Utrecht University

master Artificial Intelligence, Utrecht University

30 ECTS

Interpretation STIT Logic

Author:
Tjeu Hendriks

Student number:
3974685

Supervisor:
Benjamin G. Rin

Second corrector:
Jan M. Broersen

October 28, 2019

Abstract

STIT logic is a logic that can model choices. STIT is an abbreviation of "seeing to it that". Herzig and Troquard defined in 2006 a new version of the STIT logic that combines knowledge with STIT. This logic has a version of the STIT operator that models knowingly seeing to it. However, defining a generally accepted notion of knowledge has historically proved to be very difficult. In this thesis an alternative for a knowledge operator is suggested, namely the interpretation operator. This operator models an agent "acting as if" some proposition holds. A more abstract operator like the interpretation operator can help give new insight in the problems where a notion of knowledge seems to be necessary without defining a notion of knowledge. For instance, problems concerning responsibility can be modeled with interpretation instead of knowledge.

Contents

1 Introduction

Obligation is generally understood in a ought-to-be setting, e.g. you ought to have clean teeth every day. Most of the time statements of obligation talk about actions, e.g. you ought to clean your teeth. The problem with this understanding of obligation is that actions do not translate easily to logic, since most logics use factual statements, e.g. you have clean teeth. STIT is a logic defined to model actions together with factual statements. This logic therefore closes a bridge between obligation and logic. STIT is an abbreviation for for "seeing to it that". The logic uses an operator that models an agent seeing to it that some statement holds, e.g. you see to it that your teeth are clean. By using this operator, actions are thus modeled as someone bringing about a state of the world.

Knowledge also plays a big role in the notion of obligation, namely in the violation of responsibilities. If you are prohibited of doing something and you violate that prohibition, it can matter whether you knew you were violating or not. Herzig and Troquard[13] define a new version of the STIT logic that combines knowledge with STIT. This logic has a version of the STIT operator that models "knowingly seeing to it". To define such an operator Herzig and Troquard need to modify the frame that STIT uses to determine what agents know about the world when they take actions. However, defining a generally accepted notion of knowledge has historically proved to be very difficult. Many philosophers have proposed many different definitions of knowledge and none of them are generally accepted as the ultimate definition of knowledge. Creating a general knowledge operator is therefore a very challenging task. This in turn creates big problems for deontic logic as well, since knowledge seems to play a part in obligation. For example, how can you be obligated to ensure a proposition holds in the world, if you do not know whether that proposition holds. So if it seems to be impossible to provide a generally acceptable definition of knowledge and if it seems to be the case that knowledge plays a part in the definition of obligation, then how can we define obligation?

This thesis is based on the research of the interactions between knowledge and STIT logic. In this thesis an alternative for a knowledge operator is suggested, namely the interpretation operator. This operator models an agent "acting as if" some proposition holds. The new model, the Interpretation STIT model, is an extension of the original STIT model where the interpretations of agents are modeled as well. A more abstract operator like the interpretation operator can help shed light on problems where a notion of knowledge seems to be necessary without defining a notion of knowledge.

This thesis will start with an overview of STIT and deontic logic. In the following part earlier research into combining knowledge and STIT will be discussed. This part will also include the paper of Herzig and Troquard[13]. After that, a few examples that combine responsibility and knowledge are given. Then the Interpretation STIT model is introduced. The formal definitions of the new model are discussed and the possibilities of Interpretation STIT are discussed and some Interpretation STIT alternatives to seemingly epistemic problems are given. Eventually a soundness and completeness proof is given.

2 The definition of STIT

STIT is an abbreviation for "seeing to it that". STIT logic determines who saw to it that the world is in a certain state. The STIT logic essentially uses a Kripke frame, but in practice the frame is a tree-like structure that represents possible states of the world based on actions made before those states. This tree-like frame is seen as a Kripke frame where every route from the root to a leaf is a

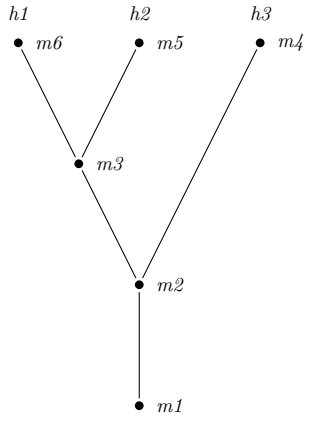
world in the Kripke frame. Nodes in this kind of Kripke frame are called "moments." Such a use of a Kripke frame is defined as a *branching time frame*. The formal definition of this type of Kripke frame is nicely stated by Horty[15]:

Definition 1. *A branching time frame is a structure* $\boldsymbol{F}$ *of the form* $\langle Tree, < \rangle$ *with* $Tree$ *a nonempty set of moments, and* $<$ *a transitive, irreflexive, and tree-like ordering on* $Tree$*.*

2.1 A branching time frame

The branching time frame is a tree-like graph that represents a branching timeline. One timeline, called a history in STIT, is one path, defined by its vertices from the root of the tree to one of the leaves. A vertex in the graph, called a moment, represents a moment in time. The name *history* is used because of the nature of most STIT logics that use these branching time frames. These logics define valuations based on outcomes of certain timelines, which makes those timelines go from a certain point in history to the moment of valuation.

Example 1. *Starting from below, an example of a branching time frame is represented as follows:*



with $m1...m6$ *being the moments of the frame and* $h1 = \{m1, m2, m3, m6\}, h2 = \{m1, m2, m3, m5\}, h3 = \{m1, m2, m4\}$ *being the histories.*

A STIT model that uses the branching time frame assigns valuations to propositions for every moment in a certain history. Such a moment in a history is written as m/h. For instance, if ϕ is some formula in the STIT model and ϕ is true at $m2$ in history $h1$ in model $\mathcal{M}$, then $\mathcal{M}, m_2/h_1 \vDash \phi$ holds. Semantically this means that viewing the subtree containing $m2$ and all future moments, namely $m2, m3, m4, m5, m6$, at the moment defining $h1$, which is $m6$, ϕ is true, so $\mathcal{M}, m_6/h_1 \vDash \phi$ also holds. The starting moment of evaluation (in the example being $m2$) is necessary to define, because of modal and STIT operators that can be contained in the syntax of the logic. For instance, at $m2$, an operator that defines the necessitation of a certain proposition ϕ will semantically be defined as being true in the model if ϕ is true for all histories that contain the moment $m2$. For this operator the evaluated history is not important. Only the moment from where the history is evaluated matters.

2.2 A STIT frame

The STIT operator in STIT models evaluates the choices made in a branching time frame. These choices are made by agents and cause branching in the branching time frame. Intuitively, this translates to the idea that if someone took an action that ensured something to happen or ensured that some proposition would be true, then he or she *saw to it that* the event would occur or that the proposition would be true. To incorporate these choices in the model, the branching time frame is extended. This extended frame is called a *STIT frame*.

The STIT frame manipulates the moments of a branching time frame by bundling the branches of a moment (node in the branching time frame) and defining those sets of branches as a result of choices. So when at a certain moment in the frame someone makes a choice, it is then defined that only a subset of the branches from that moment are possible results from that moment onward.

To define the choices of actions in the frame, a set of *agents* and a *Choice function* are added to the frame. The set *agents* defines all the entities (commonly humans) that are able to make a choice in the STIT frame. The *Choice function* is a function that defines for an agent what set of histories a certain history belongs to. Since evaluations are made at a certain moment at the end of a history, it is only relevant to know what choices are made, i.e. what other histories, beside the evaluated history, were possible when choices by agents were made.

Definition 2. *A STIT frame is a structure of the form*

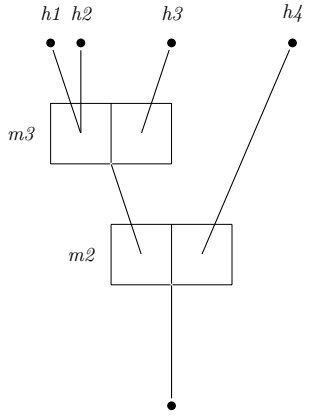
$$\langle Tree, <, Agent, Choice \rangle$$

where Tree and $<$ are the branching time frame from definition 1, Agent is a nonempty set of agents and Choice is a function mapping each agent α, moment m, and history h to a subset of H_m, where H_m is the set of all histories h where $m \in h$ holds.

When used, a choice function has the form $Choice_m^{\alpha}(h)$ where m is a moment, α is an agent and h is a history. The result of such a function is a subset of all histories that go through moment m. This subset usually contains the history h. What the other histories are depends on the choice function.

The *Choice* function is subject to the *independence requirements*. These requirements ensure that no choice by an agent can be affected by other agents. For every moment in a STIT frame it holds that only the designated agents modeled for a choice at that moment get to affect that moment. This is an example of Choices in a STIT frame:

Example 2. *Let the branching time frame be the same as in example 1. Let Agent be $\{\alpha\}$ and let the choice function be $Choice_{\alpha}^{m2}(h1) = Choice_{\alpha}^{m2}(h2) = Choice_{\alpha}^{m2}(h3) = \{h1, h2, h3\}$, $Choice_{\alpha}^{m2}(h4) = \{h4\}$, $Choice_{\alpha}^{m3}(h1) = Choice_{\alpha}^{m3}(h2) = \{h1, h2\}$ and $Choice_{\alpha}^{m3}(h3) = \{h3\}$. The STIT frame looks like this:*



As is clear in this example, the choice function divides the set of histories going through a certain moment to define a set of possible outcomes given a certain choice. Such a choice is represented by a square in the visualization of the frame.

2.3 The STIT operator

A STIT operator is defined in the syntax used with a STIT frame. This operator has the form $[\alpha \; stit : A]$ where α is an agent and A is a formula. Intuitively, the STIT operator decides whether agent α saw to it that A holds. In a STIT model this operator is always evaluated based on the choice function. How this operator is exactly evaluated depends on what kind of STIT operator is used. The different operators interpret "seeing to it that" differently.

2.3.1 cstit operator

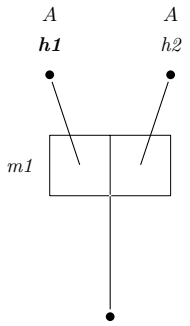
Cstit or "Chellas" STIT is defined by Brian Chellas[11] and is one of the simpler interpretations of STIT. This operator captures the intuition that you *see to it that A* when you made a choice that ensured A would hold.

Definition 3. *Let $\mathcal{M}$ be a model of a STIT frame, m be a moment in $\mathcal{M}$, h be a history in $\mathcal{M}$, α be an agent and A be a statement, then $\mathcal{M}, m/h \vDash [\alpha \; cstit : A]$ holds if and only if $Choice^m_\alpha(h) \subseteq |A|^{\mathcal{M}}_m$, where $Choice^m_\alpha(h)$ is the set of histories possible after agent α made the choice that makes h a possible history at moment m, and $|A|^{\mathcal{M}}_m$ is the set of histories in $\mathcal{M}$ and containing m where A holds.*

The cstit operator is very simple and easily definable in a STIT frame. You 'see to it that A' when you made a choice somewhere in the frame and all histories that resulted from the choice ensured A. A conceptual problem with the cstit operator is that it is possible that you technically saw to it that A because you made a choice somewhere, even though A would be ensured regardless of your choice. This may not be desirable because of situations like example 3

Example 3. *The day is sunny and Albert decided to take a hike. At a certain moment Albert bought a refreshment. Did Albert see to it that it was sunny because he bought a refreshment?*

A corresponding STIT frame to this problem is:



where $m1$ *is the choice of Albert whether to buy a refreshment or not and* A *is the proposition that is true iff the sun shines. Considering the formula* $[Albert\ cstit:\ A]$ *it is clear that, by definition 3, this holds. Intuitively Albert couldn't have seen to it that the sun shines. It just happened to be the case.*

2.3.2 dstit operator

The problem shown in example 3 can be solved by stating something about the alternative choices that an agent has. "Deliberative STIT" or *dstit* takes exactly that into account. First introduced by von Kutschera[16], *dstit* tries to express the deliberateness of a choice. An agent only deliberately sees to it that A holds if the agent makes a choice that ensures A and the agent had another choice that did not ensure A.

Definition 4. *Let* $\mathcal{M}$ *be a model of a STIT frame,* m *be a moment in* $\mathcal{M}$*,* h *be a history in* $\mathcal{M}$*,* α *be and agent and* A *be a statement, then* $\mathcal{M}, m/h \vDash [\alpha\ dstit:\ A]$ *holds if and only if* $Choice^m_\alpha(h) \subseteq |A|^{\mathcal{M}}_m$ *and* $|A|^{\mathcal{M}}_m \neq H_m$*, where* $Choice^m_\alpha(h)$ *is the set of histories possible after agent* α *made the choice that makes* h *a possible history at moment* m*, and* $|A|^{\mathcal{M}}_m$ *is the set of histories in* $\mathcal{M}$ *going through* m *where* A *holds, and where* H_m *is the set of histories that contain* m*.*

Dstit clearly solves the problem of example 3, since it is impossible to deliberately see to it that something inevitable holds. However, there are some difficulties that both *cstit* and *dstit* have to deal with. Kenny[17] challenges the modal possibility operator $\Diamond$ that both *cstit* and *dstit* required in their logic to work as expected. The problem Kenny stated was that certain axioms that define the $\Diamond$ operator in STIT do not work for a semantic notion of *ability*. Kenny found that modal possibility operators cannot assume the meaning of "being able to". The problem Kenny had was that there are axioms that are generally accepted to hold for possibility, but cannot hold for a hypothetical ability operator. The problematic axioms according to Kenny are the following:

$$T\Diamond.\ A \to \Diamond A$$

$$C\Diamond.\ \Diamond(A \vee B) \to (\Diamond A \vee \Diamond B)$$

To explain the problem with the notion of ability, Kenny gave an example of a darts player for both the axioms.

Example 4. *As a counterexample for* $T\Diamond$*, consider a poor darts player. This player throws the dart and happened to hit the bullseye by chance. According to* $T\Diamond$*, since the player hit the bullseye, it is possible for the player to hit the bullseye. However this should not mean that the player has the ability to hit the bullseye.*

Example 5. *As a counter example for $C\Diamond$, consider a darts player that is sufficiently skilled to hit a dart board, but cannot throw more precisely than that. Since the player can hit the dart board, he will either hit the top half or the bottom half. According to $C\Diamond$, since he has the ability to throw in the top or bottom half, he has the ability to throw in the top half or he has the ability to throw in the bottom half. This is in contradiction with the assumption that he could not hit more precise than a dart board.*

There are two ways to solve these examples. One of them is to not see the possibility operator $\Diamond$ as an ability operator. The consequence of that is that it is impossible to talk about ability in a solely modal sense. Another way to solve this is to create a new operator to be an ability operator. This new operator gave rise to another STIT operator called "Brown" STIT or *bstit*.

2.3.3 bstit operator

Mark Brown[10] introduced a STIT operator for ability. To solve the problems explained by example 4 and 5 for STIT, Brown introduced a new STIT operator *bstit*. The *bstit* operator is to be interpreted as *the ability to see to it that*.

Definition 5. *Let $\mathcal{M}$ be a model of a STIT frame, m be a moment in $\mathcal{M}$, h be a history in $\mathcal{M}$, α be and agent and A be a statement, then $\mathcal{M}, m/h \vDash [\alpha \ bstit : A]$ holds if and only if there is some action K in $Choice^m_\alpha(h)$ such that for every $h' \in K$ it holds that $\mathcal{M}, m/h' \vDash A$.*

Essentially, the meaning of the *bstit* operator is "the possibility to see to it that". The *bstit* operator can be seen as $\Diamond[\alpha \ cstit : A]$. One might say that this is definition for "the ability to see to it that" is too weak since you would always have the ability to see to the inevitable. Another way to describe *ability* is to use *dstit* instead of *cstit*.

The bstit operator will not be used in the rest of the thesis, but can be used in combination with the interpretation STIT model to describe some deontic problems that require both ability and knowledge to be solved.

