) First, $\vdash_{\mathcal{HHVL}} \mathtt{MisTrust}(i, j, \alpha, \varphi) \rightarrow \mathtt{Bel}_i\mathtt{After}_{j:\alpha}\neg\varphi$ by Definition 3.2.5 and Axiom $\mathbf{K_{Bel}}$. From this and Theorem 3.2.3(g), the theorem follows. □

Theorem 3.2.5. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then*

$$\vdash_{\mathcal{HHVL}} \mathtt{OccTrust}(i, j, \alpha, \varphi \wedge \psi) \rightarrow \mathtt{OccTrust}(i, j, \alpha, \varphi) \wedge \mathtt{OccTrust}(i, j, \alpha, \psi).$$

Proof. Assume $\mathtt{OccTrust}(i, j, \alpha, \varphi \wedge \psi)$. This implies, by Definition 3.2.1 and distribution of the normal universal operators $\mathtt{Bel}_i$ and $\mathtt{After}_{j:\alpha}$ over conjunction,

$$\mathtt{Choice}_i\mathtt{F}(\varphi \vee \psi) \wedge \mathtt{Bel}_i\mathtt{Does}_{j:\alpha}\top \wedge \mathtt{Bel}_i\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Bel}_i\mathtt{After}_{j:\alpha}\psi.$$

The formula $\mathtt{Choice}_i\mathtt{F}(\varphi \wedge \psi)$ implies $\mathtt{Choice}_i(\mathtt{F}\varphi \wedge \mathtt{F}\psi)$, since $\mathtt{F}$ is a normal existential operator. By distribution of $\mathtt{Choice}_i$ over conjunction, it holds that

$$\mathtt{Choice}_i\mathtt{F}(\varphi \wedge \psi) \rightarrow \mathtt{Choice}_i\mathtt{F}\varphi \wedge \mathtt{Choice}_i\mathtt{F}\psi.$$

Hence, the following formulas are valid:

$$\mathtt{OccTrust}(i, j, \alpha, \varphi \wedge \psi) \rightarrow \mathtt{Choice}_i\mathtt{F}\varphi \wedge \mathtt{Bel}_i(\mathtt{Does}_{j:\alpha}\top \wedge \mathtt{After}_{j:\alpha}\varphi)$$

and

$$\mathtt{OccTrust}(i, j, \alpha, \varphi \wedge \psi) \rightarrow \mathtt{Choice}_i\mathtt{F}\psi \wedge \mathtt{Bel}_i(\mathtt{Does}_{j:\alpha}\top \wedge \mathtt{After}_{j:\alpha}\psi),$$

which proves the theorem. □

The converse of this theorem, i.e.

$$\mathtt{OccTrust}(i, j, \alpha, \varphi) \wedge \mathtt{OccTrust}(i, j, \alpha, \psi) \rightarrow \mathtt{OccTrust}(i, j, \alpha, \varphi \wedge \psi),$$

is not valid since $\mathtt{F}\varphi \wedge \mathtt{F}\psi \rightarrow \mathtt{F}(\varphi \wedge \psi)$ is not a theorem. This is evident from the fact that $\mathtt{F}$ is a normal existential operator.

Informally, it could be the case that a truster i trusts a trustee j with achieving φ by performing α, and trusts j with achieving ψ by performing α, but it is not necessarily a goal of i that φ and ψ are true *at the same time*. For example, i might have the goal of having money so she can pay for her daughter's education, and the goal of her daughter having a good education. i thinks that borrowing money from j will ultimately lead to the achievement of both goals, but not at the same time. Borrowing money from j will provide i with money, and thus the first goal is achieved at one time, and borrowing money from j will, ultimately, ensure that i's daughter has a good education sometime in the future.

Theorem 3.2.6. *Let* $i, j \in AGT$ *and* $\alpha \in ACT$*. Then*

$$\vdash_{\mathcal{HHVL}} \texttt{DispTrust}(i,j,\alpha,\varphi,k) \land \texttt{Choice}_i\texttt{F}\varphi \land \texttt{Bel}_i k \to \texttt{OccTrust}(i,j,\alpha,\varphi).$$

See [25, p. 228] for a proof of this theorem.

The next theorem states the corresponding property of dispositional mistrust.

Theorem 3.2.7. *Let* $i, j \in AGT$ *and* $\alpha \in ACT$*. Then*

$$\vdash_{\mathcal{HHVL}} \texttt{DispMisTrust}(i,j,\alpha,\varphi,k) \land \texttt{Choice}_i\texttt{F}\varphi \land \texttt{Bel}_i k \to \texttt{MisTrust}(i,j,\alpha,\varphi).$$

Proof. First, the left hand side of the implication in the theorem implies

$$\begin{aligned}&\texttt{Choice}_i\texttt{F}\varphi \land \texttt{Bel}_i(\texttt{Choice}_i\texttt{F}\varphi \land k)\land\\ &\quad\texttt{Bel}_i\texttt{G}^*((k \land \texttt{Choice}_i\texttt{F}\varphi) \to (\texttt{Capable}_j(\alpha) \land \texttt{Intend}_j(\alpha) \land \texttt{After}_{j:\alpha}\neg\varphi)).\end{aligned}$$

The formula $\texttt{Bel}_i(\texttt{Choice}_i\texttt{F}\varphi \land k)$ is obtained from Axiom $\mathbf{4_{Choice}}$ and distribution of $\texttt{Bel}_i$ over conjunction.

The above formula implies

$$\begin{aligned}&\texttt{Choice}_i\texttt{F}\varphi \land \texttt{Bel}_i(k \land \texttt{Choice}_i\texttt{F}\varphi)\land\\ &\quad\texttt{Bel}_i((k \land \texttt{Choice}_i\texttt{F}\varphi) \to (\texttt{Capable}_j(\alpha) \land \texttt{Intend}_j(\alpha) \land \texttt{After}_{j:\alpha}\neg\varphi))\end{aligned}$$

by the definition of $\texttt{G}^*$. From the above formula, it follows that $\texttt{Choice}_i\texttt{F}\varphi \land \texttt{Bel}_i(\texttt{Capable}_j(\alpha) \land \texttt{Intend}_j(\alpha) \land \texttt{After}_{j:\alpha}\varphi)$ by the use of Axiom $\mathbf{K_{Bel}}$.

$\texttt{MisTrust}(i,j,\alpha,\varphi)$ then follows by Definiiton 3.2.5. □

3.3 Demolombe and Lorini's logic

Demolombe and Lorini [14, 29] have developed a logic (here called $\mathcal{DL}$) with the aim of formalizing aspects of the C&F theory. The formalism presented here is based on their articles [14, 29].

3.3.1 Syntax

The syntactic primitives of the logic are: a nonempty, finite set of agents $AGT = \{i_1, i_2, ..., i_k\}$; a nonempty finite set of actions $ACT = \{e_1, e_2, ..., e_l\}$; and a nonempty set of atomic propositions $ATM = \{p, q, ...\}$. Variables $i, j, ...$

denote agents, and variables $\alpha, \beta, ...$ denote actions. The language of the logic is given by the following syntax rules:

$$\varphi ::= p \,|\, \neg\varphi \,|\, \varphi \vee \varphi \,|\, \mathtt{After}_{i:\alpha}\varphi \,|\, \mathtt{Does}_{i:\alpha}\varphi \,|\, \mathtt{Bel}_i\varphi \,|\, \mathtt{Goal}_i\varphi \,|\, \mathtt{Obg}\varphi,$$

where $p \in ATM$, $i \in AGT$, and $\alpha \in ACT$. The usual connectives ($\wedge, \rightarrow, \leftrightarrow$) and *true* and *false* ($\top$ and $\bot$) are defined as in Section 3.1.1. The intended readings of the operators are as follows, where $p \in ATM$, $i \in AGT$, and $\alpha \in ACT$:

- $\mathtt{After}_{i:\alpha}\varphi$: after agent i has performed action α, φ holds;
- $\mathtt{Does}_{i:\alpha}\varphi$: agent i does action α, and φ is true afterwards;
- $\mathtt{Bel}_i\varphi$: agent i believes φ to be true;
- $\mathtt{Goal}_i\varphi$: agent i prefer (has the goal) φ to be true;
- $\mathtt{Obg}\varphi$: it is obligatory that φ.[12]

The following abbreviations are defined:

$$\begin{aligned}
\mathtt{Can}_i(\alpha) &\overset{\text{def}}{=} \neg\mathtt{After}_{i:\alpha}\bot;\\
\mathtt{Int}_i(\alpha) &\overset{\text{def}}{=} \mathtt{Goal}_i\mathtt{Does}_{i:\alpha}\top;\\
\mathtt{X}\varphi &\overset{\text{def}}{=} \bigvee_{i\in AGT,\alpha\in ACT} \mathtt{Does}_{i:\alpha}\varphi,
\end{aligned}$$

with the intuitive readings:

- $\mathtt{Can}_i(\alpha)$: agent i is capable of doing action α;
- $\mathtt{Int}_i(\alpha)$: agent i intends to do action α;
- $\mathtt{X}\varphi$: φ will be true at next time.[13]

[12] This operator is supposed to express obligation as a social or moral concept. It should not be mistaken for a property of the logic itself—if φ is obligatory, it does not mean that φ *must* be true. Consider a short example. Assume that it is obligatory to call the police if someone is committing a robbery. This is expressed in the logic as $\mathtt{Obg}(\mathit{robbery} \rightarrow \mathit{callpolice})$. I will not make use of the operator Obg in the following formalization of the C&F theory. It is included here for the sake of completeness; I want to stay as close as possible to the original formalism.

[13] The expression $\bigvee_{i=1}^{i=n} A_i$ is an abbreviation for $A_1 \vee ... \vee A_n$ [2, p. 32]. If $\bigvee_{i\in AGT,\alpha\in ACT}\mathtt{Does}_{i:\alpha}\varphi$ is true, then at least one of the formulas $\mathtt{Does}_{i:\alpha}\varphi$ is true for some $i \in AGT$ and $\alpha \in ACT$; basically, it means that there is some performance of an action $\alpha \in ACT$ by an agent $i \in AGT$, such that φ is the result of the performance. Thus, the expression has some resemblance with the existential quantifier $\exists$ in predicate logic. However, since $\mathcal{DL}$ is an extension of propositional logic without quantifiers, instead of quantifying over possibly infinite domains, finite disjunctions are used (note that both AGT and ACT are finite). Note also that the logic models time discreetly. The definition of the next time operator X satisfies $\mathtt{X}\varphi \leftrightarrow \neg\mathtt{X}\neg\varphi$, which is a standard property in temporal logic [2, p. 239]. Note also that the definition of the operator X is possible because of the Axiom **Active**, which ensures that there always is a *next world*.

Axiomatics

The following are axioms of the logic $\mathcal{DL}$:

(PC) all theorems of propositional logic;

($\mathbf{K_{Bel}}$) $(\mathtt{Bel}_i\varphi \wedge \mathtt{Bel}_i(\varphi \rightarrow \psi)) \rightarrow \mathtt{Bel}_i\psi$;

($\mathbf{D_{Bel}}$) $\neg(\mathtt{Bel}_i\varphi \wedge \mathtt{Bel}_i\neg\varphi)$;

($\mathbf{4_{Bel}}$) $\mathtt{Bel}_i\varphi \rightarrow \mathtt{Bel}_i\mathtt{Bel}_i\varphi$;

($\mathbf{5_{Bel}}$) $\neg\mathtt{Bel}_i\varphi \rightarrow \mathtt{Bel}_i\neg\mathtt{Bel}_i\varphi$;

($\mathbf{K_{Goal}}$) $(\mathtt{Goal}_i\varphi \wedge \mathtt{Goal}_i(\varphi \rightarrow \psi)) \rightarrow \mathtt{Goal}_i\psi$;

($\mathbf{D_{Goal}}$) $\neg(\mathtt{Goal}_i\varphi \wedge \mathtt{Goal}_i\neg\varphi)$;

($\mathbf{K_{Obg}}$) $(\mathtt{Obg}\varphi \wedge \mathtt{Obg}(\varphi \rightarrow \psi)) \rightarrow \mathtt{Obg}\psi$;

($\mathbf{D_{Obg}}$) $\neg(\mathtt{Obg}\varphi \wedge \mathtt{Obg}\neg\varphi)$;

($\mathbf{K_{After}}$) $(\mathtt{After}_{i:\alpha}\varphi \wedge \mathtt{After}_{i:\alpha}(\varphi \rightarrow \psi)) \rightarrow \mathtt{After}_{i:\alpha}\psi$;

($\mathbf{K_{Does}}$) $(\mathtt{Does}_{i:\alpha}\varphi \wedge \neg\mathtt{Does}_{i:\alpha}\neg\psi) \rightarrow \mathtt{Does}_{i:\alpha}(\varphi \wedge \psi)$;

(WR) $\mathtt{Goal}_i\varphi \rightarrow \neg\mathtt{Bel}_i\neg\varphi$;

(PIntr) $\mathtt{Goal}_i\varphi \rightarrow \mathtt{Bel}_i\mathtt{Goal}_i\varphi$;

(NIntr) $\neg\mathtt{Goal}_i\varphi \rightarrow \mathtt{Bel}_i\neg\mathtt{Goal}_i\varphi$;

(BelObg) $\mathtt{Obg}\varphi \rightarrow \mathtt{Bel}_i\mathtt{Obg}\varphi$;

($\mathbf{Alt_{Act}}$) $\mathtt{Does}_{i:\alpha}\varphi \rightarrow \neg\mathtt{Does}_{j:\beta}\neg\varphi$;

(Active) $\bigvee_{i\in AGT,\alpha\in ACT}\mathtt{Does}_{i:\alpha}\top$;

($\mathbf{Inc_{Act}}$) $\mathtt{Does}_{i:\alpha}\varphi \rightarrow \neg\mathtt{After}_{i:\alpha}\neg\varphi$;

(IntAct1) $(\mathtt{Int}_i(\alpha) \wedge \mathtt{Can}_i(\alpha)) \rightarrow \mathtt{Does}_{i:\alpha}\top$;

(IntAct2) $\mathtt{Does}_{i:\alpha}\top \rightarrow \mathtt{Int}_i(\alpha)$;

(MP) from $\vdash_{\mathcal{DL}} \varphi \rightarrow \psi$ and $\vdash_{\mathcal{DL}} \varphi$, infer $\vdash_{\mathcal{DL}} \psi$;

($\mathbf{Nec_{Bel}}$) from $\vdash_{\mathcal{DL}} \varphi$, infer $\vdash_{\mathcal{DL}} \mathtt{Bel}_i\varphi$;

($\mathbf{Nec_{Goal}}$) from $\vdash_{\mathcal{DL}} \varphi$, infer $\vdash_{\mathcal{DL}} \mathtt{Goal}_i\varphi$;

($\mathbf{Nec_{Obg}}$) from $\vdash_{\mathcal{DL}} \varphi$, infer $\vdash_{\mathcal{DL}} \mathtt{Obg}\varphi$;

($\mathbf{Nec_{After}}$) from $\vdash_{\mathcal{DL}} \varphi$, infer $\vdash_{\mathcal{DL}} \mathtt{After}_{i:\alpha}\varphi$;

($\mathbf{Nec_{Does}}$) from $\vdash_{\mathcal{DL}} \varphi$, infer $\vdash_{\mathcal{DL}} \neg\mathtt{Does}_{i:\alpha}\neg\varphi$.

3.3.2 Semantics

Frames in the logic are tuples $F = \langle W, R, D, B, G, O \rangle$, where the following hold:

- W is a nonempty set of possible worlds.
- R is a collection of binary relations $R_{i:\alpha}$ on W for every $i \in AGT$ and $\alpha \in ACT$. For every world $w \in W$, if $(w, w') \in R_{i:\alpha}$, then w' is a world that can be reached from w through the performance of α by i.
- D is a collection of binary relations $D_{i:\alpha}$ on W for every $i \in AGT$ and $\alpha \in ACT$. For every world $w \in W$, if $(w, w') \in D_{i:\alpha}$, then w' is the *next* world from w, which will be reached through i's performance of α.
- B is a collection of binary relations B_i on W for every $i \in AGT$. For any world $w \in W$, if $(w, w') \in B_i$, then w' is a world that is compatible with i's beliefs at world w.
- G is a collection of binary relations G_i on W for every $i \in AGT$. For any world $w \in W$, if $(w, w') \in G_i$, then w' is a world that is compatible with i's goals at world w.
- O is a binary relation on W. For any world $w \in W$, if $(w, w') \in O$, then w' is an ideal world at world w.

All operators of type $\texttt{Bel}_i$ are *KD45* modal operator, all operators of type $\texttt{Goal}_i$, and the $\texttt{Obg}$ operator are *KD* modal operators. $\texttt{After}_{i:\alpha}$ and $\texttt{Does}_{i:\alpha}$ satisfy the axioms and rules of inference of the modal system K.

The special semantic constraints for frames are the following [14]. For every $w \in W$, $i, j \in AGT$, and $\alpha, \beta \in ACT$:

S1 if $(w, w') \in D_{i:\alpha}$ and $(w, w'') \in D_{j:\beta}$, then $w' = w''$;[14]

S2 there are $i \in AGT$, $\alpha \in ACT$, and $w' \in W$, such that $(w, w') \in D_{i:\alpha}$;

S3 if $(w, w') \in D_{i:\alpha}$, then $(w, w') \in R_{i:\alpha}$;

S4 if, for all $(w, w') \in G_i$, there are w'', v such that $(w', w'') \in D_{i:\alpha}$ and $(w, v) \in R_{i:\alpha}$, then there is v' such that $(w, v') \in D_{i:\alpha}$;

S5 if there is v' such that $(w, v') \in D_{i:\alpha}$, then, for all $(w, w') \in G_i$, there is w'' such that $(w', w'') \in D_{i:\alpha}$;

S6 there is a w' such that $(w, w') \in B_i$ and $(w, w') \in G_i$;

S7 if $(w, w') \in B_i$, then, for all v, if $(w, v) \in G_i$, then $(w', v) \in G_i$;

S8 if $(w, w') \in B_i$, then, for all v, if $(w', v) \in G_i$, then $(w, v) \in G_i$.

Models in the logic are couples $M = \langle F, V \rangle$, where F is a frame, and $V : ATM \to 2^W$ is a valuation function, where $V(p)$ denotes the set of worlds $w \in W$ where p is true. The expression $M, w \vDash \varphi$ denotes that φ is true at world w in model M. Truth conditions for atomic propositions, the usual connectives, and the constants $\top$ and $\bot$ are defined as in Definition 3.1.4. The remaining truth conditions are the following:

[14] This is the *deterministic* property of relations.