2.4 Independent choices

It is also possible in STIT to model several choices of agents that are independent of each other. In this case multiple agents will make a choice at a certain moment. The outcome of the choices then depends on every choice made independently. Since the choices are independent, it does not necessarily mean that the choices are made at the same time. Choices being in the same moment does not mean that they are made in the same temporal instant. Two moments are only temporally related with respect to the dependent choices modeled in the moments. As described by Horty[15], the working of a joint Choice function is defined as follows:

Definition 6. *Let m be a moment and let $\Gamma \subseteq Agent$ be a non-empty set of agents. $Choice^m_\Gamma$ is defined as:*

$$Choice^m_\Gamma = \{\bigcap_{\alpha \in \Gamma} s(\alpha) : s \in Select_m\}$$

where $Select_m$ is defined as a set of functions defined by

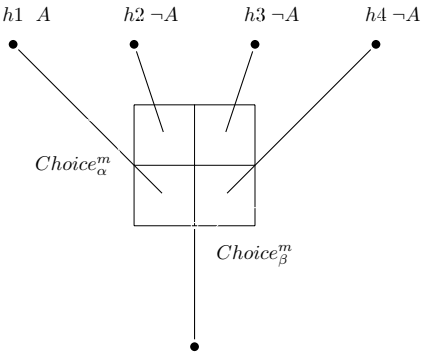
$$s(\alpha) \in Choice^m_\alpha$$

Definition 7. *Let Γ be a set of agents, let A be a formula, let $\mathcal{M}$ be a STIT model, let m be a moment in $\mathcal{M}$ and let h be a history in $\mathcal{M}$. Then, the following holds*[1]*:*

$$\mathcal{M}, m/h \vDash [\Gamma \ cstit : A] \ if \ and \ only \ if \ Choice^m_\Gamma(h) \subseteq |A|^{\mathcal{M}}_m$$

The joint choice function can be described as the intersection of the individual choice functions:

Example 6. *Let $\mathcal{F}$ be a STIT frame $\langle Tree, <, Agent, Choice\rangle$, where Agent is $\{\alpha, \beta\}$ and Choice is defined by $Choice^m_\alpha(h1) = Choice^m_\alpha(h4) = \{h1, h4\}$, $Choice^m_\alpha(h2) = Choice^m_\alpha(h3) = \{h2, h3\}$, $Choice^m_\beta(h1) = Choice^m_\beta(h2) = \{h1, h2\}$ and $Choice^m_\beta(h3) = Choice^m_\beta(h4) = \{h3, h4\}$. Let the rest of $\mathcal{F}$ be defined by the following STIT frame:*



As an example of a joint choice function, the joint choice function for α and β applied to $h1$ is:

$$Choice^m_{\{\alpha,\beta\}}(h1) = Choice^m_\alpha(h1) \cap Choice^m_\beta(h1) = \{h1, h4\} \cap \{h1, h2\} = \{h1\}$$

As an example of the evaluation, since $h1 \in |A|^{\mathcal{M}}_m$ it holds that $\mathcal{M}, m/h1 \vDash [\Gamma \ cstit : A]$. This Choice function for multiple agents can also be used for single agents. In example 6 both $\mathcal{M}, m/h2 \vDash [\alpha \ cstit : \neg A]$ and $\mathcal{M}, m/h2 \vDash [\{\alpha\} \ cstit : \neg A]$ hold. It is also clear to see that this Choice function can be used with all the types of the STIT operator.

3 Deontic logic

Deontic logic is the logic concerning obligation and permission. More specifically, deontic logic concerns the logical representation of the natural terms "obligatory", "permissible", "impermissible", "omissible" and "optional". STIT is very useful for deontic logic, since STIT together with deontic logic can model an "ought to do" statement while deontic logic alone can only model "ought to be" statements. In the past century the modeling of permission and obligation is assumed to be a problem solvable with modal logic, since it seems that the terms "obligation" and "permission" have a somewhat similar relation to each other as the box and diamond operators have. When p is obligatory, then $\neg p$ is not permissible and when $\Box p$ holds, then $\neg\Diamond\neg p$ holds. One of the first modal deontic logic was defined by Von Wright[21][18] and became later known as "Standard deontic logic" or SDL.

[1]Definition is the same for *dstit* and *bstit*

3.1 Standard deontic logic

SDL is a modal logic that uses standard propositional operators together with a OB operator. The OB operator validates the modal axioms K and D. This results in the following axiom scheme for SDL:

TAUT All the propositional logical tautologies.

K $OB(p \to q) \to (OBp \to OBq)$

D $OBp \to \neg OB \neg p$

MP if p and $p \to q$ then q.

NEC if p then OBp

The D axiom seems to be an intuitive axiom to include when assuming that permission is the dual of obligation. If it is obligatory that p, then it is permitted that p. The K axiom also seems to be an intuitive one. If it is obligatory that you brush your teeth if you go to bed, then it holds that if it is obligatory that you go to bed, then it is obligatory that you brush your teeth.

However, there are some difficulties regarding SDL. One of the problems that SDL cannot deal with is the *Free Choice Permission Paradox* by Ross[20][18]:

Problem 1. *Consider the following statements:*

1. *You may sleep on the sofa or sleep in the guest room bed.*
2. *You may sleep on the sofa and you may sleep in the guest room bed.*

These two statements are intuitively interchangeable, since they express the same permissions. In SDL, an obvious way to interpret these two statements are respectively:

1. $PE(p \vee q)$
2. $PEp \wedge PEq$

where PE is the operator semantically defined as "it is permitted to". In SDL 1 *follows from* 2*. However, in standard SDL it should seem that* 2 *also follows from* 1*, but this formula does not hold:*

$$PE(p \vee q) \to PEp \wedge PEq$$

In addition, consider this formula to be taken into SDL. If it holds then undesirable derivations could be made.

Consider PEp to be true, then by weakening[2] *$PE(p \vee q)$ is true. Then, by using the newly introduced axiom of $PE(p \vee q) \to PEp \wedge PEq$ the formula $PEp \wedge PEq$ is also true. This derivation entails that if something is permitted then everything is. This is clearly an undesirable derivation.*

Another problem with SDL is the *good samaritan paradox* from Prior[19][18]. This problem exposes a different flaw with SDL in a similar fashion as problem 1.

Problem 2. *Consider the following statements:*

[2] Weakening is considered to hold for deontic modal logic and states that $\Diamond p \to \Diamond(p \vee q)$.

1. *It ought to be the case that Jones helps Smith who has been robbed.*

2. *It ought to be the case that Smith has been robbed.*

In SDL 2. *follows from* 1. *which is clearly undesirable. The reasoning goes as follows:*
It is intuitively sound that Jones helps Smith who has been robbed if and only if Jones helps Smith and Smith has been robbed. Then the two statements are interpreted in SDL respectively:

1. $OB(p \wedge q)$

2. OBq

Now by the logical tautology $(p \wedge q) \rightarrow q$ *and the derivable rule that if* $r \rightarrow s$ *then* $OBr \rightarrow OBs$ *one can derive that* $OB(p \wedge q) \rightarrow OBq$.

3.2 Deontic logic in STIT

There are many ways to incorporate a deontic operator in STIT logic. An obvious way to incorporate it is to attach values to the histories that depend on how preferable those histories are. The standard way of deontic logic in STIT only differentiates between acceptable and not acceptable histories.

3.2.1 The standard deontic STIT model

The standard deontic STIT model uses the initial STIT model and adds a set of moment history pairs that are deemed acceptable (the definition is from Horty [15]):

Definition 8. *A standard deontic STIT frame is a 5-tuple of the form*

$$\langle Tree, <, Agent, Choice, Ought \rangle$$

where $Tree$, $<$, *Agent and* $Choice$ *are the same as in definition 2, and* $Ought$ *is a function mapping every moment* m *to a subset of* H_m.

On this frame it is then very easy to define an operator that is interpreted as "it ought to be that". The idea behind the operator would just be that it ought to be that p if and only if p is true in every acceptable history. Horty[15] defines it as:

Definition 9. *Where* m/h *is a moment history pair from a standard deontic STIT model* M,

$$M, m/h \vDash \bigcirc A$$

if and only if $M, m/h' \vDash A$ *holds for every history* $h' \in Ought(m)$.

The axiomatic rules for the $\bigcirc$ operator are standard modal rules.

1. $A \equiv B / \bigcirc A \equiv \bigcirc B$

2. $\bigcirc \top$

3. $\bigcirc(A \wedge B) \rightarrow (\bigcirc A \wedge \bigcirc B)$

4. $(\bigcirc A \wedge \bigcirc B) \rightarrow \bigcirc(A \wedge B)$

For this operator, one should also assume that if something ought to be the case, then it should also be possible that it is the case. The axiom that enforces this is:

$$\bigcirc A \rightarrow \Diamond A$$

Adding this axiom to the scheme also incidentally implies that it cannot be the case that it ought to be the case that p and it ought to be the case that $\neg p$ simultaneously.[3] Since this axiom holds, the $\bigcirc$ operator is a modal weak S5 operator. This is mostly considered to be too strong of an operator for obligation. One of the formulas that also can be derived for the obligation operator is the following:

$$\Box A \rightarrow \bigcirc A$$

Intuitively it is not desired that if a proposition is necessary, then it also ought to be the case.

3.2.2 The utilitarian STIT frame

The utilitarian STIT frame gives weights to every history based on the preference of that history. Formally it only replaces the *Ought* function from the standard deontic frame and replaces it with a *Value* function.

Definition 10. *A general deontic STIT frame is a 5-tuple*

$$\langle Tree, <, Agent, Choice, Value \rangle$$

where $Tree$, $<$, $Agent$ and $Choice$ are similar to a normal STIT frame and $Value$ function taking a moment and a history through that moment and returning a numeric value.

The *Ought* operator in a *general deontic STIT model* now models the optimal situation at any moment. This way of implementing deontic logic in STIT can be of good use for the interpretation STIT that will be explained is section 6.

Definition 11. *Let M be a general STIT model and let m/h be any moment history pair in M. Then $M, m/h \vDash \bigcirc A$ holds if and only if there is a history $h' \in H_m$ where $M, m/h' \vDash A$ holds and for every $h'' \in H_m$ it holds that $Value_m(h') \geq Value_m(h'')$.*

This version of the *Ought* operator does not have the problem that it had in the *standard deontic STIT model* since it models an *optimal* situation instead of a *necessary* situation.

4 Knowledge and deontic logic

Originally, the most obvious way to implement knowledge with deontic logic was to add the modal knowledge operator to SDL. This implementation came with some undesirable derivations of formulas combining the $\bigcirc$ operator and the knowledge operator. Aqvist[1][8] defines *The paradox of epistemic obligation* that is generally accepted as a flaw of the system. The problem shows that the axiom $KA \rightarrow A$ leads to some undesirable derivations.

Problem 3. *Consider the following statements:*

[3]Suppose for contradiction that $\bigcirc A$ and $\bigcirc \neg A$ both hold. Then $\bigcirc A \wedge \bigcirc \neg A$ also holds and by axiom 4 then $\bigcirc(A \wedge \neg A)$ holds. Since $\bigcirc A \rightarrow \diamond A$ is an axiom, $\diamond(A \wedge \neg A)$ holds. This is a contradiction.

1. *The bank is being robbed.*

2. *It ought to be the case that Jones, the guard, knows that the bank is being robbed.*

3. *It ought to be the case that the bank is being robbed.*

By using SDL and the knowledge operator K, the statements can be logically interpreted like this:

1. r

2. $\bigcirc K_J r$

3. $\bigcirc r$

By applying the axiom $Kp \rightarrow p$, we can derive 3 from 2. This is undesirable since the it intuitively should not hold that if it ought to be the case that Jones knows that the bank is being robbed, then it ought to be the case that the bank is robbed.

To solve this problem in epistemic deontic logic, either the knowledge operator or the obligation operator can be changed. The problem with changing the knowledge operator is that the axiom $Kp \rightarrow p$ is a generally accepted property of knowledge. It is very intuitive to believe that you can only know things that are true.

4.1 Combining knowledge with the STIT operator

To solve *The paradox of epistemic obligation* Broersen[6] defines a STIT logic that incorporates knowledge. The main goal of this logic is to model the statement "knowingly sees to it that". Such a statement combines the knowledge operator and the STIT operator. Broersen's version of the STIT model is an extension on a STIT frame of a logic called *XSTIT*. In this version of STIT extra operators were added to determine whether formulas were true in the future. This is essential for modeling statements such as "knowingly seeing to it", since an agent then has to be sure of the results of its actions to knowingly seeing to it. The next definitions are from Broersen [6]:

Definition 12. *A frame for the XSTIT model is a tuple $F = \langle H, S, R_\Box, \{R_A | A \subseteq Agents\}, \{\sim_a | a \in Agents\}\rangle$ where*

- *H is a non-empty set of histories.*
- *S is a non-empty set of states.*
- *$R_\Box$ is a historical necessity relation from $H \times S$ to itself where $\langle h, s\rangle R_\Box \langle h', s'\rangle$ if and only if $s \equiv s'$.*
- *R_A is an effectivity relation from $H \times S$ to itself where*
  - *R_{Ags} is a function such that if $\langle h, s\rangle R_{Ags} \langle h', s'\rangle$ then $h \equiv h'$. This forces histories to be linear sets of states.*
  - *$R_\Box \circ R_{Ags} \subseteq R_\emptyset$*
  - *$R_A \subseteq R_\Box \circ R_{Ags}$*

- $R_{Ags} \circ R_{\Box} \subseteq R_A$
- $R_A \subseteq R_B$ *for* $A \subseteq B$
- *if* $\langle h, s\rangle R_{\Box}\langle h', s\rangle$ *and* $\langle h, s\rangle R_{\Box}\langle h'', s\rangle$ *then* $\langle h, s\rangle R_{\Box}\langle h''', s\rangle$ *such that for every* A *and* B *where* $A \cap B = \emptyset$*, if* $\langle h''', s\rangle R_A\langle h'''', s'\rangle$ *then* $\langle h', s\rangle R_A\langle h'''', s'\rangle$ *and if* $\langle h''', s\rangle R_B\langle h'''', s''\rangle$ *then* $\langle h'', s\rangle R_B\langle h''''', s''\rangle$.

- $\sim_a$ *is an epistemic relation from* $H \times S$ *to itself where*
  - $\sim_a \circ R_a \subseteq \sim_a \circ R_{Ags}$
  - $R_{Ags} \circ \sim_a \subseteq \sim_a \circ R_a$

Definition 13. *Let* M *be an XSTIT model and let* s/h *be a state history pair in* M*. The validity of a formula at* s/h *is defined as:*

- $M, s/h \vDash p$ *iff* $s/h \in V(p)$
- $M, s/h \vDash \neg\phi$ *iff* $M, s/h \nvDash \phi$
- $M, s/h \vDash \phi \wedge \psi$ *iff* $M, s/h \vDash \phi$ *and* $M, s/h \vDash \psi$
- $M, s/h \vDash K_\alpha\phi$ *iff for all* s'/h' *where* $\langle s, h\rangle \sim_\alpha \langle s', h'\rangle$ *it holds that* $M, s'/h' \vDash \phi$
- $M, s/h \vDash \Box\phi$ *iff for all* s'/h' *where* $\langle s, h\rangle R_{\Box}\langle s', h'\rangle$ *it holds that* $M, s'/h' \vDash \phi$
- $M, s/h \vDash [A\ XSTIT\ :\ \phi]$ *iff for all* s'/h' *where* $\langle s, h\rangle R_A\langle s', h'\rangle$ *it holds that* $M, s'/h' \vDash \phi$

Broersen defines this logic with a version of STIT closest to the "Chellas STIT" or cstit. Broersen also argues about making a dstit version of this XSTIT operator. This seems to be possible, but gets complex very quickly, since the extra axiom of also having the opportunity to not see to it can have many versions when knowledge is added. Is it for knowingly and deliberately seeing to a proposition enough to be able to do otherwise, or do you also have to know that the other option does not ensure that proposition?

Another problem with the epistemic XSTIT logic is that obligation is still hard to incorporate, since problem 3 is still present when one would add the modal deontic operator. Later Broersen extended this logic to work with obligation [7]. He does this by defining how modes of *mens rea* are modeled. Instead of implementing a general deontic operator, Broersen defines operators that combine the knowledge, STIT and deontic operators.

4.2 The solution with regard to the modes of mens rea

Broersen [7] extends the epistemic XTIT logic with operators that model things like acting recklessly, knowingly risking and acting knowingly. These operators combine the deontic, the epistemic and the STIT operator and redefined the definition of how they work together. Broersen introduced two versions of an obligation operator and three versions of an operator that combines obligation and knowledge. The two obligation operators model the obligation to ensure and the obligation to not see to the opposite. The three combined operators model the following[4]:

[4]Broersen also defines three other operators to model the excuse for not knowing of the obligation. However, these operators are not relevant for this thesis.

Acting recklessly If you do not know that you see to it that p holds, then you are in violation since you are acting recklessly.

knowingly risking If you know that you are not seeing to it that p holds, then you are in violation by knowingly risking it not to hold.

acting knowingly If you know that you are seeing to it that p does not hold, then you are in violation by acting knowingly to let p not hold.

Definition 14. *Within the logical framework of epistemic XSTIT, the following operators are defined modeling the the violation requirements of the statements "Acting recklessly", "knowingly risking" and "acting knowingly" respectively:*

- $OK[\alpha \ xstit \ : \ \phi] \equiv_{def} \Box(\neg K_\alpha[\alpha \ xstit \ : \ \phi] \rightarrow [\alpha \ xstit \ : \ V])$
- $OK'[\alpha \ xstit \ : \ \phi] \equiv_{def} \Box(K_\alpha \neg[\alpha \ xstit \ : \ \phi] \rightarrow [\alpha \ xstit \ : \ V])$
- $OK''[\alpha \ xstit \ : \ \phi] \equiv_{def} \Box(K_\alpha[\alpha \ xstit \ : \ \neg\phi] \rightarrow [\alpha \ xstit \ : \ V])$

where V is a predefined proposition modeling the notion of violation of obligation.