- $M, w \vDash \texttt{After}_{i:\alpha}\varphi$ iff $M, w' \vDash \varphi$ for all w' such that $(w, w') \in R_{i:\alpha}$;
- $M, w \vDash \texttt{Does}_{i:\alpha}\varphi$ iff there is a world w' such that $M, w' \vDash \varphi$ and $(w, w') \in D_{i:\alpha}$;
- $M, w \vDash \texttt{Bel}_i\varphi$ iff $M, w' \vDash \varphi$ for all w' such that $(w, w') \in B_i$;
- $M, w \vDash \texttt{Goal}_i\varphi$ iff $M, w' \vDash \varphi$ for all w' such that $(w, w') \in G_i$;
- $M, w \vDash \texttt{Obg}\varphi$ iff $M, w' \vDash \varphi$ for all w' such that $(w, w') \in O$.

If there exists a model M and a world w in M where a formula φ is satisfied, φ is said to be $\mathcal{DL}$-satisfiable.

3.3.3 Formalizing the C&F theory

The formal translation of core trust is:

Definition 3.3.1.

$$\texttt{ATrust}(i, j, \alpha, \varphi) \overset{\text{def}}{=} \texttt{Goal}_i\texttt{X}\varphi \wedge \texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Can}_j(\alpha) \wedge \texttt{Int}_j(\alpha)).$$

Demolombe and Lorini do not provide definitions of the concepts of distrust, mistrust, and lack of trust. I propose the following definitions.

Definition 3.3.2.

$$\texttt{DisTrust}(i, j, \alpha, \varphi) \overset{\text{def}}{=} \texttt{Goal}_i\texttt{X}\varphi \wedge \texttt{Bel}_i(\neg\texttt{After}_{j:\alpha}\varphi \vee \neg\texttt{Can}_j(\alpha) \vee \neg\texttt{Int}_j(\alpha)).$$

Definition 3.3.3.

$$\texttt{MisTrust}(i, j, \alpha, \varphi) \overset{\text{def}}{=} \texttt{Goal}_i\texttt{X}\varphi \wedge \texttt{Bel}_i(\texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{Can}_j(\alpha) \wedge \texttt{Int}_j(\alpha)).$$

Definition 3.3.4.

$$\texttt{LackTrust}(i, j, \alpha, \varphi) \overset{\text{def}}{=} \texttt{Goal}_i\texttt{X}\varphi \wedge \neg\texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Can}_j(\alpha) \wedge \texttt{Int}_j(\alpha)).$$

I will end this section by stating a few theorems of $\mathcal{DL}$.

Theorem 3.3.1. *Let $i \in AGT$ and $\alpha \in ACT$. Then*

(a) $\vdash_{\mathcal{DL}} \texttt{Int}_i(\alpha) \wedge \texttt{Can}_i(\alpha) \leftrightarrow \texttt{Does}_{i:\alpha}\top$;

(b) $\vdash_{\mathcal{DL}} \texttt{Does}_{i:\alpha}\top \wedge \texttt{After}_{i:\alpha}\varphi \rightarrow \texttt{Does}_{i:\alpha}\varphi$;

(c) $\vdash_{\mathcal{DL}} \texttt{X}(\varphi \wedge \psi) \leftrightarrow \texttt{X}\varphi \wedge \texttt{X}\psi$.

Proof.
(a) First, the implication from left to right holds by Axiom **IntAct1**.

Second, by Axiom $\mathbf{Inc_{Act}}$, $\vdash_{\mathcal{DL}} \texttt{Does}_{i:\alpha}\top \rightarrow \neg\texttt{After}_{i:\alpha}\bot$. By the definition of $\texttt{Can}_i(\alpha)$, this is equivalent to $\vdash_{\mathcal{DL}} \texttt{Does}_{i:\alpha}\top \rightarrow \texttt{Can}_i(\alpha)$. Axiom **IntAct2** states that $\texttt{Does}_{i:\alpha}\top \rightarrow \texttt{Int}_i(\alpha)$. Thus, the implication from right to left holds.

(b) Assume the opposite, i.e. $\vdash_{\mathcal{DL}} \texttt{Does}_{i:\alpha}\top \wedge \texttt{After}_{i:\alpha}\varphi \wedge \neg\texttt{Does}_{i:\alpha}\varphi$.

By Axiom $\mathbf{Int_{Act}}$, $\vdash_{\mathcal{DL}} \texttt{After}_{i:\alpha}\varphi \rightarrow \neg\texttt{Does}_{i:\alpha}\neg\varphi$.

The assumption thus implies $\neg\texttt{Does}_{i:\alpha}\varphi \wedge \neg\texttt{Does}_{i:\alpha}\varphi$, which, by standard principles of propositional logic and distribution of $\texttt{Does}_{i:\alpha}$ over disjunction ($\texttt{Does}_{i:\alpha}$ is a normal existential modal operator) is equivalent to $\neg\texttt{Does}_{i:\alpha}(\varphi \vee \neg\varphi)$.

But $\varphi \vee \neg\varphi$ is equivalent to $\top$, and the assumption thus implies $\texttt{Does}_{i:\alpha}\top \wedge \neg\texttt{Does}_{i:\alpha}\top$, which is a contradiction. Hence, the theorem is proven.

(c) First, the implication from left to right is proved.

Assume that $\texttt{X}(\varphi \wedge \psi)$. By the definition of $\texttt{X}$, the formula $\texttt{X}(\varphi \wedge \psi)$ is equivalent to

$$\bigvee_{i \in AGT, \alpha \in ACT} \texttt{Does}_{i:\alpha}(\varphi \wedge \psi),$$

that is, there is some combination of $i \in AGT$ and $\alpha \in ACT$ such that $\texttt{Does}_{i:\alpha}(\varphi \wedge \psi)$. Since $\texttt{Does}_{i:\alpha}$ is a normal existential operator, this implies that $\texttt{Does}_{i:\alpha}\varphi \wedge \texttt{Does}_{i:\alpha}\psi$ is true. This, by the definition of $\texttt{X}$, means that $\texttt{X}\varphi \wedge \texttt{X}\psi$ is true, so $\vdash_{\mathcal{DL}} \texttt{X}(\varphi \wedge \psi) \to \texttt{X}\varphi \wedge \texttt{X}\psi$.

Second, the implication from right to left is proved.

Assume $\texttt{X}\varphi \wedge \texttt{X}\psi$, which is equivalent to

$$\bigvee_{i \in AGT, \alpha \in ACT} \texttt{Does}_{i:\alpha}\varphi \wedge \bigvee_{j \in AGT, \beta \in ACT} \texttt{Does}_{j:\beta}\psi.$$

So there is one formula $\texttt{Does}_{i:\alpha}\varphi \wedge \texttt{Does}_{j:\beta}\psi$, for $i, j \in AGT$ and $\alpha, \beta \in ACT$, that is true.

By Axiom $\mathbf{Alt_{Act}}$, $\texttt{Does}_{i:\alpha}\varphi \wedge \texttt{Does}_{j:\beta}\psi$ implies $\texttt{Does}_{i:\alpha}\varphi \wedge \neg\texttt{Does}_{i:\alpha}\neg\psi$, which in turn, by Axiom $\mathbf{K_{Does}}$ implies $\texttt{Does}_{i:\alpha}(\varphi \wedge \psi)$. By the definition of $\texttt{X}$, this implies $\texttt{X}(\varphi \wedge \psi)$.

Hence, $\vdash_{\mathcal{DL}} \texttt{X}\varphi \wedge \texttt{X}\psi \to \texttt{X}(\varphi \wedge \psi)$. □

The following theorem is proven by Demolombe and Lorini [29].

Theorem 3.3.2. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then*

$$\vdash_{\mathcal{DL}} \texttt{ATrust}(i, j, \alpha, \varphi) \to \texttt{Bel}_i\texttt{X}\varphi.$$

The following theorem highlights interesting properties of trust, distrust, and mistrust.

Theorem 3.3.3. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

(a) $\vdash_{\mathcal{DL}} \texttt{Bel}_i\neg\texttt{Can}_j(\alpha) \to \neg\texttt{ATrust}(i, j, \alpha, \varphi)$;

(b) $\vdash_{\mathcal{DL}} \texttt{Bel}_i\neg\texttt{Int}_j(\alpha) \to \neg\texttt{ATrust}(i, j, \alpha, \varphi)$;

(c) $\vdash_{\mathcal{DL}} \texttt{Bel}_i\neg\texttt{After}_{j:\alpha}\varphi \to \neg\texttt{ATrust}(i, j, \alpha, \varphi)$;

(d) $\vdash_{\mathcal{DL}} \texttt{Bel}_i\texttt{After}_{j:\alpha}\neg\varphi \to \neg\texttt{ATrust}(i, j, \alpha, \varphi)$;

(e) $\vdash_{\mathcal{DL}} \texttt{DisTrust}(i, j, \alpha, \varphi) \to \neg\texttt{ATrust}(i, j, \alpha, \varphi)$;

(f) $\vdash_{\mathcal{DL}} \texttt{MisTrust}(i, j, \alpha, \varphi) \to \neg\texttt{ATrust}(i, j, \alpha, \varphi)$.

Proof.
(a) First, $\vdash_{\mathcal{DL}}$ $\mathtt{ATrust}(i,j,\alpha,\varphi) \rightarrow \mathtt{Bel}_i\mathtt{Can}_j(\alpha)$ by Definition 3.3.1. Second, $\vdash_{\mathcal{DL}}$ $\mathtt{Bel}_i\mathtt{Can}_i(\alpha) \rightarrow \neg\mathtt{Bel}_i\neg\mathtt{Can}_j(\alpha)$ by Axiom $\mathbf{D_{Bel}}$. The theorem then follows from contraposition.

(b) The proof is similar to that of (a). By Definition 3.3.1 and Axiom $\mathbf{D_{Bel}}$ $\vdash_{\mathcal{DL}}$ $\mathtt{ATrust}(i,j,\alpha,\varphi) \rightarrow \neg\mathtt{Bel}_i\neg\mathtt{Int}_j(\alpha)$, to which the theorem is the contrapositive.

(c) Again, the same kind of proof is used as in (a) and (b).
$\vdash_{\mathcal{DL}}$ $\mathtt{ATrust}(i,j,\alpha,\varphi) \rightarrow \neg\mathtt{After}_{j:\alpha}\neg\varphi$ holds by Definition 3.3.1 and Axiom $\mathbf{D_{Bel}}$, from which the theorem follows by contraposition.

(d) Assume that the opposite hold, i.e. $\vdash_{\mathcal{DL}}$ $\mathtt{Bel}_i\mathtt{After}_{j:\alpha}\neg\varphi \wedge \mathtt{ATrust}(i,j,\alpha,\varphi)$.
From Definition 3.3.1,

$$\vdash_{\mathcal{DL}} \mathtt{ATrust}(i,j,\alpha,\varphi) \rightarrow \mathtt{Bel}_i(\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Can}_j(\alpha) \wedge \mathtt{Int}_j(\alpha)).$$

This, together with Theorem (a) yields

$$\vdash_{\mathcal{DL}} \mathtt{ATrust}(i,j,\alpha,\varphi) \rightarrow \mathtt{Bel}_i(\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Does}_{j:\alpha}\top).$$

Theorem (b) states that

$$\vdash_{\mathcal{DL}} \mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Does}_{j:\alpha}\top \rightarrow \mathtt{Does}_{j:\alpha}\varphi.$$

With Axioms $\mathbf{Nec_{Bel}}$ and $\mathbf{K_{Bel}}$, $\vdash_{\mathcal{DL}}$ $\mathtt{Bel}_i(\mathtt{After}_{j:\alpha}\wedge\mathtt{Does}_{j:\alpha}\top) \rightarrow \mathtt{Bel}_i\mathtt{Does}_{j:\alpha}\varphi$. From the above, it can be concluded that

$$\vdash_{\mathcal{DL}} \mathtt{ATrust}(i,j,\alpha,\varphi) \rightarrow \mathtt{Bel}_i\mathtt{Does}_{j:\alpha}\varphi.$$

By Axiom $\mathbf{Inc_{Act}}$, $\vdash_{\mathcal{DL}}$ $\mathtt{After}_{j:\alpha}\neg\varphi \rightarrow \neg\mathtt{Does}_{j:\alpha}\varphi$, and by Axioms $\mathbf{Nec_{Bel}}$ and $\mathbf{K_{Bel}}$, it follows that $\vdash_{\mathcal{DL}}$ $\mathtt{Bel}_i\mathtt{After}_{j:\alpha}\neg\varphi \rightarrow \mathtt{Bel}_i\neg\mathtt{Does}_{j:\alpha}\varphi$. By Axiom $\mathbf{D_{Bel}}$, $\vdash_{\mathcal{DL}}$ $\mathtt{Bel}_i\neg\mathtt{Does}_{j:\alpha}\varphi \rightarrow \neg\mathtt{Bel}_i\mathtt{Does}_{j:\alpha}\varphi$.
Thus, it follows from all of the above that

$$\vdash_{\mathcal{DL}} \mathtt{Bel}_i\mathtt{After}_{j:\alpha}\neg\varphi \wedge \mathtt{ATrust}(i,j,\alpha,\varphi) \rightarrow \bot,$$

from which it follows that the theorem holds.

(e) By Definition 3.3.2,

$$\vdash_{\mathcal{DL}} \mathtt{DisTrust}(i,j,\alpha,\varphi) \rightarrow \mathtt{Bel}_i(\neg\mathtt{After}_{j:\alpha}\varphi \vee \neg\mathtt{Can}_j(\alpha) \vee \neg\mathtt{Int}_j(\alpha)).$$

The right hand side of the implication is equivalent to

$$\mathtt{Bel}_i\neg(\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Can}_j(\alpha) \wedge \mathtt{Int}_j(\alpha))$$

by standard principles of propositional logic.
By Axiom $\mathbf{D_{Bel}}$,

$$\begin{aligned}\vdash_{\mathcal{DL}} &\mathtt{Bel}_i\neg(\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Can}_j(\alpha) \wedge \mathtt{Int}_j(\alpha))\\ &\rightarrow \neg\mathtt{Bel}_i(\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Can}_j(\alpha) \wedge \mathtt{Int}_j(\alpha)).\end{aligned}$$

Hence,

$$\vdash_{\mathcal{DL}} \texttt{DisTrust}(i, j, \alpha, \varphi) \to \neg\texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Can}_j(\alpha) \wedge \texttt{Int}_j(\alpha)).$$

By Definition 3.3.1,

$$\vdash_{\mathcal{DL}} \texttt{ATrust}(i, j, \alpha, \varphi) \to \texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Can}_j(\alpha) \wedge \texttt{Int}_j(\alpha)).$$

All the above leads to $\vdash_{\mathcal{DL}} \texttt{DisTrust}(i, j, \alpha, \varphi) \wedge \texttt{ATrust}(i, j, \alpha, \varphi) \to \bot$, from which the theorem follows by contradiction.

(f) First, form Definition 3.3.3, $\vdash_{\mathcal{DL}} \texttt{MisTrust}(i, j, \alpha, \varphi) \to \texttt{Bel}_i\texttt{After}_{j:\alpha}\neg\varphi$, from which the theorem follows by use of Theorem 3.3.3(d).

□

By using the Theorem (c), the following theorem can be proved.

Theorem 3.3.4. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then*

$$\vdash_{\mathcal{DL}} \texttt{ATrust}(i, j, \alpha, \varphi) \wedge \texttt{ATrust}(i, j, \alpha, \psi) \leftrightarrow \texttt{ATrust}(i, j, \alpha, \varphi \wedge \psi).$$

Proof. The implication from right to left is proved first.

Assume that $\texttt{ATrust}(i, j, \alpha, \varphi \wedge \psi)$. This implies, by Definition 3.3.1 and distribution of the normal universal operators $\texttt{Bel}_i$ and $\texttt{After}_{j:\alpha}$ over conjunction,

$$\texttt{Goal}_i\texttt{X}(\varphi \wedge \psi) \wedge \texttt{Bel}_i(\texttt{Does}_{i:\alpha}\top \wedge \texttt{Bel}_i\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Bel}_i\texttt{After}_{j:\alpha}\psi.$$

By Theorem (c), and distribution of $\texttt{Goal}_i$ and $\texttt{Bel}_i$ over conjunction, it follows that

$$\texttt{ATrust}(i, j, \alpha, \varphi \wedge \psi) \to \texttt{Goal}_i\texttt{X}\varphi \wedge \texttt{Bel}_i(\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\varphi)$$

and

$$\texttt{ATrust}(i, j, \alpha, \varphi \wedge \psi) \to \texttt{Goal}_i\texttt{X}\psi \wedge \texttt{Bel}_i(\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\psi),$$

which proves the theorem from right to left.

Now, the implication form left to right is proved.

Assume that $\texttt{ATrust}(i, j, \alpha, \varphi) \wedge \texttt{ATrust}(i, j, \alpha, \psi)$. This implies, by Definition 3.3.1 and distribution of the normal universal operators $\texttt{Bel}_i$ and $\texttt{After}_{j:\alpha}$ over conjunction

$$\texttt{Goal}_i(\texttt{X}\varphi \wedge \texttt{X}\psi) \wedge \texttt{Bel}_i\texttt{Does}_{j:\alpha}\top \wedge \texttt{Bel}_i\texttt{After}_{j:\alpha}(\varphi \wedge \psi).$$

This, together with Theorem (c), shows that

$$\vdash_{\mathcal{DL}} \texttt{ATrust}(i, j, \alpha, \varphi) \wedge \texttt{ATrust}(i, j, \alpha, \psi) \to \\ \texttt{Goal}_i(\varphi \wedge \psi) \wedge \texttt{Bel}_i(\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}(\varphi \wedge \psi)),$$

which proves the theorem.

□

3.4 Bonnefon, Longin, and Nguyen's logic

In their paper [5], Bonnefon, Longin, and Nguyen construct a logic (here denoted $\mathcal{BNL}$) with the purpose of reasoning about emotions, like fear, anger, and disappointment, in relation to trust (see also their earlier paper [4]). I will only consider the logic in relation to the C&F theory, and thus ignore the formalizations of emotions.

3.4.1 Syntax

The syntactic primitives of the logic are: a nonempty finite set of agents, $AGT = \{i_1, i_2, ..., i_n\}$; a nonempty finite set of atomic events $EVT = \{e_1, e_2, ..., e_p\}$; and a nonempty set of atomic propositions $ATM = \{p, q, ...\}$. Agents are denoted with variables $i, j, k, ...$ The expression $i_1 : e_1 \in AGT \times EVT$ denotes an event e_1 intentionally caused by agent i_1; such events are called *actions*. Variables $\alpha, \beta, ...$ denote actions. The language of the logic is given by the following syntax rules:

$$\varphi ::= p \,|\, i{:}\alpha\textit{-happens} \,|\, \neg\varphi \,|\, \varphi \vee \varphi \,|\, \mathtt{X}\varphi \,|\, \mathtt{X}^{-1}\varphi \,|\, \mathtt{G}\varphi \,|\, \mathtt{Bel}_i\varphi \,|\, \mathtt{Choice}_i\varphi \,|\, \mathtt{Grd}_I\varphi,$$

where $p \in ATM$, $i{:}\alpha \in AGT \times EVT$, $i{:}\alpha\textit{-happens} \in ATM$ for each $i{:}\alpha \in AGT \times EVT$, and $I \subseteq AGT$. The usual connectives $(\wedge, \rightarrow, \leftrightarrow)$ and *true* and *false* ($\top$ and $\bot$) are defined from $\neg$ and $\vee$ as in Section 3.1.1.