These new operators do not suffer from the problem (problem 3) described by Aqvist[1], since the operators are always combined with the STIT operator. If the deontic operator, the knowledge operator and the STIT operator are combined in the way described above, and that is the only way to use the knowledge operator, then the derivation of Aqvist is not a problem anymore. The only derivation that can be made in this situation is that if you ought to be knowingly seeing to it, then you also ought to be seeing to it. However, the extra operators are only defined for single agents. This solution for combining knowledge with obligation also requires a lot of new operators, which can be seen as less elegant.

4.3 Knowledge as part of the framework

Herzig and Troquard[13] define a somewhat similar version of STIT logic as Broersen, but with one extra relation. This is a relation between epistemically indistinguishable moments. If two choices in moments will look like the same choice for an agent, then this relation defines that these moments are indistinguishable for that agent. Adding to this alteration of the STIT frame, Herzig and Troquard define a new version of the STIT operator that is defined like this:

Definition 15. *Let $\mathcal{M}$ be a STIT model and let m, h be a moment and a history in $\mathcal{M}$ respectively. Let α be an agent in the model. The the following is the definition for $KSTIT$:*

$$\mathcal{M}, m/h \vDash KSTIT_\alpha \phi \Longleftrightarrow \forall m'/h' \in R_{US_\alpha}(m/h), \mathcal{M}, m'/h' \vDash \phi$$

where $R_{US_\alpha}(m/h)$ is the set of all the moment-history pairs that are indistinguishable from m/h for agent α.

This definition uniquely defines knowledge as a property of the (extended) STIT frame.

The solution from Herzig and Troquard is an efficient way to model knowledge in STIT. However, Herzig and Troquard do not explore the addition of obligation into their logic. In addition, the $KSTIT$ operator is an extension of *cstit*. A *dstit* version of this operator can create unexpected

and unintuitive outcomes. Another problem of this logic is that it models the knowledge of ability instead of action according to Broersen [7]. This results in the impossibility of modeling statements representing the modes of mens rea, since the knowledge operator and the STIT operator cannot be separated.

5 An analysis of knowledge within STIT frames

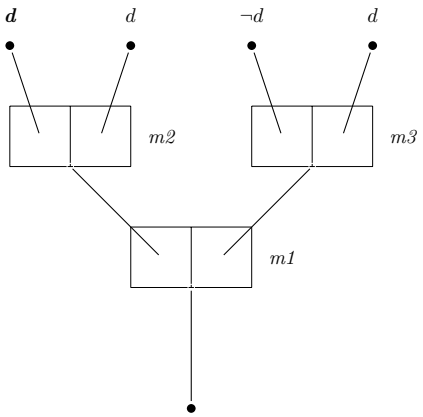
This thesis will create an extension of STIT that provides an alternative to knowledge, namely the interpretation STIT operator, that gives new insight on epistemic problems and that is not troubled by the *epistemic obligation paradox*. Before we will define such a system, it is important to note that there are many other problems that should be accounted for when creating a system that incorporates epistemic and deontic concepts. These problems will be explained with the example of the deserted traveler stated by Broersen[9].

5.1 The deserted traveler

The basis of the example is stated like this:

Example 7. *There are two assassins, $a1$ and $a2$, and a traveler. The traveler will travel tomorrow to the desert and prepared himself by setting up some bottles of water in his tent. At the start of the day $a1$ sneaks into the traveler's tent and poisons the water and sneaks back out. In the afternoon $a2$ sneaks into the tent and removes all the water. The next day the traveler goes into the desert. A few days later the traveler is found dead and the cause death is dehydration.*

The question is who killed the traveler. The frame corresponding to this example can be this:



At $m1$ $a1$ has a choice of poisoning the water or not and at $m2$ and $m3$ $a2$ has the choice of removing the water or not. d is the proposition that represents the traveler being dead and the bold face propositions represent the actual history. The formulas corresponding to the questions are $[a1\ stit : d]$ and $[a2\ stit : d]$. If the first formula holds, $a1$ killed the traveler and if the second formula holds $a2$ killed the traveler.

To first define some discussion points with relation to this example, we will look into the valuation by using the *cstit* and the *dstit* operators.

According to *cstit* both $[a1\ cstit : d]$ and $[a2\ cstit : d]$ hold, therefore both assassins saw to it that the traveler died. An obvious problem with *cstit* is easily visible in example 7. If $a2$ did not remove the water, $[a2\ cstit : d]$ would still hold and $a2$ would have still seen to it that the traveler died. This is not desirable since $a2$ intuitively would have nothing to do with the death of the traveler.

The other option is examining *dstit*. According to *dstit* only $[a1\ dstit : d]$ holds and therefore $a1$ deliberately saw to it that the traveler died. This can also be not intuitively desirable since the actions of $a2$ do not matter even though it seems that $a2$ is a culprit as well. Intuitively you could say that $a2$ also deliberately saw to it that the traveler died.

5.2 Adding knowledge

The problems with *cstit* and *dstit* in the previous example can be cleared up when considering the epistemic states of the two assassins. Let us assume that both assassins are fully aware of the STIT frame. Then, when observing that valuation of the *dstit* formula, it is obvious that $a2$ did not deliberately see to it that the traveler died, since $a2$ already knew at that moment that his actions would have no effect on the situation. For the valuation of *cstit* formulas a similar case is to be made. Assassin $a2$ will always see to it that the traveler died since he had no choice of preventing it. If you assume that $a2$ knew of his inability to prevent the death, it is easier to grasp that everyone already saw to it that the traveler died.

When partial or full lack of knowledge is assumed, the *cstit* and *dstit* operators intuitively do not represent the responsibility of the assassins anymore. When assuming the valuation of the formula $[a2\ dstit : d]$ depends on the statement that $a2$ attempted to deliberately see to it that the traveler died, it is clear that the formula would hold when $a2$ did not know of the actions of $a1$. However, in the normal STIT model this formula does not hold. For *cstit* a similar case is to be made. If $a2$ did not remove the water, it would be intuitively viable that $a2$ did not try to see to it that the traveler died. However, in the STIT model $[a2\ cstit : d]$ does hold.

One must be able to differentiate between full knowledge and partial knowledge.

5.3 Applying epistemic STIT models

One way to solve this it to use the extension from Herzig [13]. In this logic $m2$ and $m3$ can be modeled as indistinguishable for $a2$. This way $a2$ will knowingly see to it that the traveler dies if he removes the water, since $a2$ ensures the death for every possible state of the world. If $a2$ would not remove the water, then he does not knowingly see to it that the traveler lives or dies, since that depends on what $a1$ did. However, this solution cannot solve the problem if you would want that only $a1$ saw to it that the traveler died, which *dstit* can. Combining *dstit* and the logic of Herzig could solve this problem.

However, there are still more statements that are not expressible in the current epistemic STIT models that can be of value. To describe these statements an extension of the *deserted traveler* problem will be used.

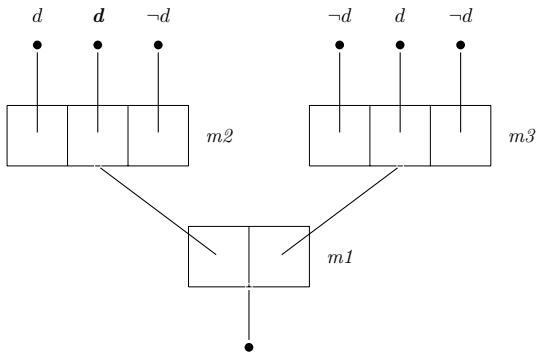
5.4 The extended deserted traveler

Knowledge seems to be an important factor for connecting STIT to responsibility. Being responsible for something without knowing why you can be responsible can be seen unintuitive in some cases. This thesis will state an extended version of the *deserted traveler* problem.

Example 8. *This problem is similar to problem 7. There are two differences from problem 7. The first difference is that a2 has three options instead of two. The first two options are the same as in problem 7. The third option is that a2 replaces the water with new, drinkable water. This option ensures that the traveler will live. Note that a2 does not necessarily know of the existence of his option to replace the water.*

The second difference is that a2 has a change of heart. a2 will go into the tent with the intention to kill the traveler, but decides to not kill him when he is in the tent. This results in a2 picking either the option to do nothing with the water or the option to replace the water.

The corresponding STIT model looks like this:



In this model only $[a2\ cstit : d]$ *and* $[a2\ dstit : d]$ *hold.* $[a1\ cstit : d]$ *and* $[a1\ dstit : d]$ *do not hold.*

In this problem the results are not intuitive. The first assassin can now never see to it that the traveler dies or not, even though his poison killed the traveler. It would be possible to assume that the second assassin saw to it that the traveler died, but only if the second assassin knew that the water was poisoned and that the assassin knew that he had the possibility to replace the water.

5.4.1 The lack of knowledge about a1

If the second assassin did know of the option to replace the water, but did not know that the water was poisoned, then a similar problem occurs as with problem 7. The second assassin does not know that the results for leaving the water and replacing the water are different. According to the assassin, there is no real reason to put in the effort of replacing the water since doing nothing will presumably hold the same result.

Herzig does solve the problem with not knowing what the result of your actions would be. In example 8, if $m2$ and $m3$ are indistinguishable for $a2$ then he would not knowingly see to it that the traveler dies if the water was poisoned and $a2$ did nothing. In addition, assassin $a2$ would knowingly see to the state of the traveler when $a2$ would take action, no matter what $a1$ did.

5.4.2 The lack of knowledge about the possibilities

If $a2$ knew that the water was poisoned, but did not know of the possibility of replacing the water, then it is not intuitive to conclude that the second assassin deliberately saw to it that the traveler died. The water was not poisoned by the second assassin and he did not know that he had the

opportunity to prevent the death. Even though Herzig can deal with the first issue, he cannot deal with the lack of knowledge of the options.

6 The interpretation STIT model

To prevent the epistemic obligation paradox it is not possible to use any isolated knowledge operator. Since the reflexive property of knowledge (if something is known then it is true) is commonly accepted, something else has to be thought of. An alternative solution to this problem could be that knowledge is not modeled anymore. In this thesis an alternative to the epistemic operator will be proposed. This alternative operator is called the *interpretation operator*. This operator is implemented with a STIT model. This new model is called the *Interpretation STIT model*. This model will is able to give some new insight in the problems stated in the previous sections. The formal definition of this model is as follows.

The frame of the new model, called the *interpretation STIT frame*, will contain one constant set of agents and a multitude of interpretations. These interpretations are STIT frames that represent an interpretation of the world. Every interpretation in this frame also has a relation defined that dictates which other interpretation it relates to depending on the agent. This relation also defines how the histories within the interpretations relate to each other. The relation from one interpretation to another loosely means that within one interpretation of the world agents interpret that interpretation in a way that corresponds to the related interpretation. Let us define the new frame as a tuple

$$\langle Agent, N^* \rangle$$

where $Agent$ is a set of agents and N^* is a set of interpretations that are defined by a 4-tuple

$$\langle Tree, \leq, Choice, R_n \rangle$$

where $Tree$ is a set of moments and $\leq$ is a tree-like relation between those moments. $Choice$ is the choice function taking moments, histories (complete branches of moments) and agents and return a set of histories defining the possible histories after a certain choice is made. R_n is the relation unique per interpretation that defines for every agent what their interpretation is of the current interpretation. Formally R_n is a function from $Agent$ to a tuple $\langle N^*, R_h \rangle$ where R_h is a relation from the set of moment, history pairs defined with $Tree$ and $\leq$ to the set of moment, history pairs defined with $Tree', \leq'$ in N^* of the result of R_n. We normally see R_n and R_h as functions, but one could also say that these are serial and deterministic relations. This is clear to see with the axioms defined in section 10.2.

The operators in the new logic are:

$p_1, ..., p_n$	The propositional symbols.
$\alpha_1, ..., \alpha_n$	The agent symbols.
$=$	The equation symbol.
$\neg, \wedge$	The logical operators. (Can be extended as usual.)
$\Box$	The historical necessity symbol.
$[\ dstit :\ ]$	The dstit operator.

$\boxtimes$ The interpretation symbol.

In the new model $M = \langle Agent, N^*, V\rangle$ (where V is a valuation function mapping formulas to histories) formulas are formulated in the following way. Where $n \in N^* = \langle Tree, \leq, Choice, R_n\rangle$ is a STIT frame in the set N^* of STIT frames, h is a maximal length branch in $\langle Tree, \leq\rangle$ and $m \in h$ is a moment in that history:

- $M, m/h/n \vDash p$ iff $m/h/n \in V(p)$, where p is an atomic proposition.
- $M, m/h/n \vDash \alpha = \beta$ iff $V(\alpha) = V(\beta)$, where $\alpha, \beta \in Agent$ are agents.
- $M, m/h/n \vDash \neg A$ iff $M, m/h/n \nvDash A$, where A is any formula.
- $M, m/h/n \vDash A \wedge B$ iff $M, m/h/n \vDash A$ and $M, m/h/n \vDash B$, where A, B are any formula.
- $M, m/h/n \vDash \Box A$ iff $M, m/h'/n \vDash A$ for every h' where $h' \in H_m$, where A is any formula.
- $M, m/h/n \vDash [\alpha\ dstit:\ A]$ iff $M, m/h'/n \vDash A$ for all $h' \in Choice^{\alpha}_{m}(h)$ and $M, m/h''/n \nvDash A$ for some $h'' \in H_m$, where A is any formula and $\alpha \in Agent$ is any agent.
- $M, m/h/n \vDash \boxtimes_{\alpha} A$ iff $M, m'/h'/n' \vDash A$ where $R_n(\alpha) = \langle n', R_h\rangle$ and $R_h(\langle m, h\rangle) = \langle m', h'\rangle$, where A is a formula and $\alpha \in Agent$ is an agent.

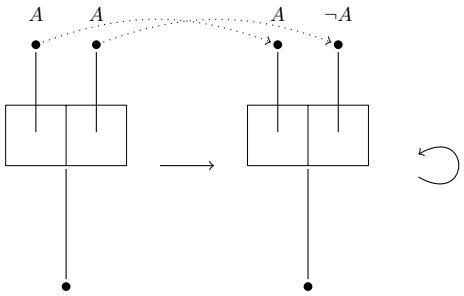
Note that formulas are now evaluated at a moment for a history within one interpretation instead of only being evaluated at a moment and a history as is the case in original STIT.

Later in this thesis another regular STIT operator is used. This operator is the Δ operator. The evaluation of this operator works as follows: $M, m/h/n \vDash \Delta\alpha p$ iff for any $h' \in_n Choice^{m}_{\alpha}(h)$ it holds that $h' \in V(p)$, where p is an atomic proposition and α is an agent. The meaning of this operator is that $\Delta\alpha p$ is true at $m/h/n$ iff p is still possible after the choice of α at m is made.

7 Intuition of the interpretation STIT model

All operators except for the interpretation operator in the interpretation STIT model are evaluated within one interpretation and work in essentially the same way as they would work in a normal STIT model. The interpretation operator works between STIT frames. The relation R_n defines how an agent interprets the current interpretation. This definition of interpreting is a purely physical one. For the interpretation STIT model $\boxtimes_{\alpha} A$ can be read as "α acts as if A". In the models of interpretation STIT the solid arrows between interpretations represent the R_n function and the dotted arrows between the histories represent the R_h function.

Example 9. *Bob tries to catch a train and acts as if the train will be on time. In reality the train will be later. Bob has two options to go to the train station. One is by bus and one is by foot. Bob thinks he will be on time when he takes the bus and Bob thinks that he will be too late when he walks. In reality he will be on time no matter what. The interpretation STIT model for this problem looks like this.*



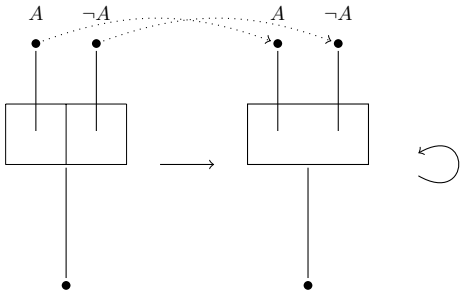
where A stands for the proposition that Bob will be on time, the left history is where Bob takes the bus and the right history is where Bob walks. As is clear to see, Bob acts as if if he would walk, then he will not be on time, but if he were to make that choice, in reality he would be on time.

The interpretation operator brings possibilities to model some new things in relation to normal STIT models.

7.1 Disregard of choice

It is possible to model an agent that acts as if a certain choice does not exist. This modeling can be done for agents that don't expect that their choices matter. Let us look at an alteration on example 9.

Example 10. *As opposed to example 9 Bob now acts as if it does not matter whether he takes the bus or goes walking, because he went from home on time. Bob does not rule out missing the train either. However, the train is early now and he has to take the bus to be on time. Then the model looks like this.*

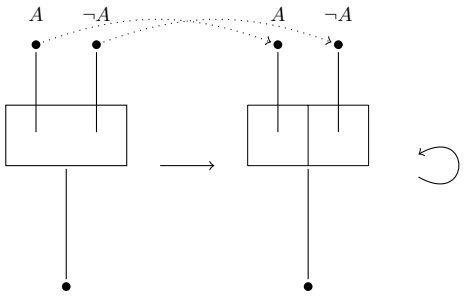


Bob acts as if the choice between the bus and the walking is not a relevant choice for catching the train.

7.2 Illusion of choice

It is possible to model an agent that acts as if a non existent choice exists.

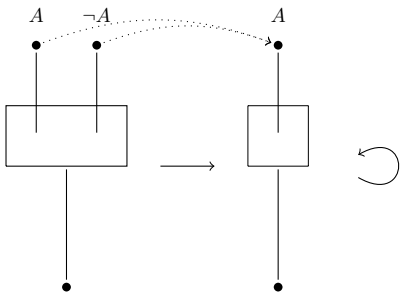
Example 11. *As oppose to example 9 it is completely out of Bob's power whether he is on time or not. The train can be early or late, but whether Bob take the bus or not does not matter. Bob acts as if that choice does matter. The following model represents this situation.*



7.3 Disregard of possibilities

It is possible to model an agent that acts as if a certain history does not exist. This can be used to model agents who don't expect certain outcomes from an action.

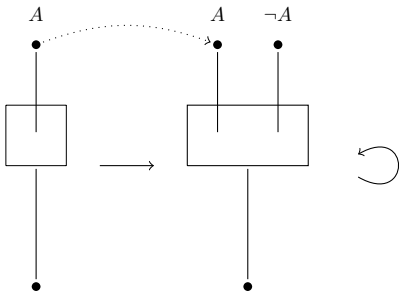
Example 12. *Consider a similar example as example 9. The difference is Bob takes the bus regardless and acts as if he will be on time. However, in reality it is possible that the bus breaks down and Bob will not be on time. Bob does not take this into account. A situation like this can be modeled like this.*



7.4 Illusion of possibilities

It is possible to model an agent that acts as if certain histories exist even though they are impossible in reality.

Example 13. *Let us consider an example similar to example 12. The difference in this example is that now Bob acts as if the bus can break down. However, for the sake of this example, the bus cannot break down. This situation can be modeled in this way.*

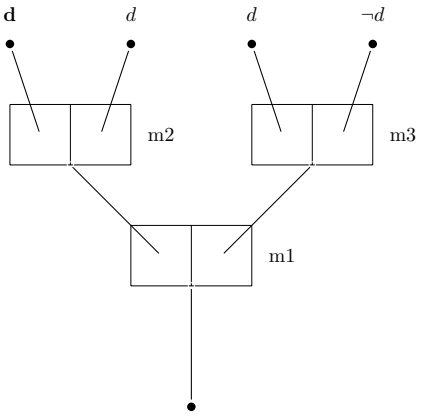


8 An interpretation STIT model view on some problems

The Interpretation STIT model is a moderately large extension on the basic STIT frame. It can incorporate the world view of agents to model more complex situations more informatively. In section 5.1 the problem of the deserted traveler was introduced in this thesis. In that section a few problems with relation to basic STIT were introduced. The Interpretation STIT model has an alternative point of view for these problems.

8.1 Broersen's deserted traveler

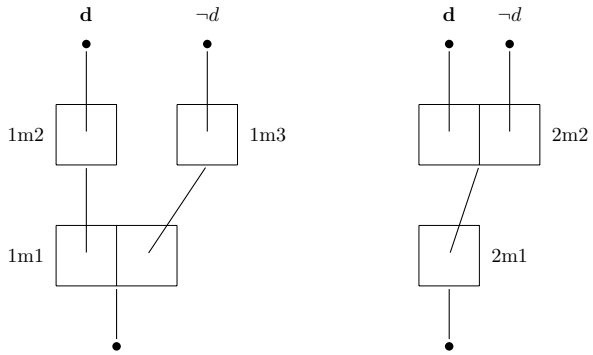
In example 7 a problem was defined where two assassins attempted to kill a desert traveler. The problem in this example was that the first assassins could not be seeing to it that the traveler died without the second assassin always seeing to it that the traveler died. Since the Interpretation STIT model does not change the workings of an individual STIT frame, it is not possible to solve this specific problem with it. However, it is possible to state whether the assassins themselves acted as if they saw to it that the traveler died. The original STIT frame will therefore be in the Interpretation STIT:



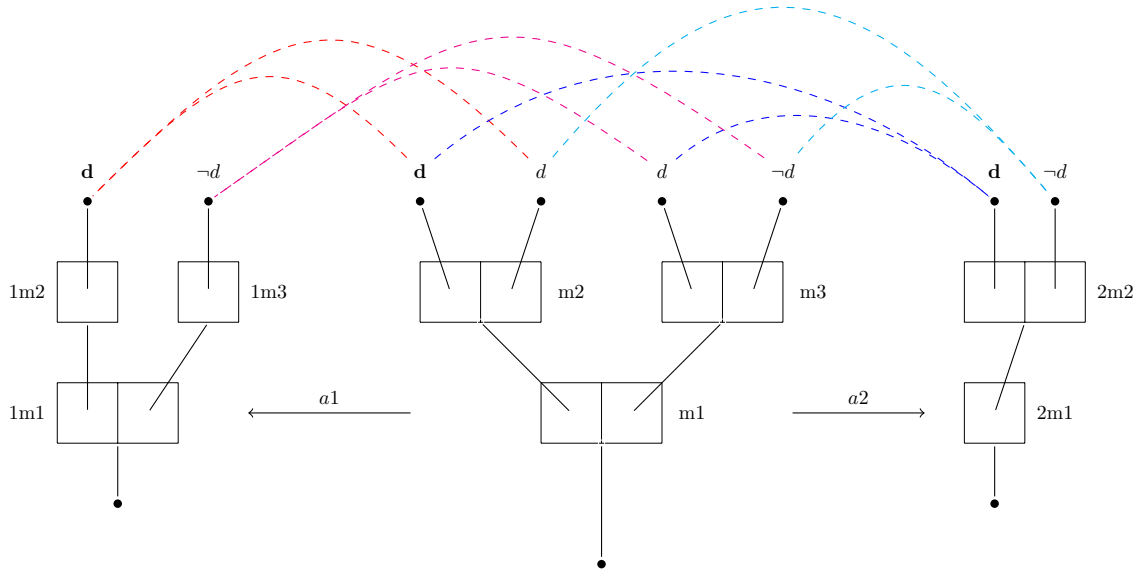
For clarification, in this frame and all the following frames in the section the bottom choices will always be the choices assassin *a1* can make, namely a subset of the choices of poisoning the water

and removing the water. The top choices will always be choices that agent *a2* can make, namely removing the water, doing nothing and replacing the water.

To extend this model to an Interpretation STIT model, the interpretation of the two assassins must be modeled in. To do this it first must be stated what misinterpretations the assassins have of the situation. In this model it will be assumed that both assassins have a disregard of the choices of the other assassin and have a disregard of the possibilities that come from those choices. The STIT frames according to assassins $a1$ and $a2$ are then respectively:



In the interpretations of the situation of the assassins it is clear to see that $[a1\ cstit : d]$ and $[a2\ cstit : d]$ will hold in their STIT models respectively if they decided to kill the traveler. It also holds that $[a1\ cstit : \neg d]$ and $[a2\ cstit : \neg d]$ will hold in the assassins' STIT models respectively if they decided not to kill the traveler. So to create the complete Interpretation STIT model the three separate models can be combined. (The relations between histories are colored to make it more easily readable):

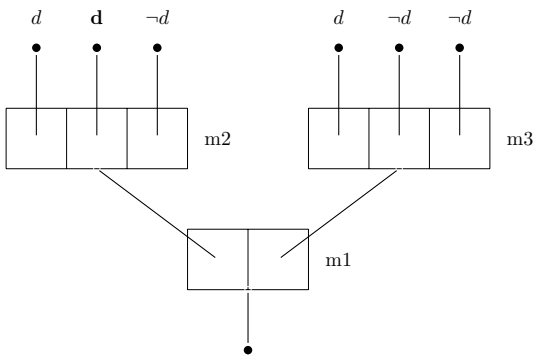


Even though in the reality (the center STIT frame) nothing has changed, the interpretations of the agents are clearer. Assuming $a1$ poisoned the water and $a2$ did not remove the water, $a2$ would be technically seeing to it (in the *cstit* sense) that the traveler died, but $a2$ acted as if he did not see to it. So in formulas $[a2\ cstit : d]$ and $\boxtimes_{a2}\neg[a2\ cstit : d]$ are true. This admittedly does not solve the problem that you would assume that $a2$ had nothing to do with the death of the traveler, but it does give some extra information to determine the responsibilities of $a1$ and $a2$. Now a question can be asked like *Is the second assassin responsible for the death of the traveler if the second assassin did see to it that the traveler died but acted as if he saw to it that the traveler did not die?*

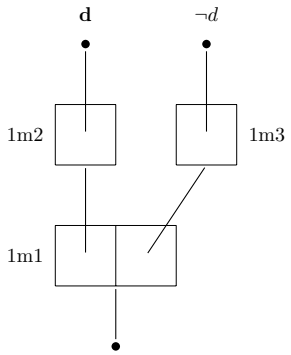
The other problem can now be interpreted as follows. Assuming both assassins tried to kill the traveler $a2$ technically did not deliberately see to it that the traveler died, but $a2$ acted as if he deliberately saw to it that the traveler died. So in formulas $\neg[a2\ dstit : d]$ and $\boxtimes_{a2}\neg[a2\ dstit : d]$. This, again, does not solve the problem that you would assume that the second assassin did deliberately see to it that the traveler died. However, another question can be asked about responsibility: *How responsible is the second assassin for the death of the traveler if he (not deliberately) saw to it that the traveler died and acted as if he deliberately saw to it that the traveler died?*

8.2 The extended deserted traveler

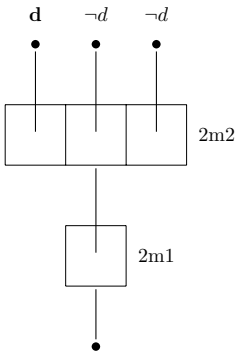
The problems of the extended traveler given in section 5.4 can be interpreted by creating a interpretation STIT model of the situation. The extended deserted traveler problem assumed that $a1$ had the possibilities of poisoning the water and doing nothing and $a2$ had the possibilities of removing the water, doing nothing and replacing the water. The basic STIT model for this problem was this:



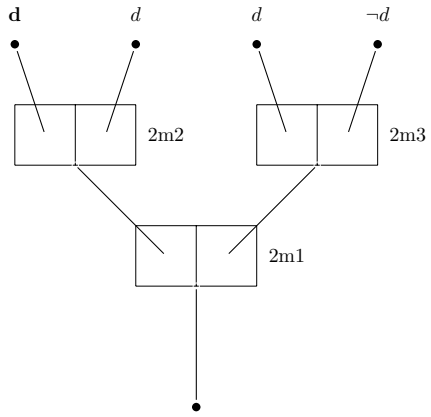
With the interpretation STIT model a difference can be made between $a2$ knowing and not knowing about the actions of $a1$ and between knowing and not knowing the existence of the possibility of removing the water. In all variations of the example the interpretation of $a1$ will not change. This interpretation will still be as follows:



Assassin $a1$ will still disregard all the choices and possibilities made by $a2$, since it is assumed that $a1$ does not know of $a2$. The interpretation of $a2$ will be different depending on the variation of the example. When $a2$ knows of the actions of $a1$ and knows of the possibility of replacing the water then the interpretation of $a2$ will be the same as the basic STIT model. If $a2$ does not know of the actions of $a1$ then the interpretation will look as follows:

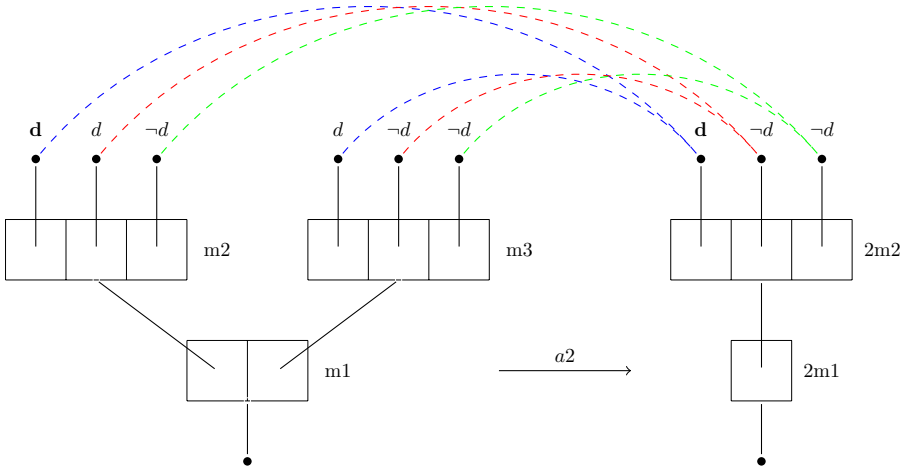


Here $a2$ disregards all choices and possibilities made by $a1$ since $a1$ only has one choice in the interpretation, and then $a2$ acts as if he can make one of the three choices described previously. If $a2$ does not know of the possibility to replace the water but does know whether the water is poisoned or not, then the interpretation will be the same as the basic STIT model of the previous section, namely:



8.2.1 The lack of knowledge about a1

If the $a2$ did know of the possibility of replacing the water and not about the existence of $a1$, then, according to $a2$, there should be no difference between replacing the water and doing nothing. This lack of difference can easily be modeled with an Interpretation STIT model as follows (the interpretation of $a1$ is omitted since it does not play a role in this variation):

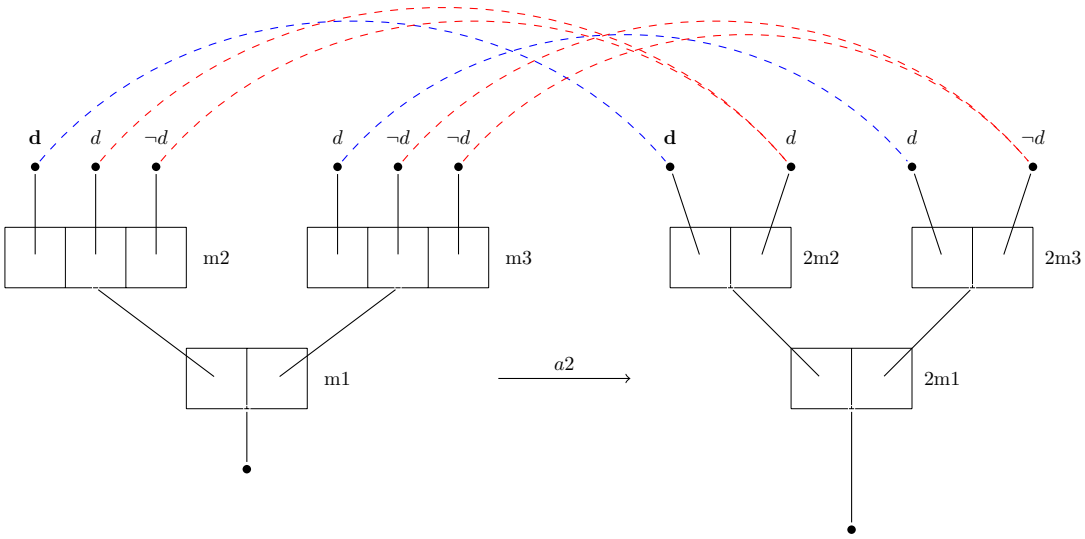


In this model it is clear to see that $a2$ acted as if replacing the water and doing nothing yielded the same results, namely that the traveler did not die. So if $a2$ did nothing then the formula's

$[a2\ dstit : d]$ and $\boxtimes_{a2}[a2\ dstit : \neg a2]$ are true. This means that even though $a2$ technically deliberately saw to it that the traveler died, he acted as if he deliberately did not see to it that the traveler died.