The operators have the following readings:

- $i{:}\alpha\textit{-happens}$: agent i is just about to perform action α;
- $\mathtt{X}\varphi$: φ will be true at the next time;
- $\mathtt{X}^{-1}\varphi$: φ was true at the previous time;
- $\mathtt{G}\varphi$: φ is true from now on;
- $\mathtt{Bel}_i\varphi$: agent i believes φ to be true;
- $\mathtt{Choice}_i\varphi$: agent i wants φ to be true;
- $\mathtt{Grd}_I\varphi$: φ is publicly grounded between agents in I.[15]

[15] Bonnefon *et al.* [5] assert that the operator $\mathtt{Grd}_I$ is a *common belief* operator. This is quite a broad characterisation, since there are several ways to express mutual beliefs in a group of agents. In the most basic sense, a common belief φ in a group I of agents is just the conjunction of the individual agents' beliefs that φ holds. However, the $\mathtt{Grd}_I$ operator used here is more close to the concept of grounded information presented and discussed by Gaudou, Herzig, and Longin [23], where the fact that it is grounded in a group I of agents that φ holds does not imply individual beliefs that φ holds for every agent in I. Section 3.4.4 contains further discussions on this topic (see also e.g. [23, 41]).

The following abbreviations are defined:

$$\begin{aligned}
\textit{i:}\alpha\textit{-done} &\overset{\text{def}}{=} \mathtt{X}^{-1}\textit{i:}\alpha\textit{-happens};\\
\mathtt{Happens}_{i:\alpha}\varphi &\overset{\text{def}}{=} \textit{i:}\alpha\textit{-happens} \wedge \mathtt{X}\varphi;\\
\mathtt{After}_{i:\alpha}\varphi &\overset{\text{def}}{=} \textit{i:}\alpha\textit{-happens} \rightarrow \mathtt{X}\varphi;\\
\mathtt{Done}_{i:\alpha}\varphi &\overset{\text{def}}{=} \textit{i:}\alpha\textit{-done} \wedge \mathtt{X}^{-1}\varphi;\\
\mathtt{F}\varphi &\overset{\text{def}}{=} \neg\mathtt{G}\neg\varphi;\\
\mathtt{Goal}_i\varphi &\overset{\text{def}}{=} \mathtt{Choice}_i\mathtt{F}(\mathtt{Bel}_i\varphi);\\
\mathtt{Intend}_i\alpha &\overset{\text{def}}{=} \mathtt{Choice}_i\mathtt{F}\textit{i:}\alpha\textit{-happens};\\
\mathtt{Capable}_i\alpha &\overset{\text{def}}{=} \neg\mathtt{After}_{i:\alpha}\bot;\\
\mathtt{Possible}_i\varphi &\overset{\text{def}}{=} \neg\mathtt{Bel}_i\neg\varphi;
\end{aligned}$$

with the intended readings:

- $\textit{i:}\alpha\textit{-done}$: agent i has done action α;
- $\mathtt{Happens}_{i:\alpha}\varphi$: agent i is doing action α, and φ will be true next;
- $\mathtt{After}_{i:\alpha}\varphi$: φ is true after the execution of α by i;
- $\mathtt{Done}_{i:\alpha}\varphi$: agent i has done action α, and φ was true at the previous time;
- $\mathtt{F}\varphi$: φ will be true at some future time;
- $\mathtt{Goal}_i\varphi$: agent i has the goal that φ be true;
- $\mathtt{Intend}_i\alpha$: agent i intends to do action α;
- $\mathtt{Capable}_i\alpha$: agent i is capable of doing action α;
- $\mathtt{Possible}_i\varphi$: agent i believes φ to be possible.

Axiomatics

The following are the axioms of $\mathcal{BNL}$:

(PC) all theorems of propositional logic;

(1) $\textit{i:}\alpha\textit{-happens} \leftrightarrow \mathtt{X}\textit{i:}\alpha\textit{-done}$;

(2) $\mathtt{X}\varphi \leftrightarrow \neg\mathtt{X}\neg\varphi$;

(3) $\varphi \leftrightarrow \mathtt{X}\mathtt{X}^{-1}\varphi$;

(4) $\varphi \leftrightarrow \mathtt{X}^{-1}\mathtt{X}\varphi$;

(5) $\mathtt{G}\varphi \leftrightarrow \varphi \wedge \mathtt{X}\mathtt{G}\varphi$;

(6) $\mathtt{G}(\varphi \rightarrow \mathtt{X}\varphi) \rightarrow (\varphi \rightarrow \mathtt{G}\varphi)$;

(7) $\mathtt{G}\varphi \rightarrow \mathtt{After}_{i:\alpha}\varphi$;

(8) $\mathtt{Happens}_{i:\alpha}\varphi \rightarrow \mathtt{After}_{j:\beta}\varphi$;

(9) $\mathtt{After}_{i:\alpha}\varphi \leftrightarrow \neg\mathtt{Happens}_{i:\alpha}\neg\varphi$;

($\mathbf{K_{Bel}}$) $(\mathtt{Bel}_i\varphi \wedge \mathtt{Bel}_i(\varphi \rightarrow \psi)) \rightarrow \mathtt{Bel}_i\psi$;

($\mathbf{D_{Bel}}$) $\mathtt{Bel}_i\varphi \rightarrow \neg\mathtt{Bel}_i\neg\varphi$;

($\mathbf{4_{Bel}}'$) $\mathtt{Bel}_i\varphi \leftrightarrow \mathtt{Bel}_i\mathtt{Bel}_i\varphi$;[16]

($\mathbf{5_{Bel}}'$) $\neg\mathtt{Bel}_i\varphi \leftrightarrow \mathtt{Bel}_i\neg\mathtt{Bel}_i\varphi$;

($\mathbf{K_{Choice}}$) $(\mathtt{Choice}_i\varphi \wedge \mathtt{Choice}_i(\varphi \rightarrow \psi)) \rightarrow \mathtt{Choice}_i\psi)$;

($\mathbf{D_{Choice}}$) $\mathtt{Choice}_i\varphi \rightarrow \neg\mathtt{Choice}_i\neg\varphi$;

($\mathbf{4_{Choice}}'$) $\mathtt{Choice}_i\varphi \leftrightarrow \mathtt{Bel}_i\mathtt{Choice}_i\varphi$;

($\mathbf{5_{Choice}}'$) $\neg\mathtt{Choice}_i\varphi \leftrightarrow \mathtt{Bel}_i\neg\mathtt{Choice}_i\varphi$;

($\mathbf{K_{Grd}}$) $(\mathtt{Grd}_I\varphi \wedge \mathtt{Grd}_I(\varphi \rightarrow \psi)) \rightarrow \mathtt{Grd}_I\psi)$;

($\mathbf{D_{Grd}}$) $\mathtt{Grd}_I\varphi \rightarrow \neg\mathtt{Grd}_I\neg\varphi$;

($\mathbf{4_{Grd}}'$) $\mathtt{Grd}_I\varphi \leftrightarrow \mathtt{Grd}_I\mathtt{Grd}_I\varphi$;

($\mathbf{5_{Grd}}'$) $\neg\mathtt{Grd}_I\varphi \leftrightarrow \mathtt{Grd}_I\neg\mathtt{Grd}_I\varphi$;

(MP) from $\vdash_{\mathcal{BNL}} \varphi \rightarrow \psi$ and $\vdash_{\mathcal{BNL}} \varphi$, infer $\vdash_{\mathcal{BNL}} \psi$;

($\mathbf{Nec_{Bel}}$) from $\vdash_{\mathcal{BNL}} \varphi$, infer $\vdash_{\mathcal{BNL}} \mathtt{Bel}_i\varphi$;

($\mathbf{Nec_{Choice}}$) from $\vdash_{\mathcal{BNL}} \varphi$, infer $\vdash_{\mathcal{BNL}} \mathtt{Choice}_i\varphi$;

($\mathbf{Nec_{Grd}}$) from $\vdash_{\mathcal{BNL}} \varphi$, infer $\vdash_{\mathcal{BNL}} \mathtt{Grd}_I\varphi$.

3.4.2 Semantics

A frame F in the logic is a 4-tuple $\langle H, B, C, G\rangle$, where:

- H is a set of *stories*, represented by a sequence of time points. Each time point is represented by an integer $z \in \mathbb{Z}$. A time point z in a story h is called a situation, and is denoted $<h, z>$.

- B is the set of all relations B_i such that $B_i(h, z)$ denotes the set of stories believed to be possible by agent i in the situation $<h, z>$.

- C is the set of all relations C_i, such that $C_i(h, z)$ denotes the set of stories h chosen by agent i in situation $<h, z>$.

- G is the set of all G_I such that $G_I(h, z)$ denotes the set of stories which are grounded between the agents in group I in the situation $<h, z>$.

[16] Note that this axiom differs from the usual **4**-type axioms in a *KD45* logic, in that it uses an equivalence ↔ instead of en implication →; hence the name $\mathbf{4_{Bel}}'$ instead of $\mathbf{4_{Bel}}$.

The following are the semantic constrains of the logic. All accessibility relations $B_i \in B$ are serial, transitive, and euclidean. All accessibility relations $G_I \in G$ are serial, transitive, and euclidean. All accessibility relations $C_i \in C$ are serial. For every $z \in \mathbb{Z}$, if $h' \in B_i(h, z)$, then $C_i(h, z) = C_i(h', z)$; if agent i believes that the story h' is possible from the story h, then i's preferred stories from h and h' are the same. In short, this means that agents are aware of their preferences; all agents believe what they prefer. A model M for the logic is a couple $\langle F, V \rangle$, where F is a frame and V is a function which associates each proposition p with a set $V(p)$ of couples (h, z) where p is true.

Truth conditions are defined as follows:

- $M, h, z \vDash p$ iff $(h, z) \in V(p)$;
- $M, h, z \vDash \mathtt{X}\varphi$ iff $M, h, z + 1 \vDash \varphi$;
- $M, h, z \vDash \mathtt{X}^{-1}\varphi$ iff $M, h, z - 1 \vDash \varphi$;
- $M, h, z \vDash \mathtt{G}\varphi$ iff $M, h, z' \vDash \varphi$ for every $z' \geq z$;
- $M, h, z \vDash \mathtt{Bel}_i\varphi$ iff $M, h', z \vDash \varphi$ for every $(h', z) \in B_i(h, z)$;
- $M, h, z \vDash \mathtt{Choice}_i\varphi$ iff $M, h', z \vDash \varphi$ for every $(h', z) \in C_i(h, z)$;
- $M, h, z \vDash \mathtt{Grd}_I\varphi$ iff $M, h', z \vDash \varphi$ for every $(h', z) \in G_I(h, z)$.

Truth conditions for the usual connectives and the constants $\top$ and $\bot$ are defined similar to Definition 3.1.4 (i.e. $M, h, z \vDash \varphi \wedge \psi$ iff $M, h, z \vDash \varphi$ and $M, h, z \vDash \psi$, etc.).

3.4.3 Formalizing the C&F theory

In the logic $\mathcal{BNL}$, Bonnefon *et al.* [5] define the following concept of trust:

Definition 3.4.1.

$$\begin{aligned}\mathtt{Trust}_{i,j}(\alpha, \varphi) \overset{\mathrm{def}}{=} & \mathtt{Goal}_i\varphi \wedge \mathtt{Bel}_i\mathtt{After}_{j:\alpha}\varphi \\ & \wedge \mathtt{Bel}_i\mathtt{Capable}_j\alpha \wedge \mathtt{Bel}_i\mathtt{Intend}_j\alpha \wedge \mathtt{Grd}_{\{i,j\}} j{:}\alpha\text{-}\mathit{happens}.\end{aligned}$$

This definition is supposed to correspond to the informal definition of core trust in the C&F theory. However, the additional clause $\mathtt{Grd}_{\{i,j\}} j{:}\alpha\text{-}\mathit{happens}$ is added, expressing that a truster i can only trust a trustee j if it is grounded between them that j is going to perform action α.

Bonnefon *et al.* [5] argue for the adding of the Grd clause in the trust definition with a variant of the following example. Suppose that a burglar breaks into an office building with the goal of stealing money from the boss's office. In another room in the building, the boss's secretary is busy filing reports. The burglar wants the secretary to stay in the other room, since it makes his stealing possible. The burglar believes that the secretary is capable of staying in the other room, and also that it is the secretary's intention of doing so. Thus, according to the C&F definition of core trust, the burglar has core trust in the secretary in relation to the action of staying in the other room and the goal of stealing the money. However, Bonnefon *et al.* [5] argue that one should be reluctant to say that the burglar actually *trusts* the secretary; the burglar merely *relies* on her. This is because there is no agreement between the burglar and the secretary that the secretary will stay in the other room when the burglar empties the safe.

The concept of distrust is defined in the following way:

Definition 3.4.2.

$$\begin{aligned}\texttt{DisTrust}_{i,j}(\alpha,\varphi) &\overset{\text{def}}{=} \texttt{Goal}_i\varphi \\ &\wedge (\texttt{Bel}_i\neg\texttt{After}_{j:\alpha}\varphi \vee (\texttt{Possible}_i\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Bel}_i\neg\texttt{Intend}_j\alpha)).\end{aligned}$$

The following are theorems of $\mathcal{BNL}$.

Theorem 3.4.1. *Let $i, j \in AGT$ and $\alpha \in EVT$. Then:*

(a) $\vdash_{\mathcal{BNL}} \texttt{Bel}_i\neg\texttt{After}_{j:\alpha}\varphi \rightarrow \neg\texttt{Trust}_{i,j}(\alpha,\varphi)$;

(b) $\vdash_{\mathcal{BNL}} \texttt{Bel}_i\neg\texttt{Intend}_j\alpha \rightarrow \neg\texttt{Trust}_{i,j}(\alpha,\varphi)$;

(c) $\vdash_{\mathcal{BNL}} \texttt{Bel}_i\neg\texttt{Capable}_j\alpha \rightarrow \neg\texttt{Trust}_{i,j}(\alpha,\varphi)$;

(d) $\vdash_{\mathcal{BNL}} \texttt{DisTrust}_{i,j}(\alpha,\varphi) \rightarrow \neg\texttt{Trust}_{i,j}(\alpha,\varphi)$.

Proof.
(a) The proof is by contradiction. Assume that the opposite holds, i.e.

$$\vdash_{\mathcal{BNL}} \texttt{Bel}_i\neg\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Trust}_{i,j}(\alpha,\varphi).$$

By definition 3.4.1, $\vdash_{\mathcal{BNL}} \texttt{Trust}_{i,j}(\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{After}_{j:\alpha}\varphi$, and by Axiom $\mathbf{D_{Bel}}$, $\vdash_{\mathcal{BNL}} \texttt{Bel}_i\texttt{After}_{j:\alpha}\varphi \rightarrow \neg\texttt{Bel}_i\neg\texttt{After}_{j:\alpha}\varphi$. Thus, the initial assumption leads to a contradiction, and hence the theorem holds.

(b) The proof is similar to that of (a). Assume the opposite, i.e.

$$\vdash_{\mathcal{BNL}} \texttt{Bel}_i\neg\texttt{Intend}_j\alpha \wedge \texttt{Trust}_{i,j}(\alpha,\varphi).$$

The following holds by Definition 3.4.1: $\vdash_{\mathcal{BNL}} \texttt{Trust}_{i,j}(\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{Intend}_j\alpha$. By Axiom $\mathbf{D_{Bel}}$, $\vdash_{\mathcal{BNL}} \texttt{Bel}_i\texttt{Intend}_j\alpha \rightarrow \neg\texttt{Bel}_i\neg\texttt{Intend}_j\alpha$, which contradicts the assumption, so that the theorem holds.

(c) The proof is similar to the proofs of (a) and (b). Assume the opposite of the proposed theorem:

$$\vdash_{\mathcal{BNL}} \texttt{Bel}_i\neg\texttt{Capable}_j\alpha \wedge \texttt{Trust}_{i,j}(\alpha,\varphi).$$

The following are theorems of $\mathcal{BNL}$: by Definition 3.4.1,

$$\vdash_{\mathcal{BNL}} \texttt{Trust}_{i,j}(\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{Capable}_j\alpha,$$

and by Axiom $\mathbf{D_{Bel}}$

$$\vdash_{\mathcal{BNL}} \texttt{Bel}_i\texttt{Capable}_j\alpha \rightarrow \neg\texttt{Bel}_i\neg\texttt{Capable}_j\alpha.$$

Thus, the assumption leads to a contradiction, which shows that the theorem holds.

(d) Assume the opposite, i.e. $\vdash_{\mathcal{BNL}} \mathtt{DisTrust}_{i,j}(\alpha,\varphi) \wedge \mathtt{Trust}_{i,j}(\alpha,\varphi)$. First, by contraposition of (a), $\vdash_{\mathcal{BNL}} \mathtt{Trust}_{i,j}(\alpha,\varphi) \rightarrow \neg\mathtt{Bel}_i\neg\mathtt{After}_{j:\alpha}\varphi$. The following holds by Definition 3.4.2:

$$\begin{aligned}\vdash_{\mathcal{BNL}} & \mathtt{DisTrust}_{i,j}(\alpha,\varphi) \\ & \rightarrow (\mathtt{Bel}_i\neg\mathtt{After}_{j:\alpha}\varphi \vee (\mathtt{Possible}_i\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Bel}_i\neg\mathtt{Intend}_j\alpha)).\end{aligned}$$

Hence, since $\mathtt{Bel}_i\neg\mathtt{After}_{i:\alpha}\varphi \wedge \neg\mathtt{Bel}_i\neg\mathtt{After}_{i:\alpha}\varphi$ is a contradiction, under the assumption it must hold that

$$\begin{aligned}\vdash_{\mathcal{BNL}} & \mathtt{DisTrust}_{i,j}(\alpha,\varphi) \wedge \mathtt{Trust}_{i,j}(\alpha,\varphi) \\ & \rightarrow \mathtt{Possible}_i\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Bel}_i\neg\mathtt{Intend}_j\alpha.\end{aligned}$$

But, by contraposition of (b), it holds that

$$\vdash_{\mathcal{BNL}} \mathtt{Trust}_{i,j}(\alpha,\varphi) \rightarrow \neg\mathtt{Bel}_i\neg\mathtt{Intend}_i\alpha,$$

which leads to a contradiction, and hence, the theorem holds.