8.2.2 The lack of knowledge about the possibilities

If $a2$ knew the water was poisoned, but did not know of the possibility to replace the water, then $a2$ should not deliberately see to it that the traveler died if $a2$ chooses to remove the water. The corresponding Interpretation STIT model for this problem is as follows:



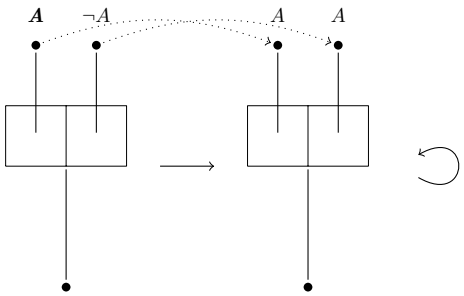
Assassin $a2$ now has a disregard of the choice of replacing the water and therefore acted as if that choice was not there. In this model the formulas $[a2\ dstit : d]$ and $\boxtimes_{a2}\neg[a2\ dstit : d]$ are true which means that $a2$ technically deliberately saw to it that the traveler died, but acted as if he did not deliberately saw to it.

8.3 Deliberation

As seen in the previous sections, the semantic definition of *dstit* does not have an intuitive meaning anymore for some of the uses of the operators. These operators are in the interpretation that did not represent an agent, but represented the actual situation. The original meaning of "deliberately seeing to it that" is that an agent sees to it and that the agent had the possibility of not seeing to it. This meaning can be misleading in Interpretation STIT since it is possible to "deliberately see to it" without noticing that you are seeing to it deliberately. The following alteration of example 9 will clarify this problem.

Example 14. *Bob tries to catch a train. Since the train is always late Bob acts as if the train will be late this time again. He has two possibilities to get to the train, namely by bus or by foot.*

However, this time the train will be on time for once and Bob will not make it if he goes by foot. Incidentally, Bob acts as if it does not matter whether he goes by bus or by foot but eventually decides to go by bus. The Interpretation STIT model for this problem looks like this:



where A stands for the proposition that Bob will be on time, the left history is where Bob takes the bus and the right history is where Bob walks. In this model $[Bob\ dstit : A]$ *holds while according to Bob it was inevitable that he would be on time, since* $\boxtimes_{Bob}\neg[Bob\ dstit : A]$ *holds.*

This example assumes that Bob deliberately sees to it that he was on time, but intuitively Bob never deliberately saw to anything, since Bob assumed he would be on time anyway.

In Interpretation STIT the original notion of "deliberately seeing to it that" can better be interpreted as "avoidably seeing to it" since the deliberations of an agent can now be defined in terms of the interpretation of the agent which is more intuitive. An alternative for the notion "deliberately seeing to it that" could divide it in two parts. One part could be $\boxtimes_\alpha[\alpha\ dstit : A]$ which describes "acting as if you avoidably see to it that." This part would describe the intentional nature of deliberation. The other part of the definition then would enforce the reality of the notion which is $[\alpha\ cstit : A]$. Intuitively "deliberately seeing to it that" does imply that you are actually seeing to it. The full new definition of "deliberately seeing to it that" would then be $\boxtimes_\alpha[\alpha\ dstit : A] \wedge [\alpha\ cstit : A]$. This new definition of "deliberately seeing to it that" can now solve a problem with the original *deserted traveler problem*, in that both assassins can now deliberately see to it that the traveler died, even though the second assassin could not change the destiny of the traveler.

9 Knowledge in Interpretation STIT

The interpretation STIT model does not have an axiomatically defined notion of knowledge. The intuition operator resembles versions of doxastic operators closely but it is not to be interpreted that way. The main difference between the interpretation operator and belief is that the semantic relation represented by the interpretation operator is deterministic. This property is not commonly given to notions of belief. It is also not intuitive to think that belief is deterministic as how belief is interpreted in standard doxastic modal logics, since people can believe different propositions. Your interpretation, however, is intuitively deterministic. You only have one interpretation of the world around you. There can be uncertainties in your interpretation but that uncertainty is within one interpretation.

The notion of interpretation as used in this STIT model is also not to be confused with your internal reasoning. Since STIT is based on the consequences of choices, interpretation is to be seen

as on what basis an agent makes choices or would have made choices. A formula such as $\boxtimes_\beta p$ is to be interpreted as *agent β made choices as if p were true*. This can entail that agent β believed that p were true, but is does not have to be so. It can also be that agent β took p as a given truth and never thinks about it.

An interpretation operator such as this one can still be very useful even though it does not directly define conscious states of agents. A lot of definitions that seem to require a defined notion of knowledge can be modeled with an interpretation STIT model. Some examples are the modes of *mens rea*.

9.1 Interpretation STIT and the modes of mens rea

As described in prior parts of this paper Broersen [7] gives interpretations of the modes of mens rea using STIT and a knowledge operator. However, the modes of mens rea could also be defined in the interpretation STIT model. A big difference is that the actual outcome of an action is separate from the intention of the action. This can be a favorable property since judicially intent can be separated from the results of an action.

9.1.1 Acting recklessly

Prior defined in section 4.2 that you are acting recklessly if you do not know that you see to it that a proposition p holds. In interpretation STIT this can be defined as:

$$([\beta\ stit\ :\ p] \wedge (\boxtimes_\beta(\neg[\beta\ stit\ :\ p] \wedge \neg[\beta\ stit\ :\ \neg p]))) \rightarrow [\beta\ stit\ :\ V]$$

where V denotes a violation the same way as in section 4.2. So $[\beta\ stit\ :\ V]$ has the meaning that β acted recklessly.

This formula is interpreted as *if agent β sees to it that p and acted as if he neither sees to it that p nor sees to it that not p, then he acted recklessly*. This interpretation does not describe anything about knowledge but it can be intuitively derived from it. The statement that β acted as if the value of p was independent of his actions can imply that β did not know that he saw to it that p held.

It is also clear to see that reality and the interpretation of agent β are separated. An advantage for this is that a different interpretation of the notion of acting recklessly where the actual outcome of actions does not matter can easily be defined in the interpretation STIT model.

9.1.2 Knowingly risking

Prior defined in section 4.2 that you are knowingly risking if you know that you are not seeing to it that a proposition p holds. In interpretation STIT this can be defined as:

$$(\boxtimes_\beta \neg[\beta\ stit\ :\ \neg p]) \rightarrow [\beta\ stit\ :\ V]$$

This formula is interpreted as *if agent β acted as if he does not see to it that not p, then he is knowingly risking*. In this formula it is clear to see that knowingly risking p does not depend on the actual value of p. One might say that p must be possible to be knowingly risking it. The current formula does not enforce that but it can be added. A definition of that form can be $(\Delta\beta p \wedge \boxtimes_\beta \neg[\beta\ stit\ :\ \neg p]) \rightarrow [\beta\ stit\ :\ V]$. This addition to the formula ensures that p must have been possible in the choice of β before the agent can knowingly risk something.

9.1.3 Acting knowingly

Prior defined in section 4.2 that you are acting knowingly if you know that you are seeing to it that a proposition p holds. In interpretation STIT this can be defined as:

$$([\beta \ stit \ : \ \neg p] \wedge (\boxtimes_\beta [\beta \ stit \ : \ \neg p])) \rightarrow [\beta \ stit \ : \ V]$$

This formula is interpreted as *if agent β sees to it that p and acted as if he does see to it that p, then he is acting knowingly.* In this definition knowledge of actions is a result of correct expectations. If an agent acts as if he sees to that something holds and it holds then it can be considered that the agent knew that he saw to it.

9.1.4 Other possible definitions

In the definition of section 9.1.3 it is clear to see that the reality of seeing to it is separate from the interpretation of the agent. The omission of the reality of seeing to it would still be an interesting statement.

$$(\boxtimes_\beta [\beta \ stit \ : \ \neg p]) \rightarrow [\beta \ stit \ : \ V]$$

Such a statement would describe a violation based on bad intent regardless of the outcome. In judicial systems formulas like these can be used for describing already existing felonies such as the intent of killing someone.

9.2 Defining knowledge

As stated before, the interpretation operator is not to be interpreted as either belief or knowledge, but rather as a conscious and subconscious state of an agent. The interpretation operator defines statements according to the state of one agent interpreted in the state of another. Defining knowledge with such an operator obviously lends itself to a description of knowledge where no one actually has absolute knowledge. The only knowledge that exists is contextual and based on an agent's interpretation of another agent's interpretation of a certain subject. Assuming agent α acts as if p is true and assuming agent α acts as if agent β acts as if p is true. According to agent α, it can be assumed that β has knowledge of p, since p is true and it seems that β thinks that p is true. In reality p might be false, but that does not seem to matter for the knowledge of β according to the ideas of α.

Another possibility of defining knowledge with the interpretation operator is to assume that there is a reality and all agents have an interpretation of that reality. If agent α would be replaced by some agent that represents nature or all determined events, then you could say that β has knowledge of p since p is true in reality and β acts as if p is true. An interpretation of knowledge based on a "reality interpretation" already bears some similarities with the *justified true belief* interpretation of knowledge. The interpretation operator can also be used to define a similar version of such a notion of knowledge.

9.2.1 Knowledge as True "Belief"

In the previous paragraph an example was given in that an agent knows p if an agent acts as if p is true and p is true in reality. This notion of knowledge bears resemblance to the definition of knowledge as *true belief*. The difference is that the interpretation operator is stronger than the

doxastic operator in the way that it is harder to acquire an interpretation of something than to believe it. A definition of a knowledge operator based on this definition looks like this:

$$K_\alpha p \leftrightarrow p \wedge \boxtimes_\alpha p$$

However, occurrences of epistemic luck could still be present when using a notion of knowledge as *true interpretation*. The reasoning behind the interpretation of something can still be different from the actual reason.

To solve the problem of epistemic luck an idea of justification must be considered. One could think that justification can be implemented with the interpretation operator very intuitively, since the intuition STIT frame already gives insight in the reasoning of an agent by representing the interpretation an agents has of a situation. Justification could be described by the prevention of the cases given in sections 7.1, 7.2, 7.3 and 7.4. This would then state that " an agent is justified if it does not have any disregard of choice, illusion of choice, disregard of possibilities or illusion of possibilities." All these cases can be excluded separately as a part of the definition of justification.

9.2.2 Disregard of choice

One could say that an agent should be aware of what mattered for something to be true before knowing it. If this were not true, then an agent might not accurately assume the causation of propositions. In other words, an agent should have no disregard of the choices made for a certain statement to know it. This can be expressed in the following way. An agent α has no disregard of choice with relation to statement A if and only if for all $\beta \in Agents$ it holds that

$$\Box([\beta\ cstit : A] \rightarrow \boxtimes_\alpha[\beta\ cstit : A])$$

This formula states that an agent α should act in any case as if some agent β saw to it that A if β actually saw to it. This results in that α is correct for all the choices relevant for A.

9.2.3 Illusion of choice

One could say that an agent should be aware of what did not matter for something to be true before knowing it. Similar to the previous section, if this were not true, then an agent might not accurately assume the causation of propositions. One could say that if something happened randomly, then you should be aware that it happened randomly if you pretend to know it. This can be expressed in the following way. An agent α has no disregard of choice with relation to statement A if and only if for all $\beta \in Agents$ it holds that

$$\Box(\boxtimes_\alpha[\beta\ cstit : A] \rightarrow [\beta\ cstit : A])$$

This formula states that if an agent α acts in any case as if some agent β saw to it that A, then it is true that β saw to it that A. This has the result that α is correct for all irrelevant choices for A.

9.2.4 Disregard of possibilities

One could say that an agent should be aware of all the possibilities that there were with regard to a certain statement before knowing it. This ties in most with excluding moral luck. If one does not have a disregard of possibilities then one acts as if there was no other option for something to

happen if there was no other option in reality. Excluding disregard of possibilities can be expressed in the following way. An agent α has no disregard of possibilities with relation to statement A if and only if it holds that

$$\Diamond A \rightarrow \boxtimes_\alpha \Diamond A$$

This formula states that if A is a possible outcome then α acts as if A is a possible outcome. The result is that α is correct about the possibility of the outcome of A.

9.2.5 Illusion of possibilities

One could say that an agent should not act as if there were other possibilities if there in fact were not any. This can be expressed in the following way. An agent α has no illusion of possibilities with relation to statement A if and only if it holds that

$$\boxtimes_\alpha \Diamond A \rightarrow \Diamond A$$

This formula states that if α acts as if A is possible then A must be possible. This results in that an agent α cannot act as if A if A could not have happened.

9.2.6 Knowledge as "justified" true "belief"

One of the stronger notions of justification is the inclusion of all four previously described formulas. If all formulas are included then the definition of knowledge would be this:

$$K_\alpha A \leftrightarrow ((\boxtimes_\alpha \diamond A \leftrightarrow \diamond A) \wedge (\Box(\boxtimes_\alpha [\beta \ cstit : A] \leftrightarrow [\beta \ cstit : A])))$$

This definition of knowledge has no term for truthfulness A or for interpretation $\boxtimes_\alpha A$ anymore. These two terms are already implied in the formula since it is necessary that the structure of the interpretation of α is identical to the real interpretation with regard to A. If a choice enforces, permits or removes the possibility of A then it must be interpreted in that way for an agent to know A. This results in a very strict definition of knowledge, since it requires the knowledge of all impacts that choices can have on a certain proposition. However, it is possible to define an even stronger notion of knowledge. One might think that it is necessary to know about other things besides A. Such a requirement is believable when thinking about the relations between certain propositions. If B bi-implies A is it then required for justification that you also know that B? To include this requirement in the notion of knowledge the complete interpretation of an agent must be identical to the real interpretation.

10 Soundness and Completeness

We prove soundness and completeness by extending Xu's[22] proof of dstit. This means that the the axiom scheme for the *interpretation STIT model* will be an extension of the axiom scheme from Xu. The proof will also make use of the more complex *dstit* instead of *cstit*. However, it is trivial that the *interpretation STIT model* will be sound and complete when using *cstit* if the model is sound an complete when using *dstit*.

10.1 Xu's axiom schema

For convenience for the *interpretation STIT model*, the axiom scheme for all operators besides the interpretation operator $\boxtimes$ will be used from Xu[22]. This is the axiom scheme from Xu for *dstit*:

The logic of the interpretation STIT model is defined by an extension of the axioms defined by Xu:

A0 $\Box(A \to B) \to (\Box A \to \Box B), \Box A \to A, \diamond A \to \Box \diamond A$

A1 $\Delta\alpha(A \to B) \to (\Delta\alpha A \to \Delta\alpha B), \Delta\alpha A \to A, \neg\Delta\alpha A \to \Delta\alpha\neg\Delta\alpha A$

A2 $[\alpha \ dstit: \ A] \leftrightarrow \neg\Box A \wedge \Delta\alpha A$

A3 $\alpha = \alpha, \alpha = \beta \to \beta = \alpha, \alpha = \beta \wedge \beta = \gamma \to \alpha = \gamma$

A4 $\alpha = \beta \to (A \to A(\alpha/\beta))$ where $A(\alpha/\beta)$ is any formula obtained from A by replacing some or all occurrences of α with β.

AIA_k $diff(\beta_0, ..., \beta_k) \wedge \diamond\Delta\beta_0 B_0 \wedge ... \wedge \diamond\Delta\beta_k B_k \to (\Delta\beta_0 B_0 \wedge ... \wedge \Delta\beta_k B_k)$ where $k \geq 1$.

APC_n $\diamond\Delta\alpha A_1 \wedge \diamond(\neg A_1 \wedge \Delta\alpha A_2) \wedge ... \wedge \diamond(\neg A_1 \wedge ... \wedge \neg A_{n-1} \wedge \Delta\alpha A_n) \to A_1 \vee ... \vee A_n$

The inference rules are as follows:

R1 Modus ponens

R2 From A to infer $\Box A$.[5]

10.2 Axioms for the interpretation operator

The interpretation operator is a modal operator. This means that the axioms can be chosen by defining modal properties to the operator. Since the interpretation operator is bound by a function in the interpretation STIT model, the operator must have that functional property as well. For relations this means that the operator is deterministic and serial. Adding these two properties results in the following addition to Xu's axiom scheme:

A5 $\boxtimes_\alpha(A \to B) \to (\boxtimes_\alpha A \to \boxtimes_\alpha B)$

This axiom is intuitively an obvious one: *If you act as if B is true when A is true, then it also holds that if you act as if A is true, then you act as if B is true.* This axiom is a modal one, namely the K axiom.

A6 $\boxtimes_\alpha A \to \neg \boxtimes_\alpha \neg A$

This is the seriality axiom and states that *If you act as if A is true, then you do not act as if A is not true.* This axiom assumes that agent modeled in the intuition STIT model do not contradict themselves and that agents always act according to some intuition and do not act randomly without basis.

[5] When A is valid in the theorem, then $\Box A$ is as well.

A7 $\neg \boxtimes_\alpha \neg A \rightarrow \boxtimes_\alpha A$

This axiom is not that intuitive and requires some argumentation. The axiom states that *if you do not act as if A is false, then you act as if A is true.* The unintuitive part of this axiom is that it prevents agents from acting on the basis of a not fully described world. This axiom separates the intuition sense of "acting as if" from believing.