□

3.4.4 Groundedness

In $\mathcal{BNL}$, the concept of trust is defined (Definition 3.4.1) as

$$\begin{aligned}\mathtt{Trust}_{i,j}(\alpha,\varphi) \overset{\text{def}}{=} & \mathtt{Goal}_i\varphi \wedge \mathtt{Bel}_i\mathtt{After}_{j:\alpha}\varphi \\ & \wedge \mathtt{Bel}_i\mathtt{Capable}_j\alpha \wedge \mathtt{Bel}_i\mathtt{Intend}_j\alpha \wedge \mathtt{Grd}_{\{i,j\}}\textit{j:}\alpha\textit{-happens}.\end{aligned}$$

This definition is in important aspects different from the definition of core trust proposed by C&F, in that it requires a kind of agreement between the truster and the trustee; it has to be grounded between them that the truster is about to perform the relevant action.

I have two lines of criticism of the above definition. The first questions whether it is necessary to publicly ground that the trustee will do the required action in order to be trusted. The second questions the particular use of the $\mathtt{Grd}_I$ operator in relation to the C&F theory.

First, consider this common language example of how the word 'trust' is sometimes used. Suppose that I have to leave my apartment in a hurry, and will not be back in a couple of days. I have a cat who needs to be fed every day. Luckily, my roommate will be home during my absence and is able to feed the cat. If I come home and find that my roommate has not fed the cat, I might say "But I trusted you!". Even though there is no agreement between me and my roommate that she would feed the cat during my absence, I *trust* her with the task. Thus, there can be trust without an agreement.

Second, the $\mathtt{Grd}_I$ operator goes against the idea of defining trust as the strictly *mental* counterpart to delegation. According to C&F theory, core trust is a mental notion, consisting of a truster's beliefs. Bonnefon *et al.* state that their operator $\mathtt{Grd}_I$ is a mutual belief operator. Informally, the most basic idea of a mutual belief in a group I of agents $i, j, ..., k$, means that each agent in the group believes a proposition to be true. Thus, if the idea of mutual belief between a truster and a trustee is introduced as part of the definition of core trust, core trust is no longer a strictly mental concept of the truster, since it would take into account beliefs of the trustee as well. Doing so would be to go against one of the main claims of the C&F theory.

Bonnefon *et al.*'s operator $\texttt{Grd}_I$ does however not express such a basic sense of mutual belief. The operator $\texttt{Grd}_I$ is more akin to the concept of mutual belief as developed in Gaudou, Herzig, and Longin [23], where some piece of information φ can be publicly grounded in a group I of agents without the need for every agent in I to agree on the truth of φ. According to Gaudou, Herzig, and Longin,

> Groundedness is an objective notion: it refers to what can be observed, and only to that. While it is related to mental states because it corresponds to the expression of Intentional states, it is not an Intentional state: it is neither a belief nor a goal, nor an intention. [23, p. 3]

Thus, even though groundedness need not be constructed as mutual belief in the basic sense, and accordingly incorporate the trustee's beliefs in the core trust definition, it is still in conflict with the chief claim of the C&F theory that trust is a strictly mental notion of the truster.

The groundedness condition in Bonnefon *et al.*'s definition of trust also limits the generality of the concept of trust. Even though my focus in this essay is trust in intentional agents, it is an important aspect of the C&F theory that trust in intentional and trust in non-intentional objects essentially are of the same kind. The groundedness condition limits the applicability of the trust-concept to cases of trust in intentional agents.

3.4.5 Intention and action

When comparing the three formalisms, it becomes clear that the logic of Bonnefon *et al.* has weak semantic constraints governing the interaction of operators, and accordingly, few interaction axioms, in comparison with the logic of Herzig *et al.* and the logic of Demolombe and Lorini. This leads to certain complications when the logic is evaluated with respect to how well it formalizes the C&F theory.

In $\mathcal{HHVL}$ and $\mathcal{DL}$, respectively, the following are theorems:

$$\vdash_{\mathcal{HHVL}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{F}\varphi;$$
$$\vdash_{\mathcal{DL}} \texttt{ATrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{X}\varphi;$$

expressing that if one trusts someone with performing an action to reach a goal, then one expects that the goal will hold in the future. These properties of the logics $\mathcal{HHVL}$ and $\mathcal{DL}$ capture the claim of the C&F theory that if one trusts a trustee, one has a *positive expectation* that one's goal will be true. A positive expectation is the combination of a prediction about the future, and the desire that the prediction will hold (see Chapter 2). A similar principle holds in $\mathcal{BNL}$.

Theorem 3.4.2. *Let $i,j \in AGT$ and $\alpha \in EVT$. Then*

$$\vdash_{\mathcal{BNL}} \texttt{Trust}_{i,j}(\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{X}\varphi.$$

However, this theorem turns out to be slightly problematic in relation to the C&F theory.

The following lemma is used in the proof of Theorem 3.4.2.

Lemma 3.4.1. *Let $i, j \in AGT$ and $\alpha \in EVT$. Then*

$$\vdash_{\mathcal{BNL}} \texttt{After}_{j:\alpha}\varphi \wedge \texttt{Capable}_j\alpha \to \texttt{X}\varphi.$$

Proof. First, assume that $\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Capable}_j\alpha$ is true. By the definitions of $\texttt{After}_{j:\alpha}$ and $\texttt{Capable}_j$ (see Section 3.4.1), this is equivalent to

$$(j{:}\alpha\textit{-happens} \to \texttt{X}\varphi) \wedge \neg(j{:}\alpha\textit{-happens} \to \texttt{X}\bot).$$

Second, by standard principles of propositional logic, the following formula is equivalent:

$$(j{:}\alpha\textit{-happens} \to \texttt{X}\varphi) \wedge j{:}\alpha\textit{-happens} \wedge \neg\texttt{X}\bot.$$

The formula $\neg\texttt{X}\bot$ is equivalent to $\texttt{X}\top$ by Axiom **2** and the definitions of $\top$ and $\bot$ (see Section 3.1.1), so the following formula is true under the assumption:

$$(j{:}\alpha\textit{-happens} \to \texttt{X}\varphi) \wedge j{:}\alpha\textit{-happens} \wedge \texttt{X}\top.$$

By standard properties of propositional logic, it holds that

$$(j{:}\alpha\textit{-happens} \to \texttt{X}\varphi) \wedge j{:}\alpha\textit{-happens} \wedge \texttt{X}\top \to \texttt{X}\varphi.$$

Consequently, the above shows that

$$\vdash_{\mathcal{BNL}} \texttt{After}_{j:\alpha}\varphi \wedge \texttt{Capable}_j\alpha \to \texttt{X}\varphi$$

holds.

□

Now, Theorem 3.4.2 can be proved.

Proof. First, $\vdash_{\mathcal{BNL}} \texttt{Trust}_{i,j}(\alpha, \varphi) \to \texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Capable}_j\alpha)$ by the definition of $\texttt{Trust}_{i,j}(\alpha, \varphi)$ (Definition 3.4.1) and distribution of $\texttt{Bel}_i$ over conjunction. Second, by Lemma 3.4.1 and Axioms $\mathbf{Nec_{Bel}}$ and $\mathbf{K_{Bel}}$, it holds that $\vdash_{\mathcal{BNL}} \texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Capable}_j\alpha) \to \texttt{Bel}_i\texttt{X}\varphi$.

The theorem follows from the above. □

However, this renders the truster's belief about the trustee's intention redundant. This goes against the claim of the C&F theory that both beliefs about the trustee's capability and intentions (or willingness) are essential in forming expectations about the trustee's behaviour [12, p. 53]. Theorem 3.4.2 is an indirect consequence of the weak semantic constraints and corresponding lack of interaction axioms in $\mathcal{BNL}$. In particular, there are no axioms governing the interaction between intentions and actions. It is natural to assume that some actions by agent i are brought about intentionally; this is also an implicit assumption in the C&F theory when dealing with trust in intentional agents.

Herzig and Lorini [31] state the principle that, if α is an action, and agent i intends to do α, then i's intention is what brings about i's attempt to do α [31]. If agent i actually *can* do α, i will succeed in her attempt. This is captured in, for instance, the logic of Demolombe and Lorini [29], where the following is an axiom:[17]

$$\mathtt{Int}_i(\alpha) \wedge \mathtt{Can}_i(\alpha) \to \mathtt{Does}_{i:\alpha}\top$$

The reason for the lack of such an interaction axiom in the logic of Bonnefon *et al.* might be found in the formalizations of emotions; however, for the rest of this essay, it suffices to conclude that the logic of Bonnefon *et al.* does not suit my purposes.

[17] On the other hand, in both $\mathcal{DL}$ and $\mathcal{HHVL}$, all actions are intentional actions. This is evident from Theorems 3.2.1 and 3.3.1(a). This claim can possibly be put to doubt, but it suffices to notice that the claim that *no* actions are intentional is much more problematic in relation to the C&F theory.

Chapter 4

Scenario-based comparison of the logics $\mathcal{HHVL}$ and $\mathcal{DL}$

The following is a simple case study of an Internet forum scenario. The case study serves two purposes: first, by formalizing a scenario using the two considered logics in parallel, grounds of comparison of the logics' performances are established. Second, the formalization shows how the C&F theory could be used to reason about formal inference and establishment of trust in an Internet communications context. Relating to the latter, a logical formalization of trust on an Internet forum could form a theoretical framework for implementing assistant-tools, with the purposes of aiding the work of moderators; such an assistant tool could, for instance, partially automate the process of finding vandalism and spot Internet trolls. The scenario is inspired by a paper by Krupa, Vercouter, and Hübner [27], where trust based assessments of Wikipedia contributors are used to partially automate the work of Wikipedia's moderators.[1]

So, what is the scenario?

4.1 The Internet forum

Consider an Internet forum on some topic. The agents involved are of two kinds: regular users and moderators. The role of the moderators is to maintain high quality on the forum. This is done by approving or not approving posts submitted by regular users, as well as warn and suspend users who causes—deliberately or not—bad quality on the forum.

Here, an Internet forum will be studied from the point of view of moderators' trust in regular users; trust in a user forms the basis of approving posts submitted by the user, while mistrust and distrust in a user form the basis for not approving the user's posts, and in some cases, provide reason for a moderator to warn or suspend the user.

I assume that the overall quality on the forum is dependent on the quality of individual posts; if a sufficient number of posts on the forum is of high quality, then the forum is of high quality. Conversely, the forum is of low quality if there

[1] Wikipedia's moderators consist of volunteers, and there are many different moderating roles. Krupa *et al.*'s article focuses on the "Recent Changes Patrol" [27].

are too many low quality posts. In order to formalize this relation, I assume that the forum has a threaded interface, i.e. the forum consists of a number of threads, which in turn consists of a number of posts.

With this construction in place, I will assume that the forum is of high overall quality if a sufficient number of threads are of high quality. A thread, in turn, is of high quality if no posts in the thread are of low quality. For example, if a thread is vandalized by a troll, the thread is considered to be of low quality. This allows for a quite natural analysis of how the actions of making a post in a thread directly affects the quality of the thread. It can now be stated that after a user j has made a post in a thread t, t is of high or low quality. Hence, the goal of a high quality forum is reduced to the set of goals of high quality in threads.

For both $\mathcal{HHVL}$ and $\mathcal{DL}$, let $T = \{t_1, t_2, ..., t_n\}$ be a finite set of *forum threads*, and let

- $hq{:}t \in ATM$ denote that thread $t \in T$ is of high quality,
- $mp{:}t \in ACT$ denote the action of making a new post in thread $t \in T$,
- $vd{:}t \in ACT$ denote the action of vandalizing the thread $t \in T$.

Let $M = \{i_1, ..., i_k\}$ be a finite set of *moderators*, and $U = \{j_1, ..., j_l\}$ be a finite set of *regular users*, such that $AGT = M \cup U$ and $M \cap U = \emptyset$.

4.2 The goal component

In BDI approaches to agency, a common distinction is that between *achievement goals* and *maintenance goals*. An achievement goal is the goal to achieve something, which does not currently hold, and a maintenance goal is the goal to maintain a certain state of affairs the way they are. Herzig *et al.* [25] formalizes this distinction in $\mathcal{HHVL}$ in the following way. Achievement goals are defined as:

$$\texttt{AGoal}_i\varphi \overset{\text{def}}{=} \texttt{Choice}_i\texttt{F}\varphi \wedge \neg\texttt{Bel}_i\varphi,$$

and maintenance goals as

$$\texttt{MGoal}_i\varphi \overset{\text{def}}{=} \texttt{Choice}_i\texttt{G}\varphi \wedge \texttt{Bel}_i\varphi.$$

In $\mathcal{DL}$, goals of agents are expressed by the formula

$$\texttt{Goal}_i\texttt{X}\varphi.$$

In the Internet forum scenario, maintenance goal sophisticates the analysis. For example, the goal of a high quality forum can be thought of as a maintenance goal (under the assumption that a sufficient number of threads on the forum is of high quality). It is natural to assume that all moderators have the goal of keeping all threads of high quality at all future times, i.e.

$$\bigwedge_{i \in M, t \in T} \texttt{Choice}_i\texttt{G}hq{:}t.$$

Note that the condition $\texttt{Bel}_i hq{:}t$ for all $t \in T$ is ignored here; this is because moderators want threads to be of high quality at all future times, even if they believe that some threads currently are of low quality.

This kind of maintenance goal requires the ability to reason about *all future times*, which is why it cannot be expressed in $\mathcal{DL}$.

The following example highlights the importance of accurately defined goals.

Example 4.2.1. Suppose that a moderator i has the goal of achieving high quality of the content in thread t. Suppose that i has the additional goal of the forum to become the largest and most comprehensive Internet community on its subject, and that i does not believe that this goal currently holds.

In $\mathcal{HHVL}$, the two goals are expressed as

$$\texttt{Choice}_i\mathrm{F}\mathit{hq}{:}t \wedge \texttt{Choice}_i\mathrm{F}\mathit{largestforum}.$$

In $\mathcal{DL}$, the corresponding formal expression is

$$\texttt{Goal}_i\texttt{X}\mathit{hq}{:}t \wedge \texttt{Goal}_i\texttt{X}\mathit{largestforum}.$$

However, it is slightly problematic to hold that i has the goal of the forum to be the largest Internet community at *the next time*. If the forum currently is far away from being the largest Internet community, then the goal of the forum being the largest Internet community is most naturally characterised as some kind of *ultimate* goal, a goal that i wants to hold at some point in the future, but not necessarily at the next instant in time.

Additionally, in $\mathcal{DL}$, the above formula is equivalent to

$$\texttt{Goal}_i\texttt{X}(\mathit{hq}{:}t \wedge \mathit{largestforum}).$$

However, i might be prepared to compromise the quality of thread t, if it means that her goal of the forum being the largest Internet community on its subject can be reached.

4.3 Trust in users

Moderators can trust users with making posts in order to contribute to the high quality of the forum. As seen, core trust is the most basic component of trust relations, and consists of the conjunction of a goal of the truster and the belief that the trustee is willing and able, and has the power/opportunity to ensure the truster's goal by performing a certain action. Trust in forum users is a little different from the general case of trust, in that moderators never doubt the capability and intention of a user when evaluating her. Typically, moderators are dependent on forum users for reaching the goal of a high quality forum (a high quality forum should be active and up-to-date, and if no one writes posts in the forum, then the forum stagnates), but they are not (weakly or strongly) dependent on particular users for reaching their goal. When trust in a forum user is inferred, it is under the condition that the user actually has submitted a post (this is also noted in [27]).

In $\mathcal{HHVL}$, Theorem 3.2.1 states that

$$\vdash_{\mathcal{HHVL}} \texttt{Capable}_i(\alpha) \wedge \texttt{Intends}_i(\alpha) \leftrightarrow \texttt{Does}_{i:\alpha}\top$$

and in $\mathcal{DL}$, Theorem 3.3.1(a) says that

$$\vdash_{\mathcal{DL}} \texttt{Can}_i(\alpha) \wedge \texttt{Int}_i(\alpha) \leftrightarrow \texttt{Does}_{i:\alpha}\top.$$

Thus, the formal definitions of trust can be simplified by replacing the conditions concerning the trustee j's capability and willingness in relation to the action α with the formula $\texttt{Does}_{i:\alpha}\top$ (the notation is the same in both $\mathcal{HHVL}$ and $\mathcal{DL}$).

A moderator i's core trust in a user j in relation to the action of making a post in thread t and the goal of high quality in thread t is expressed as follows. In $\mathcal{HHVL}$, for $i \in M, j \in U$ and $t \in T$:

$$\texttt{OccTrust}(i, j, mp{:}t, hq{:}t) \overset{\text{def}}{=} \texttt{Choice}_i\texttt{F}hq{:}t \wedge \texttt{Bel}_i(\texttt{Does}_{j:mp:t}\top \wedge \texttt{After}_{j:mp:t}hq{:}t).$$

In $\mathcal{DL}$, for $i \in M, j \in U$ and $t \in T$:

$$\texttt{ATrust}(i, j, mp{:}t, hq{:}t) \overset{\text{def}}{=} \texttt{Goal}_i\texttt{X}hq{:}t \wedge \texttt{Bel}_i(\texttt{After}_{j:mp:t}hq{:}t \wedge \texttt{Does}_{j:mp:t}\top).$$

Now, consider some examples.

Example 4.3.1. Moderator i decides to consult a colleague x about a user j, who has made a post in thread t. If i believes that x trusts j, i decides to trust j as well, i.e.

$$\texttt{Bel}_i(\texttt{OccTrust}(x, j, mp{:}t, hq{:}t) \to \texttt{OccTrust}(i, j, mp{:}t, hq{:}t)).$$

Furthermore, i believes that x has the goal of high quality in t, i believes that x believes that j has submitted a post in t, and i believes that x believes that j's post will contribute to the high quality in t, i.e.

$$\texttt{Bel}_i\texttt{Choice}_x\texttt{F}(hq{:}t)$$

and

$$\texttt{Bel}_i\texttt{Bel}_x\texttt{Does}_{j:mp:t}\top$$

and

$$\texttt{Bel}_i\texttt{Bel}_x\texttt{After}_{j:mp:t}(hq{:}t).$$

Thus, i believes that x trusts j:

$$\texttt{Bel}_i\texttt{OccTrust}(x, j, mp{:}t, hq{:}t)$$

and hence i believes that she trusts j about making a post in t:

$$\texttt{Bel}_i\texttt{OccTrust}(i, j, mp{:}t, hq{:}t),$$

which is equivalent to

$$\texttt{OccTrust}(i, j, mp{:}t, hq{:}t).$$

4.4 Distrust in users

In a way similar to the cases of trust in users, distrust in a user is only dependent on the belief about the user's power to contribute to the high quality in a thread t by making a post in t (the condition $\texttt{After}_{j:mp:t}hq{:}t$, expressed using the same symbols in both $\mathcal{HHVL}$ and $\mathcal{DL}$). Thus, a moderator i distrusts a user j if i believes that it is not the case that after j has made a post in thread t, t is of high quality.

The general definitions of distrust (Definitions 3.2.4 and 3.3.2) are as follows. In $\mathcal{HHVL}$:

$$\begin{aligned}\mathtt{DisTrust}(i,j,\alpha,\varphi) \overset{\text{def}}{=} & \mathtt{Choice}_i\mathtt{F}\varphi \\ & \wedge \mathtt{Bel}_i(\neg\mathtt{Intends}_j(\alpha) \vee \neg\mathtt{After}_{j:\alpha}\varphi \vee \neg\mathtt{Capable}_j(\alpha)),\end{aligned}$$

and in $\mathcal{DL}$:

$$\mathtt{DisTrust}(i,j,\alpha,\varphi) \overset{\text{def}}{=} \mathtt{Goal}_i\mathtt{X}\varphi \wedge \mathtt{Bel}_i(\neg\mathtt{After}_{j:\alpha}\varphi \vee \neg\mathtt{Can}_j(\alpha) \vee \neg\mathtt{Int}_j(\alpha)).$$

However, because of the circumstances just mentioned, i.e. that $\mathtt{Does}_{j:mp:t}\top$ is always true when a moderator i distrusts a user j, I propose the following definition: A moderator i distrusts a user j with contributing to the high quality in a thread t if i believes that j lacks the opportunity (or power) to achieve or maintain high quality in t by making a post in t. Formally, in $\mathcal{HHVL}$, this translates as:

$$\begin{aligned}\mathtt{DisTrust}(i,j,\mathit{mp{:}t},\mathit{hq{:}t}) \overset{\text{def}}{=} & \mathtt{Choice}_i\mathtt{F}\mathit{hq{:}t} \\ & \wedge \mathtt{Bel}_i(\mathtt{Does}_{j:mp:t}\top \wedge \neg\mathtt{After}_{j:mp:t}\mathit{hq{:}t}).\end{aligned}$$

In $\mathcal{DL}$, the corresponding definition is:

$$\begin{aligned}\mathtt{DisTrust}(i,j,\mathit{mp{:}t},\mathit{hq{:}t}) \overset{\text{def}}{=} & \mathtt{Choice}_i\mathtt{X}\mathit{hq{:}t} \\ & \wedge \mathtt{Bel}_i(\mathtt{Does}_{j:mp:t}\top \wedge \neg\mathtt{After}_{j:mp:t}\mathit{hq{:}t}).\end{aligned}$$

Example 4.4.1. Suppose that moderator i decides to distrust a user j with making a post in thread t in relation to i's goal of high quality if one of her moderator colleagues x, y, or z distrusts j:

$$\begin{aligned}\mathtt{Bel}_i(&\mathtt{DisTrust}(x,j,\mathit{mp{:}t},\mathit{hq{:}t}) \vee \mathtt{DisTrust}(y,j,\mathit{mp{:}t},\mathit{hq{:}t}) \\ & \vee \mathtt{DisTrust}(z,j,\mathit{mp{:}t},\mathit{hq{:}t}) \to \mathtt{DisTrust}(i,j,\mathit{mp{:}t},\mathit{hq{:}t})).\end{aligned}$$

Further, suppose that i believes that x distrusts j:

$$\mathtt{Bel}_i\mathtt{DisTrust}(x,j,\mathit{mp{:}t},\mathit{hq{:}t}).$$

Then, i believes that she distrusts j:

$$\mathtt{Bel}_i\mathtt{DisTrust}(i,j,\mathit{mp{:}t},\mathit{hq{:}t}).$$

4.5 Mistrust in trolls

According to Hardaker's [24], an *Internet troll* is a user of Internet communication "who constructs the identity of sincerely wishing to be part of the group in question, including professing, or conveying pseudo-sincere intentions, but whose real intention(s) is/are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement" [24, p. 237]. Thus, a troll is a forum user who intends to vandalize the forum, and by doing so, causes conflict and/or general disruption. Vandalizing a thread t in such a way leads to low quality in t.[2] Therefore, moderators mistrust users they believe to be trolls.

Note that the mistrust in a user should rely on evaluations of the user's intention (this is also acknowledged in [27]). In fact, a moderator i believing that a user j has submitted a post with the intention of vandalizing should be inclined to mistrust j, regardless of whether j actually succeeds to cause bad quality on the forum. Formally, this can be expressed by the following definitions. In $\mathcal{HHVL}$:

$$\texttt{MisTrust}(i, j, vd{:}t, hq{:}t) \overset{\text{def}}{=} \texttt{Choice}_i\mathsf{F}hq{:}t \wedge \texttt{Bel}_i\texttt{Does}_{j:vd:t}\top.$$

In $\mathcal{DL}$:

$$\texttt{MisTrust}(i, j, vd{:}t, hq{:}t) \overset{\text{def}}{=} \texttt{Goal}_i\texttt{X}hq{:}t \wedge \texttt{Bel}_i\texttt{Does}_{j:vd:t}\top.$$

However, the reason for mistrusting a user who intentionally submits vandalism is that such posts *usually* or *in most cases* causes the forum to be of low quality. As an alternative approach, one could assume that every moderator believes that every user of the forum has the opportunity to cause bad quality in t by vandalizing t, i.e., for both $\mathcal{HHVL}$ and $\mathcal{DL}$:

$$\bigwedge_{i\in M}\texttt{Bel}_i\bigwedge_{j\in U, t\in T}\texttt{After}_{j:vd:t}\neg hq{:}t.$$

A moderator i's mistrust in a user j relative to the goal of high quality in thread t and the action of vandalizing t can then be expressed as follows. In $\mathcal{HHVL}$:

$$\texttt{MisTrust}(i, j, vd{:}t, hq{:}t) \overset{\text{def}}{=} \texttt{Choice}_i\mathsf{F}hq{:}t \wedge \texttt{Bel}_i(\texttt{Does}_{j:vd:t}\top \wedge \texttt{After}_{j:vd:t}\neg hq{:}t).$$

In $\mathcal{DL}$:

$$\texttt{MisTrust}(i, j, vd{:}t, hq{:}t) \overset{\text{def}}{=} \texttt{Goal}_i\texttt{X}hq{:}t \wedge \texttt{Bel}_i(\texttt{Does}_{j:vd:t}\top \wedge \texttt{After}_{j:vd:t}\neg hq{:}t).$$

Example 4.5.1. Suppose that a moderator i has the goal of high quality in t. In $\mathcal{HHVL}$:

$$\texttt{Choice}_i\mathsf{F}hq{:}t,$$

and in $\mathcal{DL}$:

$$\texttt{Goal}_i\texttt{X}hq{:}t.$$

The moderator i also believes that a user j has submitted a post intended to vandalize thread t:

$$\texttt{Bel}_i\texttt{Does}_{j:vd:t}\top.$$

Since all moderators believe that every user has the opportunity to vandalize a thread t,

$$\texttt{Bel}_i\texttt{After}_{j:vd:t}\neg hq{:}t.$$

Hence, i mistrusts j:

$$\texttt{MisTrust}(i, j, vd{:}t, hq{:}t).$$

[2] Note that it does not need to be a goal of a troll to ensure low quality on the forum; for example, a troll's goal might just be to amuse herself. However, it is a consequence of the action of vandalizing a thread that the thread's quality drops.

4.6 Trust dispositions

The additional operators $\mathtt{G}$ and $\mathtt{F}$ allow for the concept of trust dispositions to be formally expressed. This is particularly useful in the Internet forum scenario.

If a user has made many good posts, and never attempted any trolling, moderators are inclined to trust that user with making new posts. With the definitions proposed above, moderators are forced to make new evaluations of users for every new post made, with the result that moderators can only consider a user trustworthy or not in relation to one particular post in one particular thread. Trust and mistrust dispositions allow for moderators to deem particular users trustworthy overall (within the domain of making posts in the forum).

Using the concepts of the C&F theory, in cases like the above, moderators can have trust dispositions towards users; they can make evaluations about the users' capability, willingness, and powers in relation to potential actions and/or potential goals. Such evaluations consist of beliefs that whenever a user j makes a post, the post will contribute to the goal of a high quality forum. Then, this potential evaluation forms the basis for inferring core trust in the relevant user j under the circumstances that j actually submits a post. This can be expressed in $\mathcal{HHVL}$ using the formal definition of dispositional trust.

A moderator i has dispositional trust in a user j in relation to the goal of high quality and the potential action of submitting a post in a thread t if i has the goal of high quality in t, and believes that always when j submits a post, the post will contribute to the high quality of t. In $\mathcal{HHVL}$, this translates as:

$$\begin{aligned}\mathtt{DispTrust}(i, j, mp{:}t, hq{:}t, \mathtt{Does}_{j:mp:t}\top) &\overset{\text{def}}{=} \mathtt{PotGoal}_i(hq{:}t, \mathtt{Does}_{j:mp:t}\top)\\ &\wedge \mathtt{Bel}_i\mathtt{G}^*(\mathtt{Does}_{j:mp:t}\top \wedge \mathtt{Choice}_i\mathtt{F}(hq{:}t) \rightarrow \mathtt{After}_{j:mp:t}hq{:}t).\end{aligned}$$

To express that a moderator i has dispositional trust in user j in relation to submitting posts in any thread, I propose the following definition:

Definition 4.6.1.

$$\mathtt{GenDispTrust}(i, j) \overset{\text{def}}{=} \bigwedge_{t \in T} \mathtt{DispTrust}(i, j, mp{:}t, hq{:}t, \mathtt{Does}_{j:mp:t}\top)$$

This definition expresses a type of generalized trust disposition; the trusting moderator takes the user to be generally trustworthy in relation to the goals of high quality in all threads and the actions of making posts in any thread. Note that this definition is specific for the Internet forum scenario; moderator i does not necessarily take user j to be trustworthy in *every* aspect. For example, i might take j to be trustworthy in relation to the actions of submitting posts on the forum, without thinking that she could trust j with her personal finances.

Example 4.6.1. Suppose that the moderator i takes the user j to be trustworthy:

$$\mathtt{GenDispTrust}(i, j).$$

Then, i has dispositional trust in j about j's opportunity to ensure high quality in a specific thread t_1 under the circumstances that j makes a post in t_1:

$$\mathtt{DispTrust}(i, j, mp{:}t_1, hq{:}t_1, \mathtt{Does}_{j:mp:t_1}\top).$$

Suppose further that i believes that j has submitted a post in thread t_1, and has the goal of high quality in t_1:

$$\mathtt{Bel}_i\mathtt{Does}_{j:mp:t_1}\top$$

and

$$\mathtt{Choice}_i\mathsf{F}\mathit{hq}{:}t_1.$$

Then, i trusts, "here-and-now", that j will ensure the goal of high quality in t_1 by the submission of a post in t_1:

$$\mathtt{OccTrust}(i, j, \mathit{mp}{:}t_1, \mathit{hq}{:}t_1).$$

Example 4.6.2. Suppose that a moderator i decides that she will trust a user j, who i believes has submitted a post, if two of i's moderator colleagues x and y, who have experience with posts submitted by j, have general dispositional trust in j:

$$\mathtt{Bel}_i(\mathtt{GenDispTrust}(x, j) \wedge \mathtt{GenDispTrust}(y, j) \to \mathtt{OccTrust}(i, j, \mathit{mp}{:}t, \mathit{hq}{:}t)).$$

Suppose now that i actually believes that x and y have general dispositional trust in j:

$$\mathtt{Bel}_i(\mathtt{GenDispTrust}(x, j) \wedge \mathtt{GenDispTrust}(y, j)).$$

Then, i believes that she trusts j:

$$\mathtt{Bel}_i\mathtt{OccTrust}(i, j, \mathit{mp}{:}t, \mathit{hq}{:}t),$$

and consequently, i trusts j:

$$\mathtt{OccTrust}(i, j, \mathit{mp}{:}t, \mathit{hq}{:}t).$$

4.7 Conclusion

It is clear from the above formalization of the Internet forum scenario that the time aspect is important; in particular, the modal operators G and F in $\mathcal{HHVL}$ enables reasoning about aspects of trust that is not expressible in $\mathcal{DL}$, for example combinations of goals and trust dispositions. Such dispositions are tantamount when modelling evaluations of users that are stable over time. In particular, if evaluations of users are always done "here-and-now", moderators are forced to re-evaluate users in relation to every individual post made by the users. This seems counter-intuitive, since, for instance, if a moderator believes a particular user always submits good posts, she is inclined to trust the user in relation to every post she submits. This is highlighted in Example 4.6.1.

Example 4.2.1 highlights some interesting properties of the logics $\mathcal{HHVL}$ and $\mathcal{DL}$ regarding the interaction of goals and aspects of time. In particular, the operators G and F in $\mathcal{HHVL}$ allows for a natural formalization of different kinds of goals. As seen in Example 4.2.1, treating all goals as being of the same kind leads to unintuitive consequences.

The above discussion shows that the adding of the operators G and F sophisticates the analysis of the Internet forum scenario in a useful way, even though it somewhat complicates the formalism. It should be noted that the additional operators $\mathtt{G}$ and $\mathtt{F}$ could possibly be added to $\mathcal{DL}$. Another important thing to note is that the next operator $\mathtt{X}$ cannot directly be defined in $\mathcal{HHVL}$. The reason for this is $\mathcal{HHVL}$'s lack of an axiom corresponding to $\mathcal{DL}$'s **Active**; in $\mathcal{DL}$, Axiom **Active** ensures that there always is a *next world* [29]. These things are, however, beyond the scope of this essay, which is why I have chosen the logic $\mathcal{HHVL}$ for my further analysis.

Chapter 5

A Horn fragment of $\mathcal{HHVL}$

In this chapter, a general Horn fragment for the logic $\mathcal{HHVL}$ is defined. A few further restrictions, intended to possibly avoid non-determinism, are also considered. I will then consider the Horn fragment of $\mathcal{HHVL}$ in relation to the Internet forum scenario.

5.1 Why Horn fragments?

A *fragment* of a logic S is a restriction put on S. Such a restriction can, for instance, consist in restricting which formulas will be regarded well-formed. The reason for studying fragments of multi-agent logics like $\mathcal{HHVL}$ is primarily due to the inherent *complexity* of such logics. For example, for many modal logics, checking whether any set of formulas is satisfiable requires unreasonable amounts of resources, making such logics difficult to use in practice [8]. As seen in the previous sections, a logic of trust needs to be able to differentiate between different kinds of mental attitudes, for example beliefs and goals, as well as enable reasoning about other agents' mental states. The case study in Chapter 4 also shows that the modelling of time is essential in practical applications.

Thus, there is a need to identify fragments of multi-agent logics that both have reasonable data complexity for practical uses, and at the same time are expressive enough to allow reasoning that is useful in practice.

A *Horn fragment* of a logic S is a particular restriction of S's syntax; in a Horn fragment of S only formulas of the form of *Horn clauses* (or, in some cases, as with the Horn fragment of $\mathcal{HHVL}$ defined below, *Horn formulas*) are allowed as legal formulas. Horn fragments are particularly useful, since they form the basis of logic programming and deductive databases, and in some logics, they considerably lower the data complexity of checking satisfiability and validity of sets of formulas.

Let me first define the Horn fragment of propositional logic in order to introduce the idea.

5.2 The Horn fragment of propositional logic

Consider the propositional logic *PL* with syntax and semantics defined in the usual way (see Appendix A).

The following definitions are needed for the definition of Horn clauses.

Formulas of the forms p or $\neg p$, where p is an atomic proposition, are called *literals*, and $a, b, c, ...$ are used to denote them. Literals containing a negation are called *negative* literals, and literals not containing a negation are called *positive*.

A *clause* in propositional logic is a disjunction of literals; i.e. a formula of the form

$$p_1 \vee p_2 \vee ... \vee p_k \vee \neg q_1 \vee \neg q_2 \vee ... \vee \neg q_l$$

where $p_1, ..., p_k, q_1, ..., q_l$ are atomic propositions and $k, l \geq 0$.

A *Horn clause* is a clause with at most one positive literal; i.e. Horn clauses are clauses on the form

$$\neg q_1 \vee ... \vee \neg q_l \vee p \equiv q_1 \wedge ... \wedge q_l \rightarrow p \quad \text{or} \quad \neg q_1 \vee ... \vee \neg q_l \equiv q_1 \wedge ... \wedge q_l \rightarrow \bot$$

where $p_1, ..., p_k, q_1, ..., q_l$ are atomic propositions and $k, l \geq 0$.

Horn clauses are often written as $p \leftarrow q_1, ..., q_l$, using commas instead of $\wedge$ and an implication arrow pointing to the left instead of to the right. This is because Horn clauses are often most naturally read as: to prove p, prove $q_1, ..., q_l$.

The *Horn fragment* of *PL* is the language resulting from allowing only formulas of the form of Horn clauses as well-formed formulas.

5.3 A Horn fragment of $\mathcal{HHVL}$

In this section, I define a Horn fragment of $\mathcal{HHVL}$ (here referred to as *Horn-$\mathcal{HHVL}$*), which is the fragment of $\mathcal{HHVL}$ containing only Horn-$\mathcal{HHVL}$ formulas. I then prove some properties of Horn-$\mathcal{HHVL}$. The definitions in the following sections, and the theorems and proofs in Section 5.3.1, are based on those given by Nguyen [37, 38].

For convenience, the universal modal operators in $\mathcal{HHVL}$ will be denoted in the following way:[1]

- $[B_i]\varphi$ denotes $\texttt{Bel}_i\varphi$;
- $[C_i]\varphi$ denotes $\texttt{Choice}_i\varphi$;
- $[A_{i:\alpha}]\varphi$ denotes $\texttt{After}_{i:\alpha}\varphi$;
- $[G]\varphi$ denotes $\texttt{G}\varphi$.

The operator $\texttt{Does}_{i:\alpha}$ is an existential modal operator, and will be denoted:

- $\langle D_{i:\alpha}\rangle\varphi$ denotes $\texttt{Does}_{i:\alpha}\varphi$.

The *duals* of the above operators are also taken into account. Recall that the dual of a universal modal operator is an existential modal operator, and the

[1] The same kind of notation is also used in [19].

dual of an existential modal operator is a universal modal operator. Thus, the following abbreviations are defined:

$$\begin{aligned}
\langle B_i\rangle\varphi &\overset{\text{def}}{=} \neg[B_i]\neg\varphi;\\
\langle C_i\rangle\varphi &\overset{\text{def}}{=} \neg[C_i]\neg\varphi;\\
\langle A_{i:\alpha}\rangle\varphi &\overset{\text{def}}{=} \neg[A_{i:\alpha}]\neg\varphi;\\
\langle G\rangle\varphi &\overset{\text{def}}{=} \neg[G]\neg\varphi;\\
[D_{i:\alpha}]\varphi &\overset{\text{def}}{=} \neg\langle D_{i:\alpha}\rangle\neg\varphi.
\end{aligned}$$

Note that some of the above duals were explicitly defined in Section 3.2. For example, the operator $\langle G\rangle$ is the operator $\mathtt{F}$, the operator $\langle B_i\rangle$ is $\mathtt{Poss}_i$, etc.