Example 15. *Assume that you don't know that Turkmenistan is a country. It can then be assumed that you also do not know that capital of Turkmenistan is Ashgabat. Are you then acting as if Ashgabat is the capital or are you acting as if it is not? Intuitively one would say that you are neither believing that Ashgabat is the capital or not. However, axiom* ***A7*** *states that you must be acting as if Ashgabat is the capital or it is not.*

How the interpretation sense of "acting as if" should be interpreted for cases like these is that if it would start to influence your choices you would need to know about it. Then if you neither believe that it is the capital or not you would still have an assumption about it and base your actions on that. This seems like a complex definition for the axiom, but in practice problems like this will not need solving. It does not matter whether someone acts as if A is true or not, if A does not influence the choices someone makes.

A8 $\boxtimes_\alpha \Box A \rightarrow \Box \boxtimes_\alpha A$

This axiom states that *if you act as if A is inevitable, then it is inevitable that you act as if A is true.* This axiom enforces the independence between the choices made and the interpretations. It should not be possible that the interpretation of the world can change based on the choices made. This axiom prevents undesirable statements such as that someone acts as if something is inevitable, but that it is possible that he can act as if something is false.

The extra inference rule for the $\boxtimes$ operator is the following:

R3 From A to infer $\boxtimes A$.

These axioms are chosen to model a rigid and complete world view. Other axioms can be used to describe more flexible and accepting notion of interpretation where choices can influence the interpretations of agents. These new possible sets of axioms would require a completely new form of the interpretation STIT model which will not be discussed in this thesis.

10.3 Soundness

The validity of A1 to A5, APC_n and AIA_k is already proven by Xu[22]. The same proofs hold for the interpretation STIT model as well. The next section will prove the validity of the new axioms $A5, A6$ and $A7$.

Lemma 1. $\boxtimes_\alpha(A \rightarrow B) \rightarrow (\boxtimes_\alpha A \rightarrow \boxtimes_\alpha B)$ *is valid in any interpretation STIT model.*

Proof. Let M be any interpretation STIT model and let n be any interpretation in M. Finally let h and $m \in h$ be any history and moment in n. Then we have to prove that $M, m/h/n \vDash \boxtimes_\alpha(A \rightarrow B) \rightarrow (\boxtimes_\alpha A \rightarrow \boxtimes_\alpha B)$ where A and B are any syntactically correct formula and α is any agent in M. Assume $M, m/h/n \vDash \boxtimes_\alpha(A \rightarrow B)$. Then we have to prove that $M, m/h/n \vDash \boxtimes_\alpha A \rightarrow \boxtimes_\alpha B$ holds. Assume that $M, m/h/n \vDash \boxtimes_\alpha A$. Then we have to prove that $M, m/h/n \vDash \boxtimes_\alpha B$. Since

$M, m/h/n \vDash \boxtimes_\alpha(A \to B)$ and $M, m/h/n \vDash \boxtimes_\alpha A$ hold, by definition of the interpretation STIT model it also holds that $M, m'/h'/n' \vDash A \to B$ and $M, m'/h'/n' \vDash A$ where $R_n(\alpha) = \langle n', R_h\rangle$ and $R_h(\langle m, h\rangle) = \langle m', h'\rangle$. Since $M, m'/h'/n' \vDash A \to B$ and $M, m'/h'/n' \vDash A$ hold, by definition of implication $M, m'/h'/n' \vDash B$ where $R_n(\alpha) = \langle n', R_h\rangle$ and $R_h(\langle m, h\rangle) = \langle m', h'\rangle$. Hence, by definition of the interpretation STIT model $M, m/h/n \vDash \boxtimes_\alpha B$ holds. □

Lemma 2. *$\boxtimes_\alpha A \to \neg \boxtimes_\alpha \neg A$ is valid in any interpretation STIT model.*

Proof. Let M be any interpretation STIT model and let n be any interpretation in M. Finally let h and $m \in h$ be any history and moment in n. Then we have to prove that $M, m/h/n \vDash \boxtimes_\alpha A \to \neg \boxtimes_\alpha \neg A$ where A is any syntactically correct formula and α is any agent in M. Assume $M, m/h/n \vDash \boxtimes_\alpha A$. Then we have to prove that $M, m/h/n \vDash \neg \boxtimes_\alpha \neg A$. Since $M, m/h/n \vDash \boxtimes_\alpha A$, by definition of the interpretation STIT model it also holds that $M, m'/h'/n' \vDash A$ where $R_n(\alpha) = \langle n', R_h\rangle$ and $R_h(\langle m, h\rangle) = \langle m', h'\rangle$. Since $M, m'/h'/n' \vDash A$, it also holds that $M, m'/h'/n' \nvDash \neg A$. Hence $M, m/h/n \nvDash \boxtimes_\alpha \neg A$ holds. Therefore, $M, m/h/n \vDash \neg \boxtimes_\alpha \neg A$ holds. □

Lemma 3. *$\neg \boxtimes_\alpha \neg A \to \boxtimes_\alpha A$ is valid in any interpretation STIT model.*

Proof. Let M be any interpretation STIT model and let n be any interpretation in M. Finally let h and $m \in h$ be any history and moment in n. Then we have to prove that $M, m/h/n \vDash \neg\boxtimes_\alpha \neg A \to \boxtimes_\alpha A$ where A are any syntactically correct formula and α is any agent in M. Assume $M, m/h/n \vDash \neg \boxtimes_\alpha \neg A$. Then we have to prove that $M, m/h/n \vDash \boxtimes_\alpha A$. Since $M, m/h/n \vDash \neg \boxtimes_\alpha \neg A$ holds, by definition of negation $M, m/h/n \nvDash \boxtimes_\alpha \neg A$ holds. Then, by definition of the interpretation STIT model it also holds that $M, m'/h'/n' \nvDash \neg A$ where $R_n(\alpha) = \langle n', R_h\rangle$ and $R_h(\langle m, h\rangle) = \langle m', h'\rangle$. By definition of negation, it also holds that $M, m'/h'/n' \vDash A$. Hence, by definition of the interpretation operator $M, m/h/n \vDash \boxtimes_\alpha A$. □

Lemma 4. *$\boxtimes_\alpha \Box A \to \Box \boxtimes_\alpha A$ is valid in any interpretation STIT model.*

Proof. Let M be any interpretation STIT model and let n be any interpretation in M. Finally let h and $m \in h$ be any history and moment in n. Then we have to prove that $M, m/h/n \vDash \boxtimes_\alpha \Box A \to \Box \boxtimes_\alpha A$ where A are any syntactically correct formula and α is any agent in M. Assume that $M, m/h/n \vDash \boxtimes_\alpha \Box A$. Then we have to prove that $M, m/h/n \vDash \Box \boxtimes_\alpha A$. So by definition of the $\Box$ operator, we have to prove that $M, m/h'/n \vDash \boxtimes_\alpha A$ for every h' where $h' \in H_m$. Since $M, m/h/n \vDash \boxtimes_\alpha \Box A$, by definition of the interpretation STIT model it also holds that $M, m''/h''/n'' \vDash \Box A$ where $R_n(\alpha) = \langle n'', R_h\rangle$ and $R_h(\langle m, h\rangle) = \langle m'', h''\rangle$. Since $M, m''/h''/n'' \vDash \Box A$, by definition of the $\Box$ operator, it holds that $M, m''/h'''/n'' \vDash A$ for every h''' where $h''' \in H_{m'}$. Since $M, m''/h'''/n'' \vDash A$ holds for every h''' where $h''' \in H_{m''}$ and it holds that $R_n(\alpha) = \langle n'', R_h\rangle$ and $R_h(\langle m, h\rangle) = \langle m'', h''\rangle$, it also holds that for every $R_h(\langle m, h'\rangle) = \langle m'', h'''\rangle$ it is so that $M, m''/h'''/n'' \vDash A$. Therefore $M, m/h'/n \vDash \boxtimes_\alpha A$ holds for every h' where $h' \in H_m$. □

Theorem 1. *The axiom scheme from section 10.1 and 10.2 is sound with respect to the interpretation STIT frame. Hence, for every consistent set of formulas Γ and any formula ϕ it holds that if ϕ is derivable from Γ according to the axioms from section 10.1 and 10.2, then ϕ is derivable from Γ in the interpretation STIT frame.*

Proof. Since all axioms in the axiom scheme from section 10.1 and 10.2 are valid in the interpretation STIT model, it follows that for every consistent set of formulas Γ and any formula ϕ it holds that if ϕ is derivable from Γ according to the axioms from section 10.1 and 10.2, then ϕ is derivable from Γ in the interpretation STIT frame. □

10.4 Completeness

The original canonical frame Xu[22] uses to prove completeness for *dstit* is

$$\langle X, [\beta_0], [\beta_1], ..., R_{[\beta_0]}, R_{[\beta_1]}, ...\rangle$$

- where X is an equivalence class of R_L. R_L is the relation on W_L, the set of all maximally consistent sets of any language L, where for each $w, w' \in W_L, wR_Lw'$ iff $\{A : \Box A \in w\} \subseteq w'$.
- where $[\beta_0], [\beta_1], ...$ are $\cong$-equivalence classes which $\beta_0, \beta_1, ...$ belong to respectively. $\cong$ is a relation where for each agent symbols α, β, $\alpha \cong \beta$ iff $\alpha = \beta \in w$ for all $w \in X$.
- where $R_{[\beta_0]}, R_{[\beta_1]}, ...$ are relations of agent terms on X such that for every $w, w' \in X$, $wR_{[\beta_0]}w'$ iff $\{A : \Delta\beta_0 A \in w\} \subseteq w'$.

10.4.1 Canonical Frame

Let $L = L_n$ for $n \geq 0$ be the axiomatic system defined by the axioms described in section 10.1. Let W_L be the set of all maximally consistent sets for L. Let R_L be a relation from W_L to itself where for each $w_1, w_2 \in W_L$ it holds that $w_1R_lw_2 \in R_L$ iff $\{A : \Box A \in w_1\} \subseteq w_2$. This is proven to be an equivalence relation[22]. Let X be any R_L equivalence class. X is a universal relation. Let $\cong$ be a relation between agents where $\alpha \cong \beta$ iff $\alpha = \beta \in w$ and $\alpha \ncong \beta$ iff $\alpha \neq \beta \in w$. By $A4$ $\cong$ is an equivalence relation. Let $[\alpha]$ denote the term used for the set of all agents α' where $\alpha \cong \alpha'$. Let $\beta_1, \beta_2, ...$ represent all $\cong$ equivalence classes. The *Simple agent-frame for L on X* is the sequence $\langle X, [\beta_0], [\beta_1], ...\rangle$.

For every $\cong$ equivalence class $[\beta]$ let $R^X_{[\beta]}$ be a relation from X to itself where for every $w_1, w_2 \in X$ it holds that $w_1R^X_{[\beta]}w_2$ iff $\{A : \Delta\beta A \in w_1\} \subseteq w_2$. $R^X_{[\beta]}$ is an equivalence relation. Let $E^X_{[\beta]}$ be the set of all $R^X_{[\beta]}$ equivalence classes. Let the *simple canonical frame of L with relation to X* be the sequence $\langle X, [\beta_1], [\beta_2], ..., R^X_{[\beta_1]}, R^X_{[\beta_2]}, ...\rangle$.

To extend the frame to work with multiple STIT frames, more R_L equivalence classes are needed. Let $X_{\mathcal{N}}$ be the set of all R_L equivalence classes such that for every $X^n \in X^{\mathcal{N}}$ it holds that for every agent term α and β, and all $w \in X$ and all $w' \in X^n$, $\alpha = \beta \in w$ iff $\alpha = \beta \in w'$. This restricts every element in $X^{\mathcal{N}}$ to only have agents that exist in X. Let us use X', X'' for the elements in $X^{\mathcal{N}}$.

For every $\cong$ equivalence class $[\beta]$, for every $X' \in X^{\mathcal{N}}$ let $R^{X'}_{[\beta]}$ be a relation from X' to itself where for every $w_1, w_2 \in X'$, $\langle w1, w2\rangle \in R^{X'}_{[\beta]}$ iff $\{A : \Delta\alpha A \in w_1\} \subseteq w_2$. $R^{X'}_{[\beta]}$ is an equivalence relation as well. Let $E^{X'}_{[\beta]}$ be the set of all $R^{X'}_{[\beta]}$ equivalence classes.

Let for every $\cong$ equivalence class $[\beta]$, $R^{[\beta]}_N$ be a relation from $X^{\mathcal{N}}$ to itself where for every $X', X'' \in X^{\mathcal{N}}$, $X'R^{[\beta]}_N X''$ iff for all $w' \in X'$ there exists a $w'' \in X''$ such that $\{A : \boxtimes_\beta A \in w'\} \subseteq w''$.

We call the *canonical frame of L* the sequence:

$$\langle [\beta_1], [\beta_2], ..., \langle X_1, R^{X_1}_{[\beta_1]}, R^{X_1}_{[\beta_2]}, ...\rangle, \langle X_2, R^{X_2}_{[\beta_1]}, R^{X_2}_{[\beta_2]}, ...\rangle, ..., R^{[\beta_1]}_N, R^{[\beta_2]}_N, ...\rangle$$

10.4.2 Completeness theorem

Before we start proving completeness, we first have to prove some lemmas that will make the final proof more convenient. When observing the function R_n in the model and the relations $R^{[\beta]}_N$ in the canonical frame, some functional properties must hold for R_n. The first property is that for

any equivalence class X and $\cong$ equivalence class $[\beta]$ there is only one equivalence class X' where $XR_N^{[\beta]}X'$ holds. The second property is that for any equivalence class X and $\cong$ equivalence class $[\beta]$ and the only one equivalence class X' for which it holds that $XR_N^{[\beta]}X'$, for any $w \in X$ there is exactly one $w' \in X'$ where $\{A : \boxtimes_\beta A \in w\} \subseteq w'$.

Lemma 5. *Let X be any $\Box$ equivalence class and let $[\beta]$ be any $\cong$ equivalence class. Then there is at most one $\Box$ equivalence class X' for which it holds that $XR_N^{[\beta]}X'$.*

Proof. Let X be any $\Box$ equivalence class and let $[\beta]$ be any $\cong$ equivalence class in the canonical frame $\langle [\beta_1], [\beta_2], ..., \langle X_0, R^{X_0}_{[\beta_1]}, R^{X_0}_{[\beta_2]}, ...\rangle, \langle X_1, R^{X_1}_{[\beta_1]}, R^{X_1}_{[\beta_2]}, ...\rangle, \langle X_2, R^{X_2}_{[\beta_1]}, R^{X_2}_{[\beta_2]}, ...\rangle, ..., R_N^{[\beta_1]}, R_N^{[\beta_2]}, ...\rangle$. Then let us for proof of contradiction assume that there are two different $\Box$ equivalence classes X' and X'' where $XR_N^{[\beta]}X'$ and $XR_N^{[\beta]}X''$ hold. By definition of the canonical frame for all $w \in X$ there exists a $w' \in X'$ such that $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ and for all $w \in X$ there exists a $w'' \in X''$ such that $\{A : \boxtimes_\beta A \in w\} \subseteq w''$. Since X' and X'' are different, there is a formula B such that for all MCS $w'_B \in X'$ it holds that $\Box B \in w'_B$ and for all MCS $w''_B \in X''$ it hold that $\neg\Box B \in w''_B$. Let $w_B \in X$ be any MCS. By tautology either $\boxtimes_\beta \Box B \in w_B$ or $\boxtimes_\beta \Box B \notin w_B$. We prove that both claims will lead to a contradiction.