Let $\Sigma = \{B_i, C_i, G, A_{i:\alpha}, D_{i:\alpha} : i \in AGT, \alpha \in ACT\}$. For convenience, $[R]$ and $\langle R\rangle$, $R \in \Sigma$, denote universal operators and existential operators, respectively.

Formulas of the form p or $\neg p$, where $p \in ATM$, are called *classical literals*, and $a, b, c, ...$ are used to denote them.

Formulas of the form $p, [R]p$, or $\langle R\rangle p$, where $R \in \Sigma$ and $p \in ATM$, are called *atoms* and $A, B, C, ...$ are used to denote them.

Definition 5.3.1. A formula is in *negation normal form* if it does not contain the connective $\to$ and the connective $\neg$ can only occur immediately before an atomic proposition of ATM.

Every formula of $\mathcal{HHVL}$ can be translated into its negation normal form by applying the following equivalences:

$$\begin{aligned}
\varphi \leftrightarrow \psi &\equiv (\varphi \to \psi) \wedge (\psi \to \varphi);\\
\varphi \to \psi &\equiv \neg\varphi \vee \psi;\\
\neg\neg\varphi &\equiv \varphi;\\
\neg(\varphi \wedge \psi) &\equiv \neg\varphi \vee \neg\psi;\\
\neg(\varphi \vee \psi) &\equiv \neg\varphi \wedge \neg\psi;\\
\neg[R]\varphi &\equiv \langle R\rangle\neg\varphi;\\
\neg\langle R\rangle\varphi &\equiv [R]\neg\varphi,
\end{aligned}$$

where $R \in \Sigma$. The first five equivalences are standard tautologies of propositional logic. The rest of the rules, concerning the modal operators of $\mathcal{HHVL}$, holds because all operators are normal (see Section 3.1.1).

Note also the following equivalences, which can be applied to formulas where the constants $\top$ or $\bot$ occur:

$$\begin{aligned}
\varphi \wedge \top &\equiv \varphi;\\
\varphi \vee \top &\equiv \top;\\
\varphi \wedge \bot &\equiv \bot;\\
\varphi \vee \bot &\equiv \varphi.
\end{aligned}$$

Definition 5.3.2. A formula is *positive* if it is constructed from $\top$ and atomic propositions of ATM using $\wedge, \vee$, and the modal operators $[R]$ and $\langle R\rangle$, $R \in \Sigma$. A formula φ is *negative* if the negation normal form of $\neg\varphi$ is positive. If a formula is not positive, it is called *non-positive.*

Note that both positive and negative formulas are in negation normal form, and that positive formulas contain no $\neg$, while in negative formulas every occurrences of $\top$ and every atomic proposition of ATM are preceded by a negation.

Definition 5.3.3. A *modal context* (denoted $\boxdot$) is a possibly empty sequence of universal modal operators $[R]$, $R \in \Sigma$.

Note that a modal context can be distributed over conjunction. Suppose that $\varphi = \boxdot(\phi \wedge \psi)$ holds. Then, the rightmost operator $[R], R \in \Sigma$ occurring in $\boxdot$ can be distributed over the conjunction to, so that $\varphi \equiv \boxdot'([R]\phi \wedge [R]\psi)$, where $\boxdot = \boxdot'[R]$. Distribution can be applied again in a similar way for the rightmost operator in $\boxdot'$, etc., until the formula $\varphi \equiv \boxdot\phi \wedge \boxdot\phi'$ is reached.

Definition 5.3.4. A *clause* in $\mathcal{HHVL}$ is a formula of the form

$$\boxdot(A_1 \vee ... \vee A_n \vee \neg B_1 \vee ... \neg B_m)$$

where $m, n \geq 0$, $\boxdot$ is a modal context, and $A_1, ..., A_n, B_1, ..., B_m$ are atoms.

Definition 5.3.5. A formula φ is a *Horn-$\mathcal{HHVL}$ formula* (hereafter referred to simply as *Horn formula*) if it is of one of the following forms:

- $\varphi = \top$;
- φ is a proposition of ATM;
- φ is a negative formula;
- $\varphi = [R]\psi$ or $\varphi = \langle R\rangle\psi$ or $\varphi = \psi \wedge \zeta$, where $R \in \Sigma$ and ψ, ζ are Horn formulas;
- $\varphi = \psi \rightarrow \zeta$, where ψ is a positive formula and ζ is a Horn formula;
- $\varphi = \psi \vee \zeta$, where ψ is a negative formula and ζ is a Horn formula.

Definition 5.3.6. A clause is a *Horn-$\mathcal{HHVL}$ clause* (hereafter referred to as *Horn clause*) if it is a Horn formula.

Theorem 5.3.1. *Every Horn clause is of one of the forms*

$$\boxdot(\neg B_1 \vee ... \vee \neg B_m) \equiv \boxdot(B_1 \wedge ... \wedge B_m \rightarrow \bot) \quad \textit{or}$$
$$\boxdot(\neg B_1 \vee ... \vee \neg B_m \vee A) \equiv \boxdot(B_1 \wedge ... \wedge B_m \rightarrow A)$$

where $m \geq 0$, $\boxdot$ *is a modal context, and* $B_1, ..., B_m, A$ *are atoms.*

Proof. Since $[R]\varphi$ is a Horn formula if φ is a Horn formula for every $R \in \Sigma$, it suffices to prove that a clause with empty modal context and more than one positive atom is not a Horn formula.

First, all formulas $\neg B_i$ and A_j, $i, j \geq 0$ are Horn formulas. Every A_j is of the form p or $[R]p$ or $\langle R\rangle p$, where $R \in \Sigma$ and $p \in ATM$. Every formula $\neg B_i$ is of the form $\neg p$ or $\neg[R]p$ or $\neg\langle R\rangle p$, where $R \in \Sigma$ and $p \in ATM$. In the two latter cases formulas of the form $\neg[R]p$ are equivalent to $\langle R\rangle\neg p$ and formulas of the form $\neg\langle R\rangle p$ are equivalent to $[R]\neg p$, which by definition are Horn formulas. Further, all formulas $\neg B_i$ are negative formulas. The negative normal form of the clause $\neg B_1 \vee ... \vee \neg B_m$, $m \geq 0$ is a negative formula, hence a Horn formula.

The clause $\neg B_1 \vee ... \vee \neg B_m \vee A$, $m \geq 0$, is a disjunction of a negative formula (i.e. $\neg B_1 \vee ... \vee \neg B_m$) and a Horn formula (i.e. A), hence a Horn formula.

The theorem is proved by induction over the length of the disjunction of positive atoms occurring in a clause; it is shown that a clause $\neg B_1 \vee ... \vee \neg B_m \vee A_1 \vee ... \vee A_n$ is not a Horn formula for $n \geq 2$.

Base step. Consider a clause with empty modal context and with two positive atoms: $(\neg B_1 \vee ... \vee \neg B_m \vee A_1 \vee A_2)$. The negative normal form of $\varphi = \neg B_1 \vee ... \vee \neg B_m$ is a negative formula. Since A_1 is a Horn formula, $\psi = \varphi \vee A_1$ is a Horn formula. It is, however, not a negative formula, since the atomic proposition occurring in A_1 is not prefixed by $\neg$. This means that $\psi \vee A_2$ is not a Horn formula.

Induction step. Assume that $\zeta = \neg B_1 \vee ... \vee \neg B_m \vee A_1 \vee ... \vee A_k$, $k \geq 2$ is not a Horn formula. Then $\zeta' = \zeta \vee A_{n+1}$ is not a Horn formula, since ζ is not a negative formula.

Thus, the theorem is proved by induction on the length of the disjunction of positive atoms occurring in a clause. □

Definition 5.3.7. A *positive program* is a set of Horn clauses of the form

$$\boxdot(B_1 \wedge ... \wedge B_m \rightarrow A)$$

where $\boxdot$ is a modal context, $m \geq 0$, and $B_1, ..., B_m, A$ are atoms. If $m > 0$, the clause is called a *program clause.* If $m = 0$, the resulting clause $\rightarrow A$ is called a *fact.*

Clauses in positive programs are often written using a reversed implication symbol and commas instead of $\wedge$, i.e. $\boxdot(A \leftarrow B_1, ..., B_m)$. The expression $P \vDash \varphi$ is used to denote that φ is a logical consequence of the positive program P.[2]

5.3.1 The expressiveness of the language of Horn clauses in $\mathcal{HHVL}$

In this section, I prove that the language of Horn clauses in $\mathcal{HHVL}$ is as expressive as the language of Horn formulas in $\mathcal{HHVL}$. The proof of the theorem stating this property is essentially the same as the proof given in Nguyen [37].

Definition 5.3.8. A set $X = \{\phi_1, ..., \phi_k\}$ of formulas in $\mathcal{HHVL}$ is *satisfiable* if there is a model M in which all formulas in X are satisfiable. In other words, a set X of formulas is satisfiable if all formulas in X can be true at the same time.

It will be useful to regard a set of formulas as a conjunction of the formulas in the set, i.e. $\{\phi_1, ..., \phi_k\}$ as $\phi_1 \wedge ... \wedge \phi_k$. This is fine, since by the definition of satisfiability of conjunction, a formula $\varphi \wedge \psi$ is satisfiable if and only if φ is satisfiable and ψ is satisfiable.

Definition 5.3.9. Two sets of formulas X and Y are *equisatisfiable* in $\mathcal{HHVL}$ if (X is satisfiable if and only if Y is satisfiable).

[2] If the logic S in question is not clear from context, $P \vDash_S \varphi$ is used to express that φ is a consequence of the program P in the logic S.

Note that equisatisfiability is different from logical equivalence. Two equisatisfiable formulas need not be equivalent, since they might be satisfiable in different models. Equisatisfiability is useful when checking whether a set of formulas X is satisfiable—if X can be shown to be equisatisfiable to a set Y of formulas, checking whether X is satisfiable amounts to checking whether Y is satisfiable.

The following theorem is a variant of a similar theorem concerning monomodal normal logics proven by Nguyen [37, pp. 35–36] (see also Mints [35]). The proof is essentially the same as that of Nguyen [37], since all operators $[R]$ (and their duals $\langle R\rangle$), $R \in \Sigma$, are normal modal operators.

Theorem 5.3.2. *For any set X of Horn formulas in $\mathcal{HHVL}$, there exists a set Y of Horn clauses such that X and Y are equisatisfiable.*

In order to prove this theorem, the following lemma is needed [37]. For the sake of convenience, the following notation is introduced:

- $[\phi_1, ..., \phi_k]$ denote the disjunction $\phi_1 \vee ... \vee \phi_k$;
- $\phi_1; ...; \phi_k$ denote the set $\{\phi_1, ..., \phi_k\}$;
- if X and Y are sets of formulas, $X; Y$ are used to denote $X \cup Y$;
- $X; \varphi$ denotes $X \cup \{\varphi\}$.

Thus, $X \cup \{\boxdot(\psi \vee \zeta \vee \xi)\}$ is denoted $X; \boxdot[\psi, \zeta, \xi]$, etc.

Lemma 5.3.1. *Let p and q be fresh propositions, which means that they only occur where indicated. Then, for any $R \in \Sigma$, the following pairs of sets of formulas are equisatisfiable in $\mathcal{HHVL}$:*

(a) $X; \boxdot[\psi, \zeta \vee \xi]$ *and* $X; \boxdot[\psi, \zeta, \xi]$;

(b) $X; \boxdot[\psi, \zeta \wedge \xi]$ *and* $X; \boxdot[\psi, \neg p]; \boxdot[p, \zeta]; \boxdot[p, \xi]$;

(c) $X; \boxdot[\psi, \zeta \wedge \xi]$ *and* $X; \boxdot[\psi, q]; \boxdot[\neg q, \zeta]; \boxdot[\neg q; \xi]$;

(d) $X; \boxdot[\phi, [R]\psi]$ *and* $X; \boxdot[\phi, [R]p]; \boxdot[R][\neg p, \psi]$;

(e) $X; \boxdot[\phi, [R]\psi]$ *and* $X; \boxdot[\phi, [R]\neg q]; \boxdot[R][q, \psi]$;

(f) $X; \boxdot[\phi, \langle R\rangle\psi]$ *and* $X; \boxdot[\phi, \langle R\rangle p]; \boxdot[R][\neg p, \psi]$;

(g) $X; \boxdot[\phi, \langle R\rangle\psi]$ *and* $X; \boxdot[\phi, \langle R\rangle\neg q]; \boxdot[R][q, \psi]$.

Proof. What needs to be shown to prove the theorem is that, for every pair of sets of formulas (a)–(g), if the left hand side (LHS) is satisfiable, then the right hand side (RHS) is satisfiable, and if the RHS is satisfiable, then the LHS is satisfiable.

First, the implication from left to right is proved (if the LHS is satisfiable, then the RHS is satisfiable).

Fix one of the pairs. Suppose that the LHS of the pair is satisfied in a model $M = \langle F, V\rangle$, where $F = \langle W, A, B, C, D, G\rangle$ is a frame. Let $M' = \langle F, V'\rangle$ be a model such that, where p and q are fresh atomic propositions,

- for $x \neq p$ and $x \neq q$, $x \in V'(w)$ if and only if $x \in V(w)$,

- $p \in V'(w)$ if and only if $M, w \vDash \psi$,
- $q \in V'(w)$ if and only if $M, w \vDash \neg\psi$.

It is easily seen that the right hand side of the fixed pair is satisfied in M'. Here is a sketch of the reasoning involved for pair (b). If $M, w \vDash \psi \vee (\zeta \wedge \xi)$, then $M, w \vDash \psi \vee \zeta$ and $M, w \vDash \psi \vee \xi$. If $M, w \vDash \psi$, then the left hand side trivially holds, since $M', w \vDash \psi$ and $M', w \vDash p$ hold. If $M, w \nvDash \psi$ holds, then $M', w \vDash \zeta$, $M', w \vDash \xi$, and $M', w \vDash \neg p$ hold.

Second, the implication from right to left is proved (if the RHS is satisfiable, then the LHS is satisfiable).

Pair (a): This is straightforward, since disjunction is associative; $\psi \vee \zeta \vee \xi$ implies $\psi \vee (\zeta \vee \xi)$. By axioms **Nec** and **K** for all $R \in \Sigma$, $\boxdot(\psi \vee \zeta \vee \xi) \rightarrow \boxdot(\psi \vee (\zeta \vee \xi))$ is a theorem.

Pair (b): It is a tautology of propositional logic that $(\psi \vee \neg p) \wedge (p \vee \zeta) \wedge (p \vee \xi) \rightarrow \psi \vee (\zeta \wedge \xi)$. Since all modal operators in $\mathcal{HHVL}$ are normal, $\boxdot((\psi \vee \neg p) \wedge (p \vee \zeta) \wedge (p \vee \xi) \rightarrow \psi \vee (\zeta \wedge \xi))$ is a theorem, which (by the **Nec** and **K** axioms and distribution of $[R]$ over conjunction for all $R \in \Sigma$) makes $\boxdot(\psi \vee \neg p) \wedge \boxdot(p \vee \zeta) \wedge \boxdot(p \vee \xi) \rightarrow \boxdot(\psi \vee (\zeta \wedge \xi))$ a tautology.

Pair (c): $\boxdot(\psi \vee q) \wedge \boxdot(\neg q \vee \zeta) \wedge \boxdot(\neg q \vee \xi) \rightarrow \boxdot(\psi \vee (\zeta \wedge \xi))$ is a tautology by similar reasoning as for pair (b).

Pair (d): $(\phi \vee [R]p) \wedge [R](\neg p \vee \psi) \rightarrow (\phi \vee [R]\psi)$ is a tautology. To see why, assume the opposite, i.e. $(\phi \vee [R]p) \wedge [R](\neg p \vee \psi) \wedge \neg\phi \wedge \neg[R]\psi$ which by standard equivalences of propositional logic and distribution of $[R]$, $R \in \Sigma$ over conjunction is equivalent to $(\phi \vee [R]p) \wedge ([R]p \rightarrow [R]\psi) \wedge \neg\phi \wedge \neg[R]\psi$. This, in turn, leads to a contradiction, since $[R]p$ must hold, and hence also $[R]\psi \wedge \neg[R]\psi$. Since $(\phi \vee [R]p) \wedge [R](\neg p \vee \psi) \rightarrow \phi \vee [R]\psi$ is a tautology, $\boxdot(\phi \vee [R]p) \wedge \boxdot[R](\neg p \vee \psi) \rightarrow \boxdot(\phi \vee [R]\psi)$ is a tautology by Axioms **Nec** and **K** for $R \in \Sigma$.

Pair (e): $\boxdot(\phi \vee [R]\neg q) \wedge \boxdot[R](q \vee \psi) \rightarrow \boxdot(\phi \vee [R]\psi)$ is a tautology by similar reasoning as for pair (d).

Pair (f): Assume the opposite of $(\phi \vee \langle R\rangle p) \wedge [R](\neg p \vee \psi) \rightarrow (\phi \vee \langle R\rangle\psi)$, i.e. $(\phi \vee \langle R\rangle p) \wedge [R](\neg p \vee \psi) \wedge \neg\phi \wedge \neg\langle R\rangle\psi$ which, by standard properties of propositional logic and the normal operators $[R], R \in \Sigma$, is equivalent to $(\phi \vee \langle R\rangle p) \wedge \neg\langle R\rangle p \wedge \neg\langle R\rangle\neg\psi \wedge \neg\phi \wedge \neg\langle R\rangle\psi$. This formula implies $\langle R\rangle p \wedge \neg\langle R\rangle p$, which is a contradiction. Hence $(\phi \vee \langle R\rangle p) \wedge [R](\neg p \vee \psi) \rightarrow (\phi \vee \langle R\rangle\psi)$ is a tautology. By Axioms **Nec** and **K**, for $R \in \Sigma$, $\boxdot(\phi \vee \langle R\rangle p) \wedge \boxdot[R](\neg p \vee \psi) \rightarrow \boxdot(\phi \vee \langle R\rangle\psi)$ is a tautology.

Pair (g): $\boxdot(\phi \vee \langle R\rangle\neg q) \wedge \boxdot[R](q \vee \psi) \rightarrow \boxdot(\phi \vee \langle R\rangle\psi)$ is a tautology by similar reasoning as for pair (f).

□

With Lemma 5.3.1 in place, Theorem 5.3.2 can be proved.

Proof. Let X be a set of Horn-$\mathcal{HHVL}$ formulas. Translation of all formulas in X into negative normal form yields a new set of formulas X'. Then, the pairs of equisatisfiable sets of formulas in Lemma 5.3.1 are used as translation rules from left to right.

Since all formulas in X' are in negative normal form, every formula in X' is of the forms $\boxdot(\varphi \wedge \varphi')$, or of one of the forms represented by the left hand side of the pairs (a)–(g).

If a formula is of the form $\boxdot(\varphi \wedge \varphi')$, it can be regarded as the set of formulas $\{\boxdot\varphi, \boxdot\varphi'\}$.