$\boxtimes_\beta \Box B \in w_B$ Since $\boxtimes_\beta \Box B \in w_B$ and $XR_N^{[\beta]}X''$ hold, there is an MCS $w_{\Box B} \in X''$ where $\Box B \in w_{\Box B}$. Since for all $w''_B \in X''$ it hold that $\neg\Box B \in w''_B$ it also holds that $\neg\Box B \in w_{\Box B}$. This leads to a contradiction. $\lightning$

$\boxtimes_\beta \Box B \notin w_B$ Since $\boxtimes_\beta \Box B \notin w_B$ holds, $\neg \boxtimes_\beta \Box B \in w_B$ holds by definition of MCS. By a derivation of axiom $A7$ it then holds that $\boxtimes_\beta \neg\Box B \in w_B$. Since $\boxtimes_\beta \neg\Box B \in w_B$ and $XR_N^{[\beta]}X'$ hold, there is an MCS $w_{\Box B} \in X'$ where $\neg\Box B \in w_{\Box B}$. Since for all $w'_B \in X'$ it hold that $\Box B \in w'_B$ it also holds that $\Box B \in w_{\Box B}$. This leads to a contradiction. $\lightning$

$\Box$

Lemma 6. *Let X be any $\Box$ equivalence class and let $[\beta]$ be any $\cong$ equivalence class. Then for every MCS $w \in X$ there is an MCS w' where $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds.*

Proof. Let X be any $\Box$ equivalence class and let $[\beta]$ be any $\cong$ equivalence class in the canonical frame $\langle [\beta_1], [\beta_2], ..., \langle X_0, R^{X_0}_{[\beta_1]}, R^{X_0}_{[\beta_2]}, ...\rangle, \langle X_1, R^{X_1}_{[\beta_1]}, R^{X_1}_{[\beta_2]}, ...\rangle, \langle X_2, R^{X_2}_{[\beta_1]}, R^{X_2}_{[\beta_2]}, ...\rangle, ..., R_N^{[\beta_1]}, R_N^{[\beta_2]}, ...\rangle$. Also let w be any MCS in X. Then we have to prove that there is an MCS w' such that $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds. Assume for contradiction that there is no MCS w' such that $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds. By the definition of an MCS it holds that $\{A : \boxtimes_\beta A \in w\}$ is inconsistent[6]. Therefore there exist a formula B such that $B \in \{A : \boxtimes_\beta A \in w\}$ and $\neg B \in \{A : \boxtimes_\beta A \in w\}$. That means that $\boxtimes_\beta B \in w$ and $\boxtimes_\beta \neg B \in w$. By axiom A6 $\neg \boxtimes_\beta \neg B \in w$ holds. Since w is consistent $\boxtimes_\beta \neg B \notin w$ should also hold but is in contradiction with the prior statement $\boxtimes_\beta \neg B \in w$. $\lightning$ $\Box$

[6]The set $\{A : \boxtimes_\beta A \in w\}$ cannot be an MCS since there is no w' that contains it. Therefore the set must be either not maximal or not consistent. We can show that the set is maximal and therefore not consistent. Assume for contradiction that $\{A : \boxtimes_\beta A \in w\}$ is not maximal. Then there is a new formula C that can be added to the set without making it inconsistent. We know that w is maximal and therefore contains either $\boxtimes_\beta C$ or $\neg \boxtimes_\beta C$. If $\boxtimes_\beta C$ is in w then C is already in $\{A : \boxtimes_\beta A \in w\}$ and therefore not new, so $\neg \boxtimes_\beta C$ is in w. By axiom A7 we also know that $\boxtimes_\beta \neg C$ is in w and therefore $\neg C$ is in $\{A : \boxtimes_\beta A \in w\}$. This means that adding C would lead to a contradiction and therefore we know that $\{A : \boxtimes_\beta A \in w\}$ must be maximal and inconsistent.

Lemma 7. *Let X be any $\Box$ equivalence class and let $[\beta]$ be any $\cong$ equivalence class. Then there is at least one $\Box$ equivalence class X' for which it holds that $XR_N^{[\beta]}X'$.*

Proof. Let X be any $\Box$ equivalence class and let $[\beta]$ be any $\cong$ equivalence class in the canonical frame $\langle [\beta_1], [\beta_2], ..., \langle X_0, R^{X_0}_{[\beta_1]}, R^{X_0}_{[\beta_2]}, ...\rangle, \langle X_1, R^{X_1}_{[\beta_1]}, R^{X_1}_{[\beta_2]}, ...\rangle, \langle X_2, R^{X_2}_{[\beta_1]}, R^{X_2}_{[\beta_2]}, ...\rangle, ..., R_N^{[\beta_1]}, R_N^{[\beta_2]}, ...\rangle$. Then we have to prove that there is a $\Box$ equivalence class X' where holds that for every $w \in X$ there is a $w' \in X'$ such that $\{A : \boxtimes_\beta A \in w\} \subseteq w'$. Assume for contradiction that there is no $\Box$ equivalence class X' such that for all $w \in X$ there is a $w' \in X'$ such that $\{A : \boxtimes_\beta A \in w\} \subseteq w'$. Since, by lemma 6, it holds that for all $w \in X$ there is an MCS w' such that $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ and since, by definition of the $\Box$ equivalence class, for some $\Box$ equivalence class X' it holds that $w' \in X'$, it can only be so that there are two distinct MCS's $w_1, w_2 \in X$ such that there is an MCS $w_1' \in X_1'$ where it holds that $\{A : \boxtimes_\beta A \in w_1\} \subseteq w_1'$ and such that there is an MCS $w_2' \in X_2'$ where it holds that $\{A : \boxtimes_\beta A \in w_2\} \subseteq w_2'$. It must also hold that X_1' is not the same $\Box$ equivalence class as X_2'. Since X_1' an X_2' are not the same equivalence classes, there is a formula B such that $\Box B \in w_1'$ and $\neg\Box B \in w_2'$ [7]. Since $\Box B \in w_1'$ holds, by axiom A0 $\Box\Box B \in w_1'$ also holds.[8] This means that $\boxtimes_\beta\Box\Box B \in w_1$ holds as well.[9] By axiom A8 it then also holds that $\Box \boxtimes_\beta \Box B \in w_1$. By the definition of the $\Box$ equivalence class $\boxtimes_\beta\Box B \in w_2$ also holds. But since $\neg\Box B \in w_2'$ it also holds that $\boxtimes_\beta\neg\Box B \in w_2$ and therefore by axiom A6 $\neg \boxtimes_\beta \Box B \in w_2$ also holds which contradicts with the previous statement of $\boxtimes_\beta\Box B \in w_2$.↯ □

Corollary 1. *Let X be any $\Box$ equivalence class and let $[\beta]$ be any $\cong$ equivalence class. Then there is exactly one $\Box$ equivalence class X' for which it holds that $XR_N^{[\beta]}X'$.*

Lemma 8. *Let X be any $\Box$ equivalence class, let $w \in X$ be any MCS and let $[\beta]$ be any $\cong$ equivalence class. Let X' be the $\Box$ equivalence class where $XR_N^{[\beta]}X'$ holds. Then there is at most one unique $w' \in X'$ where $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds.*

Proof. Let X be any $\Box$ equivalence class, let $w \in X$ be any MCS and let $[\beta]$ be any $\cong$ equivalence class. Let X' be the $\Box$ equivalence class where $XR_N^{[\beta]}X'$ holds. Then we have to prove that there is at most one unique $w' \in X'$ where $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds.

Let us assume for proof of contradiction that there exists $w', w'' \in X'$ that are unique where both $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ and $\{A : \boxtimes_\beta A \in w\} \subseteq w''$ hold. Since w' and w'' are unique, there is a formula B where $B \in w'$ and $B \notin w''$. Then since $B \in w'$, and w' is consistent, $\neg B \notin w'$. Therefore, since $\neg B \notin w'$ and $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ it holds that $\boxtimes_\beta\neg B \notin w$. From this and the assumption that w is maximal follows that $\neg \boxtimes_\beta \neg B \in w$. Then by definition of axiom $A7$ it follows that $\boxtimes_\beta B \in w$. Since $\boxtimes_\beta B \in w$ and $\{A : \boxtimes_\beta A \in w\} \subseteq w''$ hold, $B \in w''$ holds. This is in contradiction with the assumption that $B \notin w''$. ↯ □

[7] Assume for contradiction that w_1' and w_2' are not in the same $\Box$ equivalence class and that there is no formula B such that $\Box B \in w_1'$ and $\neg\Box B \in w_2'$. In other words, for all formulas B it holds that $\Box B \in w_1'$ if and only if $\Box B \in w_2'$. By axiom A0 we know that if $\Box B \in w_2'$ then $B \in w_2'$. It then follows that if $\Box B \in w_1'$ then $B \in w_2'$ or $\{B : \Box B \in w_1'\} \subseteq w_2'$. By definition it then holds that $w_1' R_l w_2'$ and thus w_1' and w_2' are in the same $\Box$ equivalence class.↯

[8] The box operator is in S5 which means that it is also transitive and therefore the derived rule $\Box A \rightarrow \Box\Box A$ holds in the current axiom system.

[9] Assume for contradiction that $\boxtimes_\beta\Box\Box B \notin w_1$ holds. Then $\neg \boxtimes_\beta \Box\Box B \in w_1$ holds since w_1 is maximal. Then by axiom A7 it holds that $\boxtimes_\beta\neg\Box\Box B \in w_1$. Therefore $\neg\Box\Box B \in w_1'$ should hold which would make w_1' inconsistent.

Lemma 9. *Let X be any $\Box$ equivalence class, let $w \in X$ be any MCS and let $[\beta]$ be any $\cong$ equivalence class. Let X' be the $\Box$ equivalence class where $XR_N^{[\beta]}X'$ holds. Then there is at least one $w' \in X'$ where $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds.*

Proof. Let X be any $\Box$ equivalence class, let $w \in X$ be any MCS and let $[\beta]$ be any $\cong$ equivalence class. Let X' be the $\Box$ equivalence class where $XR_N^{[\beta]}X'$ holds. Then we have to prove that there is at least one $w' \in X'$ where $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds. Since $w \in X$ and $XR_N^{[\beta]}X'$ hold, by lemma 6, there is a $w' \in X'$ such that $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds. $\Box$

Corollary 2. *Let X be any $\Box$ equivalence class, let $w \in X$ be any MCS and let $[\beta]$ be any $\cong$ equivalence class. Let X' be the $\Box$ equivalence class where $XR_N^{[\beta]}X'$ holds. Then there is exactly one unique $w' \in X'$ where $\{A : \boxtimes_\beta A \in w\} \subseteq w'$ holds.*

Theorem 2. *Every L_0 consistent set θ of formulas is satisfiable in a structure; and every L_n consistent set θ of formulas ($n \geq 1$) is satisfiable in a structure in which every agent has at most n possible choices at every moment.*

Proof. Let $L = L_n$ for any $n \geq 0$ be an axiomatic system defined by the axioms described in section 10.1, and let θ be any L consistent set of formulas. Then by the Lindenbaum Lemma there is an maximally consistent set w containing θ. Let X_0 be the equivalence class to which w belongs and let $\langle [\beta_1], [\beta_2], ..., \langle X_0, R_{[\beta_1]}^{X_0}, R_{[\beta_2]}^{X_0}, ...\rangle, \langle X_1, R_{[\beta_1]}^{X_1}, R_{[\beta_2]}^{X_1}, ...\rangle, \langle X_2, R_{[\beta_1]}^{X_2}, R_{[\beta_2]}^{X_2}, ...\rangle, ..., R_N^{[\beta_1]}, R_N^{[\beta_2]}, ...\rangle$ be the canonical frame. Then we have to prove that there is a interpretation STIT model M where for every formula ϕ it holds that $\phi \in w$ if and only if $M \vDash \phi$.

Let us use $X^{\mathcal{N}}$ for the set of all elements $X_0, X_1, X_2, ...$ from the canonical frame. We convert the canonical frame in a interpretation $S = \langle Agent, \langle Tree^0, \leq^0, Choice^0, R_n^0\rangle, \langle Tree^1, \leq^1, Choice^1, R_n^1\rangle, ...\rangle$ in the following way:

- $Agent = \{[\beta_1], [\beta_2], ...\}$
- For all $k \geq 0$ and all $X_k \in X^{\mathcal{N}}$:
  - $Tree^k = \{X_k\} \cup X_k$
  - $\leq^k = \{\langle X_k, w\rangle : w \in X_k\} \cup \{\langle X_k, X_k\rangle\} \cup \{\langle w, w\rangle : w \in X_k\}$
  - $Choice^k$ is defined by the following functional relations for every $[\beta] \in Agent$:
    - $Choice_{X_k}^{k[\beta]}(\{X_k\}) = \{H : \exists e \in E_{[\beta]}^{X_k} \wedge H = \{h_w : w \in e\}\}$ where $h_w = \{X_k, w\}$ for any $w \in X_k$.
    - $Choice_{w'}^{k[\beta]}(h_w) = \{h_w\}$ where $h_w = \{X_k, w\}$ for any $w \in X_k$ where $w' \in \{w, X_k\}$
  - R_n^k is defined by the following functional relation for every $[\beta] \in Agent^k$:
    - $R_n^k([\beta]) = \langle N, R_h\rangle$ such that where $h_w = \{X_k, w\}$ for any $w \in X_k$ and $m \geq 0$:
      - N is a node $\langle Tree^m, \leq^m, Choice^m, R_n^m\rangle$ in S where $\{X_m\} \in Tree^m$ and $X_k R_N^{[\beta]} X_m$ for any $m \geq 0$.
      - $R_h(h_w) = h_{w'}$ where $h_{w'} = \{X_m, w'\}$ where $w' \in \{w'' \in X_m : \{A : \boxtimes_{[\beta]} A \in w\} \subseteq w'\}$

      Note that $R_n^k([\beta])$ is a function because by the corollaries 1 and 2, it is ensured that exactly one N and exactly one $h_{w'}$ per h_w exists.

Let us define the model $M = \langle S, V\rangle$ such that for each agent term β, $V(\beta) = [\beta]$ and for each propositional variable p and each history h_w , $\langle X_k, h_w, N^k\rangle \in V(p)$ iff $p \in w$ for $k \geq 0$ where $N^k = \langle Tree^k, \leq^k, Choice^k, R_n^k\rangle$.

Now we have to prove that for every $w \in X_0$ it holds that $\phi \in w$ if and only if $M, X_0/h_w/N^0 \vDash \phi$. Now we give a proof by induction over ϕ that for every $w \in X_0$ and for every formula $\phi \in w$ it holds that $M, X_0/h_w/N^0 \vDash \phi$.

atomic proposition Assume p is an atomic proposition and $p \in w$, then we have to prove that $M, X_0/h_w/N^0 \vDash p$. By the definition of atomic propositions in the model, we have to prove $\langle X_0/h_w/N^0\rangle \in V(p)$. According to the valuation function where for each propositional variable p and each history h_w , $\langle X_k, h_w, N^k\rangle \in V(p)$ iff $p \in w$ for $k \geq 0$, $\langle X_0/h_w/N^0\rangle \in V(p)$.

agent equality Assume α and β are agents terms and $\alpha = \beta \in w$, then we have to prove that $M, X_0/h_w/N^0 \vDash \alpha = \beta$. By the definition of the $\cong$ equivalence class, we know that $[\alpha]$ and $[\beta]$ are agents in the model and since $\cong$ is an equivalence relation we know that since $\alpha = \beta$, $[\alpha] = [\beta]$. Therefore, by definition of the evaluation rule, we know that $V(\alpha) = V(\beta)$. Thus, by the agent equality definition of the model, $M, X_0/h_w/N^0 \vDash \alpha = \beta$.

negation Assume ϕ is some formula for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ for any k. Also assume that $\neg\phi \in w$. Then we have to prove that $M, X_0/h_w/N^0 \vDash \neg\phi$. By definition of the logical model, we have to prove that $M, X_0/h_w/N^0 \nvDash \phi$.

Since $\neg\phi \in w$ and w is an MCS, then $\phi \notin w$. By the assumption of induction it holds that $M, X_0/h_w/N^0 \nvDash \phi$.

conjunction Assume ϕ, ψ are some formulas for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ and $\psi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \psi$ for any k. Also assume that $\phi \wedge \psi \in w$. Then we have to prove that $M, X_0/h_w/N^0 \vDash \phi \wedge \psi$. By definition of the logical model, it is sufficient to prove that $M, X_0/h_w/N^0 \vDash \phi$ and $M, X_0/h_w/N^0 \vDash \psi$.

Since $\phi \wedge \psi \in w$ by definition of the logical tautologies, $\phi \in w$ and $\psi \in w$. By the assumption of induction, it then holds that $M, X_0/h_w/N^0 \vDash \phi$ and $M, X_0/h_w/N^0 \vDash \psi$.

box Assume ϕ is some formula for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ for any k. Also assume that $\Box\phi \in w$. Then we have to prove that $M, X_0/h_w/N^0 \vDash \Box\phi$. By the definition of the interpretation STIT model, we know that it is sufficient to prove that $M, X_0/h'/N^0 \vDash \phi$ for every h' where $h' \in H_{X_0}$. By definition of the canonical frame, we know that $H_{X_0} = \{h_{w'} : w' \in X_0\}$. Therefore we have to prove that $M, X_0/h_{w'}/N^0 \vDash \phi$ for every $w' \in X_0$.

By definition X in the canonical frame, we know that for all $w' \in X_0$ it holds that if for any $w'' \in X_0$ it holds that $\Box\phi \in w''$ then $\phi \in w'$. Since by assumption $\Box\phi \in w$, it holds that $\phi \in w'$ for every $w' \in X_0$. By the assumption of induction, it is derivable that $M, X_0/h_{w'}/N^0 \vDash \phi$ for every $w' \in X_0$.

dstit Assume ϕ is some formula for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ for any k. Also assume α is some agent and $[\alpha\ dstit:\ \phi] \in w$. Then we have to prove that $M, X^n/h_w/X \vDash [\alpha\ dstit:\ \phi]$. By definition of the interpretation STIT model, we know that is is sufficient to prove that $M, X_0/h'/N^0 \vDash \phi$ for all $h' \in Choice^0\,{}^{\alpha}_{X_0}(h_w)$ and $M, X_0/h''/N^0 \nvDash \phi$ for

some $h'' \in H_{X_0}$. By definition of the canonical frame, we know that $H_{X_0} = \{h_{w''} : w'' \in X_0\}$. By definition of the canonical frame, we also know that $Choice^0 \overset{\alpha}{}_{X_0}(h_w) = \{h_w\}$.