The translation rules (a)–(g) are applied to X' until no more changes can be made to X'. The rules (d)–(g) are only used when ψ is not a classical literal, and if both (b) and (c), both (d) and (e), or both (f) and (g) are applicable, then the appropriate one must be chosen so that the resulting set contains only Horn formulas.

The resulting set Y is a set containing only Horn clauses.

□

5.3.2 Possible further restrictions

In order to avoid undesirable properties, some further restrictions of Horn-$\mathcal{HHVL}$ might be warranted.

Consider the following program P in Horn-$\mathcal{HHVL}$ (a similar example can be found in [39]):

$$\begin{aligned}
&[G]p \leftarrow;\\
&q \leftarrow \langle G\rangle p;\\
&s \leftarrow [G]r.
\end{aligned}$$

If there is an accessible world from the world of evaluation, then $\langle G\rangle p$ is true, and hence also q. If there is not an accessible world, then $[G]r$ is true, and hence also s. Thus, it is the case that $P \vDash q \vee s$ but $P \nvDash q$ and $P \nvDash s$. The problem is that if there is no accessible world, then $[G]r$ might *undesirably* become true, since r is trivially true at every accessible world (since there are no accessible worlds), even though $[G]r$ does not follow from the program. This is because the operator $[G]$ is not *serial*.[3]

The operators $[B_i]$ and $[C_i]$, $i \in AGT$, satisfy the principles

$$[B_i]\varphi \rightarrow \langle B_i\rangle\varphi \quad \text{and} \quad [C_i]\varphi \rightarrow \langle C_i\rangle\varphi$$

which correspond to the semantic constraint placed on frames known as *seriality*. Recall that a relation R is serial if, for every $w \in W$ there is a $v \in W$ such that $R(w, v)$. The operators (their corresponding accessibility relation) $[G]$, $[A_{i:\alpha}]$ and $[D_{i:\alpha}]$ do not have this property.

To avoid this kind of "non-determinism", one can restrict Horn-$\mathcal{HHVL}$ by allowing $[G]$, $[A_{i:\alpha}]$, and $[D_{i:\alpha}]$ on the left hand side of implications only in conjunction with their existential duals.

[3] When I say that an operator is serial, I mean that the accessibility relation corresponding to the operator is serial. Note also that if the seriality property applies to an operator, it also applies to the operator's dual, since dual operators share the same accessibility relation.

The following operators are defined:

$$\begin{aligned}[G]_s\varphi &\overset{\text{def}}{=} [G]\varphi \wedge \langle G\rangle\varphi;\\ [A_{i:\alpha}]_s\varphi &\overset{\text{def}}{=} [A_{i:\alpha}]\varphi \wedge \langle A_{i:\alpha}\rangle\varphi;\\ [D_{i:\alpha}]_s\varphi &\overset{\text{def}}{=} [D_{i:\alpha}]\varphi \wedge \langle D_{i:\alpha}\rangle\varphi.\end{aligned}$$

Horn-$\mathcal{HHVL}$ can then be restricted by disallowing $[G], [A_{i:\alpha}]$, and $[D_{i:\alpha}]$, and instead use $[G]_s, [A_{i:\alpha}]_s$ and $[D_{i:\alpha}]_s$ on the left hand side of implications. Note that $[G], [A_{i:\alpha}]$, and $[D_{i:\alpha}]$ are still allowed in on the right hand side of implications, and in formulas where no implications occur. The legal occurrences of the existential operators $\langle G\rangle$, $\langle A_{i:\alpha}\rangle$, and $\langle D_{i:\alpha}\rangle$ are not restricted in any way.

For the operator $[G]$, it may not be that problematic to adopt seriality, i.e. to use $[G]_s$ on the left hand side of implications. This is because it is quite natural to assume that if φ holds at all future times, then there is a future time where φ holds.

However, it is not always desirable to adopt seriality for the dynamic operators $[A_{i:\alpha}]$ and $[D_{i:\alpha}]$ (and their duals), since, for example,

$$\begin{aligned}[D_{i:\alpha}]\varphi \to \langle D_{i:\alpha}\rangle\varphi &\equiv \neg[D_{i:\alpha}]\varphi \vee \langle D_{i:\alpha}\rangle\varphi\\ &\equiv \langle D_{i:\alpha}\rangle\neg\varphi \vee \langle D_{i:\alpha}\rangle\varphi\\ &\equiv \langle D_{i:\alpha}\rangle(\neg\varphi \vee \varphi)\\ &\equiv \langle D_{i:\alpha}\rangle\top.\end{aligned}$$

Thus, imposing seriality on the operator $[D_{i:\alpha}]$ amounts to stating that every agent $i \in AGT$ always performs an action $\alpha \in ACT$.

5.4 Horn formulas and the Internet forum scenario

In this section, I consider the Internet forum scenario expressed using Horn-$\mathcal{HHVL}$.

5.4.1 Trust, distrust, and mistrust

Let $mp{:}t \in ACT$, $hq{:}t \in ATM$, and $t \in T$. First, note that

$$[C_i]\langle G\rangle hq{:}t \quad \text{and} \quad [B_i]\langle D_{j:mp:t}\rangle\top$$

are Horn formulas.

A moderator j's trust in a user i about the action of making a post and the goal of high quality is expressed by the formula

$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:mp:t}\rangle\top \wedge [B_i][A_{j:mp:t}]hq{:}t.$$

This formula is a Horn formula, since it is a conjunction of known Horn formulas, and $[B_i][A_{j:mp:t}]hq{:}t$, which also is a Horn formula.

A moderator i's distrust in a user j is expressed by the formula

$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:mp:t}\rangle\top \wedge [B_i]\neg[A_{j:mp:t}]hq{:}t.$$

This is equivalent to a Horn formula, since it is a conjunction of Horn formulas, as seen above, and the formula $[B_i]\neg[A_{j:mp:t}]hq{:}t$, which is a Horn formula, since $\neg[A_{j:\alpha}]hq{:}t \equiv \langle A_{j:\alpha}\rangle\neg hq{:}t$ is a Horn formula.

A moderator i's mistrust in a user j is expressed by the formula

$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:vd:t}\rangle\top \wedge [B_i][A_{j:vd:t}]\neg hq{:}t.$$

This formula is a Horn formula, since $[B_i]\langle D_{j:vd:t}\rangle\top$ and $[B_i][A_{j:vd:t}]\neg hq{:}t$ are easily seen to be Horn formulas.

5.4.2 Trust dispositions

A moderator i's potential goal of high quality in thread t under the circumstances that a user j submits a post is expressed by the formula

$$\begin{aligned}&\texttt{PotGoal}_i(hq{:}t, \langle D_{j:mp:t}\rangle) \overset{\text{def}}{=} \\ &\qquad \langle B_i\rangle(\langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t \wedge \langle G\rangle(\langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t)),\end{aligned}$$

which is clearly a Horn formula, since it is a conjunction of Horn formulas preceded by $\langle B_i\rangle$.

A moderator's dispositional trust in a user j in relation to the potential goal of high quality in t under the circumstances that j submits a post, and j's action of actually submitting a post, is defined as the conjunction of a potential goal and the formula

$$\begin{aligned}&[B_i](\langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t \to [A_{j:mp:t}]hq{:}t) \\ &\qquad\qquad \wedge [B_i][G](\langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t \to [A_{j:mp:t}]hq{:}t).\end{aligned}$$

The above formula is Horn, since

$$\varphi = \langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t \to [A_{j:mp:t}]hq{:}t$$

is a Horn formula (the left hand side of the implication is a positive formula, and the right hand side is a Horn formula), which makes $[B_i]\varphi$ and $[B_i][G]\varphi$ Horn formulas.

5.4.3 Inferring trust

Let, for all $i \in M, j \in U, t \in T$:

- $trust_{i,j,t} \in ATM$ denote moderator i's core trust in user j about ensuring high quality in thread t by submitting a post in t;
- $distrust_{i,j,t} \in ATM$ denote moderator i's distrust in user j about ensuring high quality in thread t by submitting a post in t;
- $mistrust_{i,j,t} \in ATM$ denote moderator i's mistrust in user j in relation to the goal of high quality in t and the action of vandalizing t;
- $disptrust_{i,j,t} \in ATM$ denote moderator i's dispositional trust that user j will contribute to the high quality in t under the circumstances that j actually submits a post.

Trust are inferred according to the rule:

$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:mp:t}\rangle\top \wedge [B_i][A_{j:mp:t}]hq{:}t \to \mathit{trust}_{i,j,t}.$$

This is a Horn formula, since the right hand side is a proposition of ATM and the left hand side is a positive formula.

However, since $[B_i]\neg[A_{j:mp:t}]hq{:}t \equiv [B_i]\langle A_{j:mp:t}\rangle\neg hq{:}t$ and $[B_i][A_{j:vd:t}]\neg hq{:}t$ are not positive formulas, the inference rules

$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:mp:t}\rangle\top \wedge [B_i]\neg[A_{j:mp:t}]hq{:}t \to \mathit{distrust}_{i,j,t}$$

and

$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:vd:t}\rangle\top \wedge [B_i][A_{j:vd:t}]\neg hq{:}t \to \mathit{mistrust}_{i,j,t}$$

are not Horn formulas.

In the case of mistrust, the second approach discussed in Section 4.5, consisting in defining mistrust as a conjunction of the moderator's goal of high quality in t and the moderator's belief that the user deliberately has submitted a vandalizing post in t while ignoring the consequences of the particular action of vandalizing, can be used to express the following rule:

$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:vd:t}\rangle\top \to \mathit{mistrust}_{i,j,t}.$$

This formula is easily seen to be a Horn formula, since the left hand side consists of a conjunction of positive formulas.

However, in the case of distrust in the context of the Internet forum scenario, distrust is clearly based on the belief about the user's lack of opportunity/power to ensure high quality.

Inference of trust dispositions cannot be expressed as a Horn formula either. To see why, note that the Horn formula

$$\begin{aligned}&[B_i](\langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t \to [A_{j:mp:t}]hq{:}t)\\&\qquad\qquad \wedge [B_i][G](\langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t \to [A_{j:mp:t}]hq{:}t)\end{aligned}$$

is equivalent to

$$\begin{aligned}&[B_i](\neg\langle D_{j:mp:t}\rangle\top \vee \neg[C_i]\langle G\rangle hq{:}t \vee [A_{j:mp:t}]hq{:}t)\\&\qquad\qquad \wedge [B_i][G](\neg\langle D_{j:mp:t}\rangle\top \vee \neg[C_i]\langle G\rangle hq{:}t \vee [A_{j:mp:t}]hq{:}t),\end{aligned}$$

which is not a positive formula. Hence

$$\begin{aligned}&([B_i](\langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t \to [A_{j:mp:t}]hq{:}t)\\&\qquad \wedge [B_i][G](\langle D_{j:mp:t}\rangle\top \wedge [C_i]\langle G\rangle hq{:}t \to [A_{j:mp:t}]hq{:}t)) \to \mathit{disptrust}_{i,j,t}\end{aligned}$$

is not a Horn formula.

This means that the Internet forum scenario cannot be fully expressed using Horn-$\mathcal{HHVL}$.

5.4.4 Examples

In this section, I will express some of the examples from Section 4 in Horn-$\mathcal{HHVL}$.

Example 5.4.1. Consider Example 4.3.1 again. This example can be expressed by a set X containing the Horn formulas:

$$[B_i](trust_{x,j,t} \to trust_{i,j,t});$$
$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:mp:t}\rangle\top \wedge [B_i][A_{j:mp:t}]hq{:}t \to trust_{i,j,t};$$
$$[B_i]trust_{i,j,t} \to trust_{i,j,t};$$
$$[B_i][C_x]\langle G\rangle hq{:}t;$$
$$[B_i][B_x][A_{j:mp:t}]hq{:}t;$$
$$[B_i][B_x]\langle D_{j:mp:t}\rangle\top.$$

In this case, it is clear that $X \vDash trust_{i,j,t}$.

Example 5.4.2. Consider Example 4.4.1. This example can be formalized in Horn-$\mathcal{HHVL}$ as the set X containing the Horn formulas

$$[B_i](distrust_{x,j,t} \vee distrust_{y,j,t} \vee distrust_{z,j,t} \to distrust_{i,j,t});$$
$$[B_i]distrust_{i,j,t} \to distrust_{i,j,t};$$
$$[B_i]distrust_{x,j,t},$$

which logically implies $distrust_{i,j,t}$.

Example 5.4.3. Consider Example 4.5.1, and consider the set X containing the following Horn formulas:

$$[C_i]\langle G\rangle hq{:}t \wedge [B_i]\langle D_{j:vd:t}\rangle\top \to mistrust_{i,j,t};$$
$$[B_i]\langle D_{j:vd:t}\rangle\top;$$
$$[C_i]\langle G\rangle hq{:}t.$$

It is clear that $X \vDash mistrust_{i,j,t}$.

Example 5.4.4. This example is a variant of Example 4.6.2. Consider the set of Horn formulas X:

$$[B_i](disptrust_{x,j,t} \wedge disptrust_{y,j,t} \to trust_{i,j,t});$$
$$[B_i]trust_{i,j,t} \to trust_{i,j,t};$$
$$[B_i](disptrust_{x,j,t} \wedge disptrust_{y,j,t}),$$

from which the consequence $trust_{i,j,t}$ follows logically.

Chapter 6

Summary and conclusions

In this thesis, the three different logics $\mathcal{HHVL}$, $\mathcal{DL}$, and $\mathcal{BNL}$, all intended to formalize the C&F theory of trust, were presented and evaluated along two lines.

First, key concepts of the C&F theory were formally defined, and some interesting properties resulting from these definitions were proved. I proposed new formal definitions for the concepts of mistrust, lack of trust, and dispositional mistrust. The proven properties were then compared to properties of the informal formulation of the C&F theory, resulting in the conclusion that the logics $\mathcal{HHVL}$ and $\mathcal{DL}$ were best suited for formalization of the C&F theory.

Second, the performances of the logics $\mathcal{HHVL}$ and $\mathcal{DL}$ were compared within a case study consisting of trust assessments of users of an Internet forum. It was concluded that $\mathcal{HHVL}$ allowed a more sophisticated analysis.

Then, a Horn fragment of $\mathcal{HHVL}$ was defined. The Horn fragment was shown to be too restrictive to express the Internet forum scenario.

It is clear that a trust logic intended to formalize the C&F theory of trust in order to reason about trust in intentional entities must be able to formalize *intentional action*. The C&F theory states that the truster's evaluation of the trustee's intention is equally important as the evaluation of the trustee's capability and opportunity for predicting the trustee's behaviour. The logic $\mathcal{BNL}$ does not capture this property, primarily because there are no interaction axioms governing the connection between intention and action. Further, in $\mathcal{BNL}$, the operators $\mathtt{After}_{i:\alpha}$ and $\mathtt{Happens}_{i:\alpha}$ are duals, which allows for some unwanted properties. For instance, capability implies action:

$$\vdash_{\mathcal{BNL}} \neg\mathtt{After}_{i:\alpha}\bot \rightarrow \mathtt{Happens}_{i:\alpha}\top.$$

This problem is avoided in $\mathcal{HHVL}$ and $\mathcal{DL}$ by the use of two different accessibility relations for the action operators $\mathtt{After}_{i:\alpha}$ and $\mathtt{Does}_{i:\alpha}$, which allows for the formula $\mathtt{After}_{i:\alpha}\varphi \leftrightarrow \neg\mathtt{Does}_{i:\alpha}\neg\varphi$ to be falsifiable.

Another critical point when formalizing the C&F theory is the time aspect, and in particular the interaction of goals and time. The logics $\mathcal{HHVL}$ and $\mathcal{DL}$ both model time in a linear temporal logic fashion, but while $\mathcal{HHVL}$ incorporates the *always* and *eventually* operators $\mathtt{G}$ and $\mathtt{F}$ as basic, $\mathcal{DL}$ uses the *next time* operator $\mathtt{X}$, defined in terms of actions. This leads to different modelling of agents' goals.

Regarding the Horn fragment of $\mathcal{HHVL}$, the analysis in Section 5.4 shows that inference of $\texttt{OccTrust}(i, j, \alpha, \varphi)$ (corresponding to core trust in the C&F theory) can be expressed by a Horn formula, but that inference of mistrust, distrust, and trust dispositions cannot be expressed by Horn formulas.

6.1 Ideas for further research

The following are ideas for further inquiries in trust logics and their Horn fragments.

- How could the rest of the basic concepts of the C&F theory be formally defined in a MAS oriented modal logic? In this thesis, I considered the formal translations of the concepts of core trust, mistrust, distrust, lack of trust, and trust dispositions. In order to reason about the connection between trust, reliance, and delegation, these concepts need to be formally defined.
- What is the data complexity for checking satisfiability for the logic $\mathcal{HHVL}$ and the Horn fragment of $\mathcal{HHVL}$ defined in Section 5? Are further restrictions needed to reach a tractable fragment?
- Another line of inquiry is to further investigate how to allow negations on the left hand side of implications (see the analysis in Section 5.4), in order to express inference of distrust, mistrust, and dispositional trust.
- It would also be interesting to investigate the construction of least models for Horn-$\mathcal{HHVL}$, as well as study the possibility of providing Horn-$\mathcal{HHVL}$ with a proof calculus. In that case, further restrictions along the lines proposed in Section 5.3.2 are probably needed (see e.g. [37, 39]).

Appendix A

Propositional logic

This appendix is a short introduction to propositional logic. The material is based on [2]. Other good introductory texts on propositional logic are Asratian, Björn, and Turesson's [1], Mårtensson's [32], and Prawitz's [40].

Propositional logic studies *propositions*. A proposition is a sentence that is either true or false. For example, "It is raining in Linköping" is a proposition, but "How many tigers are there in Alaska?" is not a proposition, since it does not have a truth value.

Consider the two propositions "It is raining in Linköping" and "2+2=4". These propositions are either true or false. The two propositions can be combined in different ways, yielding new sentences which are either true or false. Consider the following combination: "It is raining in Linköping and 2+2=4". This sentence is true when "It is raining in Linköping" and "2+2=4" are both true. The word "and" is an example of a *connective*. Other such connectives in common English are "or", "if ..., then" and "not". The purpose of propositional logic is to formalize the way in which words like "and" and "not" combines propositions, and how such combinations enable formal reasoning from premises to a conclusion.

A.1 Syntax

The language of propositional logic (denoted PL) consists of a nonempty set of *atomic propositions* $ATM = \{p, q, ...\}$, and the *logical connectives* presented in Table A.1.

An atomic proposition is a proposition without an internal logical structure; in other words, it is a proposition without any logical connectives.

Name	Symbol	Meaning
negation	$\neg$	not
conjunction	$\wedge$	and
disjunction	$\vee$	or
implication	$\rightarrow$	if ... , then ...
equivalence	$\leftrightarrow$	if and only if

Table A.1: The logical connectives in propositional logic.

All connectives except $\neg$ are *binary*, which means that they connect two formulas. The connective $\neg$ is applied to a single formula. The following definition gives the rules of how formulas are constructed.

Definition A.1.1. A *well-formed formula* (hereafter denoted simply "formula") is constructed from the elements of ATM and the logical connectives from Table A.1 according to the following rules, where $p \in ATM$:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \rightarrow \varphi \mid \varphi \leftrightarrow \varphi.$$

The above expression means that all atomic propositions are formulas, a formula preceded by $\neg$ is a formula, and two formulas connected by $\wedge, \vee, \rightarrow$, or $\leftrightarrow$ is a formula.

In the meta-language, $\varphi, \psi, ...$ are used to denote formulas. This means that $\varphi, \psi, ...$ are not symbols in the language of PL, but convenient symbols used as abbreviations for formulas. For example, the formula $(p \vee q) \wedge (r \vee s)$ can be abbreviated $\varphi \wedge \psi$, where $\varphi = (p \vee q)$ and $\psi = (r \vee s)$.

Parentheses "(" and ")" are used to separate subformulas in formulas. In addition, *precedence* and *associativity* conventions for formulas are defined, just like in arithmetic; for example, one recognizes $a \cdot b \cdot c + d \cdot e$ as $(((a \cdot b) \cdot c) + (d \cdot e))$. The connectives in propositional logic have the following order of precedence:

1. $\neg$;

2. $\wedge$ and $\vee$;

3. $\rightarrow$ and $\leftrightarrow$.

Additional parentheses can be used to clarify formulas, even when it is not strictly necessary. Further, the connectives $\wedge$, $\vee$, and $\leftrightarrow$ are associative, so it is possible to omit parentheses in formulas that have repeated occurrences of these connectives. For example, instead of writing $((p \vee q) \vee r)$, one can write $p \vee q \vee r$. The connective $\rightarrow$ is not associative, so parentheses must be used when implication occur repeatedly in a formula.

A.2 Semantics

The semantics of PL defines the meaning of formulas in PL.

Definition A.2.1. Let φ be a formula, and let P_φ be the set of all atomic propositions appearing in φ. An *interpretation* for a formula φ is a full mapping $f_\varphi : P_\varphi \rightarrow \{T, F\}$ assigning a truth value *true* (denoted T) or *false* (denoted F) to each proposition in P_φ.

Definition A.2.2. Truth values of a formula φ under an interpretation f_φ, denoted $v_{f_\varphi}(\varphi)$, (whenever the formula φ is clear from context, $v_{f_\varphi}(\varphi)$ will be abbreviated by $v_f(\varphi)$), are defined as follows:

- $v_f(\varphi) = f(\varphi)$ if φ is an atomic proposition;

- $v_f(\neg\varphi) = T$ if $v_f(\varphi) = F$,
  $v_f(\neg\varphi) = F$ if $v_f(\varphi) = T$;

p	$\neg p$
T	F
F	T

p	q	$p \wedge q$	$p \vee q$	$p \rightarrow q$	$p \leftrightarrow q$
T	T	T	T	T	T
T	F	F	T	F	F
F	T	F	T	T	F
F	F	F	F	T	T

Table A.2: The truth table for the logical connectives

- $v_f(\varphi \wedge \psi) = T$ if $v_f(\varphi) = T$ and $v_f(\psi) = T$,
  $v_f(\varphi \wedge \psi) = F$ otherwise;
- $v_f(\varphi \vee \psi) = F$ if $v_f(\varphi) = F$ and $v_f(\psi) = F$,
  $v_f(\varphi \vee \psi) = T$ otherwise;
- $v_f(\varphi \rightarrow \psi) = F$ if $v_f(\varphi) = T$ and $v_f(\psi) = F$,
  $v_f(\varphi \vee \psi) = T$ otherwise;
- $v_f(\varphi \leftrightarrow \psi) = T$ if $v_f(\varphi) = v_f(\psi)$,
  $v_f(\varphi \leftrightarrow \psi) = F$ if $v_f(\varphi) \neq v_f(\psi)$.

The meaning of the logical connectives can be illustrated with the help of *truth tables*. Consider the formula $\neg p$. There is only one atomic proposition in this formula, which means that there are two possible interpretations for the formula; either p is true, or p is false. Consider the truth table for negation given in Table A.2. Each row corresponds to an interpretation of the formula occurring on the far right of the top row. Thus, if the atomic proposition p is true, then the resulting formula when the connective $\neg$ is applied to p is false, and vice versa. The idea is to show how the truth value of a formula containing a logical connective depends on its parts. This is done by calculating the truth value under every possible interpretation of the formula using the rules in Definition A.2.2. The truth table for the binary connectives are also given in Table A.2.

A.3 Satisfiability and validity

Definition A.3.1. Let φ be a formula.

- φ is *satisfiable* if there is some interpretation such that $v_f(\varphi) = T$;
- φ is *unsatisfiable* if it is not satisfiable;
- φ is *valid* (denoted $\models \varphi$) if $v_{f_\varphi}(\varphi) = T$ for every interpretation;
- φ is *falsifiable* (denoted $\not\models \varphi$) if it is not valid.

A valid formula is also called a *tautology*, and an unsatisfiable formula is also called a *contradiction*. Note the following properties of the above definition:

- φ is unsatisfiable if and only if $\neg\varphi$ is valid;
- φ is satisfiable if and only if $\neg\varphi$ is falsifiable.

p	q	$p \wedge q$	$\neg(p \wedge q)$	$\neg p$	$\neg q$	$\neg p \vee \neg q$	φ
T	T	T	F	F	F	F	T
T	F	F	T	F	T	T	T
F	T	F	T	T	F	T	T
F	F	F	T	T	T	T	T

Table A.3: The truth table for the formula $\varphi = \neg(p \wedge q) \leftrightarrow \neg p \vee \neg q$.

p	q	$p \vee q$	ψ
T	T	T	T
T	F	T	T
F	T	T	F
F	F	F	T

Table A.4: The truth tables for the formula $\psi = p \vee q \to p$.

Truth tables can be used to check whether a given formula is satisfiable, unsatisfiable, valid or falsifiable. Consider, as an example, the formula $\varphi = \neg(p \wedge q) \leftrightarrow \neg p \vee \neg q$. The truth table is constructed by considering the truth values of all the subformulas of φ under all possible interpretations of φ. Since there are two atomic propositions occurring in φ, there are $2^2 = 4$ possible interpretations. The truth table for φ is presented in Table A.3. The column on the far right in the above truth table consists of four Ts. This means that φ is true under any interpretation, i.e. φ is a valid formula (a tautology).

Consider the formula $\psi = p \vee q \to p$. There are four possible interpretations of this formula, since it contains two atomic propositions: p and q. The truth table for this formula is presented in Table A.4. The formula ψ is thus not valid, since it is false under the interpretation represented by the third row in the above truth table; this means that ψ is falsifiable. ψ is also satisfiable since there are interpretations under which it is true (the interpretations represented by the first, second, and fourth rows).

A.4 Logical equivalence

It is often useful to substitute a formula for a logically equivalent formula, for example to simplify a complex formula. Two formulas are logically equivalent if they have the same truth-value under all interpretations. Formally, this is defined as:

Definition A.4.1. Let φ, ψ be formulas. φ and ψ are *logically equivalent* (denoted $\varphi \equiv \psi$) if $v_f(\varphi) = v_f(\varphi)$ for all interpretations f.

It is important to note that $\equiv$ is not a symbol of PL; rather, it is a symbol used in the metalanguage to reason about PL.

However, as seen in Definition A.2.2, the connective $\leftrightarrow$ behaves in a way reminding of the above definition of logical equivalence; $\varphi \leftrightarrow \psi$ is true under an interpretation if and only if φ and ψ have the same truth-value under the interpretation. This suggests that there is an interesting relation between $\equiv$ and $\leftrightarrow$. In fact, the following is a theorem:

Theorem A.4.1. *$\varphi \equiv \psi$ if and only if $\varphi \leftrightarrow \psi$ is valid (i.e. true under any interpretation).*

The proof is quite simple, and can be found in [2, p. 22].

Truth tables and Theorem A.4.1 can be used to check whether two formulas are equivalent. Consider the example formula $\varphi = \neg(p \wedge q) \leftrightarrow \neg p \vee \neg q$ from Section A.3. It was shown in Section A.3 that φ is a valid formula. By Theorem A.4.1, $\neg(p \wedge q) \equiv \neg p \vee \neg q$.

Theorem A.4.2. *Let φ be a formula containing logical connectives, and let ψ be a formula occurring in φ. Assume that $\psi \equiv \psi'$. Let φ' be the formula resulting from φ by substitution of every occurrence of ψ with ψ'. Then $\varphi \equiv \varphi'$.*

This theorem describes *substitution*, and shows that it is possible to form new equivalences from older ones.

A.4.1 Examples of useful equivalences

There are several logical equivalences that are particularly useful. Here is a short list of such equivalences:

$$\begin{aligned}
\neg\neg\varphi &\equiv \varphi;\\
\varphi \wedge \psi &\equiv \neg(\neg\varphi \vee \neg\psi);\\
\varphi \vee \psi &\equiv \neg(\neg\varphi \wedge \neg\psi);\\
\neg(\varphi \wedge \psi) &\equiv \neg\varphi \vee \neg\psi;\\
\neg(\varphi \vee \psi) &\equiv \neg\varphi \wedge \neg\psi;\\
\varphi \rightarrow \psi &\equiv \neg(\varphi \wedge \neg\psi);\\
\varphi \rightarrow \psi &\equiv \neg\varphi \vee \psi;\\
\varphi \rightarrow \psi &\equiv \neg\psi \rightarrow \neg\varphi;\\
\neg(\varphi \rightarrow \psi) &\equiv \varphi \wedge \neg\psi;\\
\varphi \leftrightarrow \psi &\equiv (\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi);\\
\varphi \leftrightarrow \psi &\equiv \neg\varphi \leftrightarrow \neg\psi.
\end{aligned}$$

Truth tables can be used to prove the above equivalences.

Bibliography

1. Asratian, A., Björn, A., and Turesson, B. O., *Diskret matematik*, Linköpings universitet, Linköping, 2011.
2. Ben-Ari, M., *Mathematical Logic for Computer Science*, Springer, London, 2012.
3. Blackburn, P. and van Benthem, J., Modal Logic: A Semantic Perspective, in *Handbook of Modal Logic* (P. Blackburn, J. van Benthem, and F. Wolter, eds.), pp. 2–84, Elsevier, Amsterdam, 2007.
4. Bonnefon, J.-F., Longin, D., and Nguyen, M.-H., Relation of Trust and Social Emotions: A Logical Approach, in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* **2**, pp. 289–292, IEEE Computer Society, Washington, 2009.
5. Bonnefon, J.-F., Longin, D., and Nguyen, M.-H., A Logical Framework for Trust-related Emotions, in *Proceeding of the Third International Workshop on Formal Methods for Interactive Systems (FMIS)*, Electronic Communications of the EASST **22**, 2010.
6. Bratman, M. E., *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge, 1987.
7. Briggs, R., Normative Theories of Rational Choice: Expected Utility, in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), The Metaphysics Research Lab, Stanford University, Stanford, 2014, `http://plato.stanford.edu/archives/fall2014/entries/rationality-normative-utility/`
8. Bulling, N. and Hindriks, K. V., Taming the Complexity of Linear Time BDI Logics, in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, pp. 275–282, Taipei, 2011.
9. Castelfranchi, C. and Falcone, R., Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification, in *Proceedings of the Third International Conference on Multi-Agent Systems, (ICMAS-98)*, pp. 72–79, IEEE Press, New York, 1998.
10. Castelfranchi, C. and Falcone, R., Trust is Much More Than Subjective Probability, in *32nd Hawaii International Conference on System Sciences – Mini-Track on Software Agents, Maui*, IEEE Press, New York, 2000.
11. Castelfranchi, C. and Falcone, R., Social Trust: A Cognitive Approach, in *Trust and Deception in Virtual Societies* (C. Castelfranchi and Y. H. Tan, eds.), pp. 55–90, Kluwer, Dordrecht, 2001.

12. Castelfranchi, C. and Falcone, R., *Trust Theory: A Socio-Cognitive and Computational Model*, Wiley, Chichester, 2010.
13. Castelfranchi, C., Falcone, R. and Lorini, E., A Non-reductionistic Approach to Trust, in *Computing With Social Trust* (J. Goldbeck, ed.), pp. 45–72, Springer, London, 2009.
14. Demolombe, R. and Lorini, E., A Locial Account of Trust in Information Sources, in *11th Internationl Workshop on Trust in Agent Societies*, Estoril, 2008.
15. Dennett, D., *The Intentional Stance*, MIT Press, Cambridge, 1987.
16. Dignum, V. and Padget, J., Multiagent Organizations, in *Multiagent Systems* (G. Weiss, ed.), pp. 51–98, MIT Press, Boston, 2013.
17. Doherty, P., Heintz, F., and Landén, D., A Delegation-Based Architecture for Collaborative Robotics, in *Agent-Oriented Software Engineering XI: 11th International Workshop, AOSE 2010* (D. Weyns and M.-P. Gleizes, eds.), pp. 205–247, Toronto, 2011.
18. Doherty, P. and Meyer, J. J. C., On the Logic of Delegation – Relating Theory and Practice, in *The Goals of Cognition: Essays in honour of Cristiano Castelfranchi* (F. Paglieri, L. Tummolini, R. Falcone, and M. Miceli, eds.), pp. 467–496, College Publications, London, 2012.
19. Dunin-Kęplicz, B., Nguyen, L. A. and Szałas, A., Horn-TeamLog: A Horn Fragment of TeamLog with PTime Data Complexity, in *Computational Collective Intelligence. Technologies and Applications* (C. Bădică, N.-T. Nguyen, and M. Brezovan, eds.), Lecture Notes on Computer Science **8083**, pp. 143–153, Springer, Berlin, 2013.
20. Dunin-Kęplicz, B., Nguyen, L. A. and Szałas, A., A Framework for Graded Beliefs, Goals and Intentions, *Fundamenta Informaticae* **100** (2010), 53–76.
21. Dunin-Kęplicz, B., and Verbrugge, R., *Teamwork in Multi-agent Systems: A Formal Approach*, Wiley, Hoboken, 2010.
22. Gambetta, D., Can We Trust Trust?, in *Trust: Making and Breaking Cooperative Relations* (D. Gambetta, ed.), pp. 213–237, Blackwell, Cambridge, 1988.
23. Gaudou, B., Herzig, A., and Longin, D., A Logical Framework for Grounding-based Dialogue Analysis, in *Proceedings of the Third International Workshop on Logic and Communication in Multi-Agent Systems (LCMAS)*, Electronic Notes in Theoretical Computer Science **157**, pp. 117–137, Elsevier, Amsterdam, 2006.
24. Hardaker, C., Trolling in Asynchronous Computer-mediated Communication: From User Discussions to Academic Definitions, *Language, Behaviour, Culture* **6** (2010), 215–242.
25. Herzig, A., Lorini, E., Hübner, J. F., and Vercouter, L., A Logic of Trust and Reputation, *Logic Journal of IGPL* **18** (2010), 212–244.
26. van der Hoek, W. and Wooldridge, M., Logics for Multiagent Systems, in *Multiagent Systems* (G. Weiss, ed.), pp. 762–810, MIT Press, Boston, 2013.
27. Krupa, Y., Vercouter, L., Hübner, J. F., and Herzig, A., Trust Based Evaluation of Wikipedia's Contributors, in *Proceedings of the 10th International Workshop on Engineering Societies in the Agents World X*, ESAW '09, Lecture Notes in Computer Science **5881**, pp. 148–161, Springer, Berlin, 2009.

28. Kurucz, A., Combining Modal Logics, in *Handbook of Modal Logic* (P. Blackburn, J. van Benthem, and F. Wolter, eds.), pp. 869–924, Elsevier, Amsterdam, 2007.
29. Lorini, E. and Demolombe, R., Trust and Norms in the Context of Computer Security: A Logical Formalization, in *Deontic Logic in Computer Science* (R. van der Meyden and L. van der Torre, eds.), Lecture Notes in Computer Science **5076**, pp. 50–64, Springer, Berlin, 2008.
30. Lorini, E. and Demolombe, R., From Binary Trust to Graded Trust in Information Sources: A Logical Perspective, in *Trust in Agent Societies* (R. Falcone, S. Barber, J. Sabater-Mir, and M. Singh, eds.), Lecture Notes in Computer Science **5396**, pp. 205–225, Springer, Berlin, 2008.
31. Lorini, E. and Herzig, A., A Logic of Intention and Attempt, *Synthese* **163** (2008), 45–77.
32. Mårtensson, B., *Logik: En introduktion*, Studentlitteratur, Lund, 2009.
33. Meyer, J.-J. and Veltman, F., Intelligent Agents and Common Sense Reasoning, in *Handbook of Modal Logic* (P. Blackburn, J. van Benthem, and F. Wolter, eds.), pp. 991–1029, Elsevier, Amsterdam, 2007.
34. McKnight, D. H. and Chervany, N. L., Trust and Distrust Definitions: One Bite at a Time, in *Trust in Cyber-societies* (R. Falcone, M. Singh, and Y. H. Tan, eds.), pp. 27–54, Springer, Berlin, 2001.
35. Mints, G., Gentzen-type Systems and Resolution Rules Part I: Propositional Logic, in *COLOG-88* (P. Martin-Löf and G. Mints eds.), Lecture Notes on Computer Science **417**, pp. 198–231, Springer, Berlin, 1988.
36. Needham, P., *A First Course in Modal Logic*, Department of Philosophy, University of Stockholm, Stockholm, 1999.
37. Nguyen, L. A., Constructing the Least Models for Positive Modal Logic Programs, *Fundamenta Informaticae* **42** (2000), 29–60.
38. Nguyen, L. A., On the Complexity of Fragments of Modal Logics, in *Advances in Modal Logic* 5, pp. 249–268, King's College Publications, London, 2005.
39. Nguyen, L. A., Multimodal Logic Programming, *Theoretical Computer Science* **360** (2006), 247–288.
40. Prawitz, D., *ABC i symbolisk logik*, Thales, Stockholm, 2010.
41. Stalnaker, R., Common Ground, *Linguistics and Philosophy* **25** (2002), 701–721.
42. Ullmann-Margalit, E., Trust out of Distrust, *The Journal of Philosophy* **10** (2002), 532–548.
43. Ullmann-Margalit, E., Trust, Distrust, and in Between, in *Distrust* (R. Hardin, ed.), pp. 60–82, Russell Sage Foundation, New York, 2004.
44. Weiss, G. (ed.), *Multiagent Systems*, MIT Press, Boston, 2013.
45. Wooldrigde, M., *An Introduction to Multiagent Systems* (2nd ed.), Wiley, Chichester, 2009.