Since it holds that $[\alpha\ dstit:\ \phi] \in w$, we also know that $\Delta\alpha\phi \in w$ by $A2$. Then, by definition of the canonical frame and $A1$, $\phi \in w'$ for every world in the equivalence class $R^{X_0}_{[\alpha]}$ where $w \in R^{X_0}_{[\alpha]}$. Therefore, $\phi \in w$. By definition of the valuation function, it then holds that $M, X_0/h_w/N^0 \vDash \phi$, thus $M, X_0/h'/N^0 \vDash \phi$ for all $h' \in Choice^0 \overset{\alpha}{}_{X_0}(h_w) = \{h_w\}$.

For the second part we use the assumption $[\alpha\ dstit:\ \phi] \in w$ again. Since $[\alpha\ dstit:\ \phi] \in w$, we also know that $\neg\Box\phi \in w$ by $A2$. Since w is an MCS, it can be inferred that $\Box\phi \notin w$. Since $\Box\phi \notin w$ and $w \in X_0$ by assumption, and w is an MCS, and since by the characteristics of X_0 being a $\Box$ equivalence class, it must hold that there is a MCS $w' \in X_0$ where $\neg\phi \in w'$. Since $w' \in X_0$ and $\neg\phi \in w'$, by definition of the canonical model and the induction assumption, there is a history $h_{w'}$ for which it holds that $MX_0, h_{w'}, N^0 \vDash \neg\phi$. By logical definition of negation $MX_0, h_{w'}, N^0 \nvDash \phi$. Since $w' \in X_0$, by definition of the canonical model $h_{w'} \in H_{X_0}$. Therefore, we know that $M, X_0/h''/N^0 \nvDash \phi$ for some $h'' \in H_{X_0}$, namely $h_{w'}$.

Interpretation operator Assume ϕ is some formula for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ for any k and assume α is some agent. Also assume $\boxtimes_\alpha\phi \in w$. Then we have to prove that $M, X_0/h_w/N^0 \vDash \boxtimes_\alpha\phi$. By the logical definition of $\boxtimes$ it is sufficient to prove that $M, m'/h'/n' \vDash \phi$ where $R_n(\alpha) = \langle n', R_h\rangle$ and $R_{h_w}(\langle X_0, h_w\rangle) = \langle m', h'\rangle$.

By definition of the canonical frame, we know that $R^0_n(\alpha) = \langle n'', R_h\rangle$ such that:

- n'' is the node $\langle Tree'', \leq'', Choice'', R''_n\rangle$ in S where $\{X''\} \in Tree''$ and $X_0 R^{[\alpha]}_N X''$.
- $R_h(h_w) = h_{w''}$ where $h_{w''} = \{X'', w''\}$ where $w'' \in \{w' \in X'' : \{A : \boxtimes_{[\alpha]} A \in w\} \subseteq w'\}$

Since $\boxtimes_{[\alpha]}\phi \in w$ we know that $\phi \in w''$. By the induction assumption and definition of the canonical model, we know that $MX'', h_{w''}, n'' \vDash \phi$.[10]

Now we give a proof by induction over ϕ that for every $w \in X_0$ and for every formula ϕ it holds that if $M, X_0/h_w/N^0 \vDash \phi$ then $\phi \in w$.

atomic proposition Assume p is an atomic proposition and $M, X_0/h_w/N^0 \vDash p$, then we have to prove that $p \in w$. Since $M, X_0/h_w/N^0 \vDash p$ it holds by definition that $\langle X_0/h_w/N^0\rangle \in V(p)$. By assumption, then it also holds that $p \in w$.

agent equality Assume α and β are agents terms and $M, X_0/h_w/N^0 \vDash \alpha = \beta$, then we have to prove that $\alpha = \beta \in w$. Since $M, X_0/h_w/N^0 \vDash \alpha = \beta$ it holds by definition that $V(\alpha) = V(\beta)$. By definition, it also holds that $[\alpha] = [\beta]$. Thus, by definition of the $\cong$ equivalence class, $\alpha \cong \beta$ holds. Since $\alpha \cong \beta$, it also holds that $\alpha = \beta \in w$ by the definition of the canonical frame.

negation Assume ϕ is some formula for which it holds that that $M, X_k/h_{w^k}/N^k \vDash \phi$ iff $\phi \in w^k$ for any k. Also assume that $M, X_0/h_w/N^0 \vDash \neg\phi$. Then we have to prove that $\neg\phi \in w$. Since $M, X_0/h_w/N^0 \vDash \neg\phi$ it holds that $M, X_0/h_w/N^0 \nvDash \phi$. Therefore, since $M, X_0/h_w/N^0 \nvDash \phi$, it also holds that $\phi \notin w$. Since w is maximal, it holds that $\neg\phi \in w$.

[10] For proof of the independence of agents and possible choices, see Xu(1994). His proofs still work for the interpretation STIT logic.

conjunction Assume ϕ, ψ are some formulas for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ and $\psi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \psi$ for any k. Also assume that $M, X_0/h_w/N^0 \vDash \phi \wedge \psi$. Then we have to prove that $\phi \wedge \psi \in w$. By definition of the model, it holds that $M, X_0/h_w/N^0 \vDash \phi$ and $M, X_0/h_w/N^0 \vDash \psi$. Then, by definition of the canonical model it holds that $\phi \in w$ and $\psi \in w$. By the tautologies of propositional logic and the fact that w is maximally consistent, it holds that $\phi \wedge \psi \in w$.

box Assume ϕ is some formula for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ for any k. Also assume that $M, X_0/h_w/N^0 \vDash \Box\phi$. Then we have to prove that $\Box\phi \in w$. By definition of the interpretation STIT model, since $M, X_0/h_w/N^0 \vDash \Box\phi$ it holds that $M, X_0/h'/N^0 \vDash \phi$ for every $h' \in H_{X_0}$. Thus, by definition of the canonical model and the induction assumption, $\phi \in w'$ holds for every $w' \in X_0$. By definition of the canonical frames $\Box$ equivalence classes and the fact that w is maximally consistent, it holds that $\Box\phi \in w'$ for every $w' \in X_0$. Since $w \in X_0$ it holds that $\Box\phi \in w$.

dstit Assume ϕ is some formula for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ for any k. Also assume α is some agent and $M, X^n/h_w/X \vDash [\alpha \ dstit: \ \phi]$. Then we have to prove that $[\alpha \ dstit: \ \phi] \in w$. By definition of the interpretation STIT model, it holds that $M, X_0/h'/N^0 \vDash \phi$ for all $h' \in Choice^0 \, {}^{\alpha}_{X_0}(h_w)$ and $M, X_0/h''/N^0 \nvDash \phi$ for some $h'' \in H_{X_0}$. By definition of the canonical model and the induction hypothesis it holds that $\phi \in w$ and $\phi \notin w'$ for some $w \in X_0$. Since $h' \in Choice^0 \, {}^{\alpha}_{X_0}(h_w)$ it also holds by definition of the canonical frame that $\Delta\alpha\phi \in w$. Since $\phi \in w$ and $\phi \notin w'$ for some $w \in X_0$ for all $w'' \in X_0$ it also holds that $\neg\Box\phi \in w''$. By the $A2$ axiom, it also holds that $[\alpha \ dstit: \ \phi] \in w$.

Interpretation operator Assume ϕ is some formula for which it holds that that $\phi \in w^k$ iff $M, X_k/h_{w^k}/N^k \vDash \phi$ for any k and assume α is some agent. Also assume $M, X_0/h_w/N^0 \vDash \boxtimes_\alpha\phi$. Then we have to prove that $\boxtimes_\alpha\phi \in w$. By the logical definition of $\boxtimes$ it holds that $M, m'/h'/n' \vDash \phi$ where $R_n(\alpha) = \langle n', R_h\rangle$ and $R_{h_w}(\langle X_0, h_w\rangle) = \langle m', h'\rangle$. By the induction hypothesis and the definition of the canonical model, it holds that $\phi \in w'$ where $w' \in X_m$ such that $X_0 R^{[\alpha]} X_m$ and $\{A : \boxtimes_{[\alpha]} A \in w\} \subseteq w'$. By corollary 2 and since $w' \in X_m$ such that $X_0 R^{[\alpha]} X_m$ and $\{A : \boxtimes_{[\alpha]} A \in w\} \subseteq w'$ and $\phi \in w'$ and w is maximally consistent, it holds that $\boxtimes_\alpha\phi \in w$.

□

Theorem 3. *The axiom scheme from section 10.1 and 10.2 is complete with respect to the interpretation STIT frame. Hence, for every consistent set of formulas* Γ *and any formula* ϕ *it holds that if* ϕ *is derivable from* Γ *in the interpretation STIT frame, then* ϕ *is derivable from* Γ *according to the axioms from section 10.1 and 10.2.*

Proof. Assume that ϕ is not derivable from Γ according to the axioms from section 10.1 and 10.2. Then we have to prove that ϕ is not derivable from Γ in the interpretation STIT frame. Since ϕ is not derivable from Γ according to the axioms, $\Gamma \cup \neg\phi$ is a consistent set. Since $\Gamma \cup \neg\phi$ is consistent, there is a maximally consistent set Π such that $\Gamma \cup \neg\phi \subseteq \Pi$. By theorem 2 we know that there is an interpretation STIT model M where Π is derivable. Therefore M also derives $\Gamma \cup \neg\phi$. Thus Since M derives Γ and M does not derive ϕ it holds that ϕ is not derivable from Γ in the interpretation STIT frame. □

Corollary 3. *The axiom scheme from section 10.1 and 10.2 is sound and complete with respect to the interpretation STIT frame.*

11 Conclusion

This thesis started with an overview of STIT and deontic logic. After that earlier research into combining knowledge and STIT was discussed. Here, the problem of implementing a independent knowledge operator, the problem with combining deontic and epistemic logic and the problems with the deserted traveler example were stated. Then the new interpretation STIT logic was introduced and described. This new logic does not have the problem with implementing a knowledge operator and combining the knowledge operator with the deontic operator, since it does not have a knowledge operator. Also, some new insight was given in the deserted traveler example. After that, some new suggestions for implementing a knowledge operator were given. Finally, soundness and completeness were proven for the interpretation STIT logic.

A lot of new situations seem to be able to be modeled with Interpretation STIT. The operator representing the term "acts as if" gives a new perspective to reasoning. However, it is hard to correctly work with Interpretation STIT, since the connection between "acting as if" and "reasoning about" is still to be worked out more specifically.

To work out how knowledge should work with relation to a physical interpretation is not yet clear cut. As discussed in this thesis, a starting point can be to take the notion of justified true belief, since the notion of interpretation in the new STIT model resembles belief a lot. This also has its downsides. In the Interpretation STIT frame it is impossible to interpret the interpretation operator as "α believes that p". Believing implies a difference between a mental state and the physical state of an agent. This difference is impossible to model with interpretation STIT. Interpretation STIT can only model that physical choices one assumes to make. Therefore, Interpretation STIT only models the part of the mental state that does cause the choices to be made. It is believable (and debatable) that there is such a thing in the mental state that does not influence the choices an agent makes. More research to define the philosophical place of the interpretation operator with relation to knowledge could help solve this issue with Interpretation STIT.

Another notion that is to be reconsidered in Interpretation STIT is that of deliberateness. This notion already has implications about the mental state of an agent that conflicts with the part of the mental state regular STIT is trying to model as stated in section 8.3. A new semantic definition for the operator modeling the statement "α deliberately sees to it that p" could help clarify what is meant by that operator. The *dstit* should incorporate the interpretation operator to make it more in line with our intuition about deliberation.

The axiom schema corresponding to the current interpretation STIT frame is also debatable. For this version it is chosen that, since values are evaluated at the end of the interpretations, the interpretations of agents cannot be influenced by choices already made. An agent could not reason that if another choice was made somewhere in the model, then the agent would have another interpretation of the situation. Another idea is to make a version of the interpretation STIT logic that allows for interpretations to differ based on choices made. However, such a logic will probably be even more complex than this version. This brings us to another point of improvement with this interpretation STIT logic.

The way of depicting the model could also be improved. Currently, the complexity of the models can increase extremely fast. For example, in section 8.2 the examples already omit a part of the complete model because the visualization became too complex otherwise. However, the intuition

of the models is significantly clearer that the models would suggest. It therefore seems likely that there are equivalent models to be thought of that scale better than an Interpretation STIT model.

There are a few problems with interpretation STIT that are still to overcome. How to connect the notion of knowledge and the notion of interpretation logically is still unclear. It can also be possible that some generally preferred notion of knowledge cannot even be modeled in an Interpretation STIT model. The usability of Interpretation STIT for more complex problems can also pose a practical issue. With the current model that Interpretation STIT uses it is difficult to clearly represent a model with more than two agents. However, there is a possibility that there can be an equivalent and more visually simplified definition of Interpretation STIT.

On the other hand, interpretation STIT gives a new way to look at epistemic problems. The possibilities of the new interpretation operator seem to be quite extensive. Some notions of awareness relating STIT, namely *the disregard of choice, the illusion of choice, the disregard of possibilities* and *the illusion of possibilities*, can be defined with the new type of model. Also, decisions of obligation can be made about problems where knowledge plays a role with Interpretation STIT and the defined notions of awareness. It also seems possible to define varying knowledge operators by using the interpretation operator and STIT. Such flexibility can be essential for solving the problem that the disagreement about knowledge brings. Different notions of knowledge can be used for different situations.

References

[1] Aqvist, L. (1967). Good Samaritans, *contrary-to-duty imperatives, and epistemic obligations. Noûs*, 1(4), 361-379.

[2] Belnap, N., & Perloff, M. (1993). In the realm of agents. *Annals of Mathematics and Artificial Intelligence*, 9(1-2), 25-48.

[3] Belnap, N. D., Perloff, M., & Xu, M. (2001). *Facing the future: agents and choices in our indeterminist world.* Oxford University Press on Demand.

[4] Broersen, J., Herzig, A., & Troquard, N. (2006, September). A stit-extension of atl. In *European Workshop on Logics in Artificial Intelligence* (pp. 69-81). Springer, Berlin, Heidelberg.

[5] Broersen, J., Herzig, A., & Troquard, N. (2007, June). A normal simulation of coalition logic and an epistemic extension. In *Proceedings of the 11th conference on Theoretical aspects of rationality and knowledge* (pp. 92-101). ACM.

[6] Broersen, J. (2008, May). A complete STIT logic for knowledge and action, and some of its applications. In *International Workshop on Declarative Agent Languages and Technologies* (pp. 47-59). Springer, Berlin, Heidelberg.

[7] Broersen, J. (2011). Deontic epistemic STIT logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2), 137-152.

[8] Broersen, J., & van der Torre, L. (2012). Ten problems of deontic logic and normative reasoning in computer science. In *Lectures on Logic and Computation* (pp. 55-88). Springer, Berlin, Heidelberg.

[9] J. Broersen (personal communication, October 25, 2018) explained The Desert Traveler.

[10] Brown, M. A. (1988). On the logic of ability. *Journal of philosophical logic*, 17(1), 1-26.

[11] Chellas, B. F. (1969). The Logical Form of Imperatives, Phd.

[12] Hansson, B. (1969). An analysis of some deontic logics. *Nous*, 373-398.

[13] Herzig, A., & Troquard, N. (2006, May). Knowing how to play: uniform choices in logics of agency. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems* (pp. 209-216). ACM.

[14] Horty, J. F. (1989). An alternative STIT operator. *Manuscript, Philosophy Department, University of Maryland.*

[15] Horty, J. F. (2001). *Agency and deontic logic*. Oxford University Press.

[16] Von Kutschera, F. (1986). Bewirken. *Erkenntnis*, 24(3), 253-281.

[17] Kenny, A. (1976). Human abilities and dynamic modalities. *Essays on explanation and understanding* (pp. 209-232). Springer, Dordrecht.

[18] McNamara, P. (2019, June 21) *Deontic Logic*. Retrieved from https://plato.stanford.edu/archives/sum2019/entries/logic-deontic/

[19] Prior, A. N. (1958). Escapism: The Logical Basis of Ethics. *Melden*, 135-146.

[20] Ross, A. (1941). Imperatives and Logic. *Theoria*, 7, 53-71.

[21] Von Wright, G. H. (1951). Deontic logic. *Mind, 60*(237), 1-15.

[22] Xu, M. (1994, July). Decidability of deliberative STIT theories with multiple agents. *In International Conference on Temporal Logic* (pp. 332-348). Springer, Berlin, Heidelberg.