On the Use of Causal Rules to Specify How Trust Impacts Change in Knowledge and Belief

Aaron Hunter
British Columbia Institute of Technology, Burnaby, Canada

Abstract

We are concerned with models of trust, and the way that trust impacts changes in knowledge and belief. In order to address this problem, we consider transformations on Kripke structures in modal logic. We start with simple trust rules that specify how reported information impacts the truth of a modal formula. Trust rules are defined with respect to a particular source. So an individual trust rule basically indicates when a source's reported information will cause the truth value of a modal formula to change. In the context of epistemic logic, this means that a trust rule indicates when a report will affect the knowledge or belief of the recipient. We demonstrate how a set of trust rules defines a model transformation in which the underlying accessibility relation is modified to ensure all rules are satisfied. This transformation captures how much the underlying source is trusted to impact the recipients perspective on the world. Model transformations of this kind are commonly used to capture belief change in Dynamic Epistemic Logic. What is distinct in our approach is that we show how simple rules can give a compact and flexible specification of trust in an information source. Our approach is also general, in the sense that we consider arbitrary modal change, which means we can capture different conceptions of knowledge and belief. We compare our approach with related work, particularly on the formaliziation of sensing actions. Future directions and applications are also considered.

Keywords: Trust, belief revision, modal logic, knowledge representation

1. Introduction

The notion of trust plays a key role in many areas, ranging from commonsense reasoning to information security. The fundamental question is the following: when one agent receives information from another, how should their beliefs change? In order to answer this question, we need to consider the perceived expertise of the source as well as their honesty.

In this paper, we are concerned with defining a compact formal representation of the trust that one agent holds in another. The basic building blocks of our model are simple causal rules, inspired by classic action languages [1, 2]. We demonstrate how a set of so-called trust rules defines a transformation function over Kripke structures. As such, given a set of trust rules, we can determine what an agent will believe after receiving a report from another agent.

While our focus in this paper is essentially on the interaction of trust and belief, we define our formal approach in a general modal setting. In other words, trust rules are explicitly defined with respect to an arbitrary modal logic. We are then able to determine which modal formulas are true after a report, regardless of the meaning of the modality. It turns out that this can be problematic in general, because the set of trust rules might define a transformation that does not preserve important modal properties. In the case of standard epistemic logic, however, we will see that it is easy to define sets of rules that preserve the equivalence relation. Moreoover, we will see that the trust rules for epistemic logic can actually capture infallible trust in sensing actions.

*aaron_hunter@bcit.ca

This paper makes two main contributions to the literature on trust and epistemic change. First, while existing modal approaches to epistemic change are based on model transformations, this work provides a straightfoward way to specify a model transformation through simple causal rules. Second, by explictly addressing modal change, we provide a general approach that applies to a variety of different epistemic operators. We take the position that incrementally building a model transformation through simple causal rules provides a flexible and natural way to define a full picture of trust based on a few known causal relationships.

2. **Preliminaries**

2.1. **Motivation**

Suppose that we are receiving reported information from a baker. It makes sense to believe them when they send reports about bread, or even food in general. On the other hand, it does not make sense to trust them when they send reports about Python programming. We are interested in providing a simple, compact representation of trust in situations where a source that is only trusted over certain domains.

Fundamentally, the important thing about trust is that it allows us to determine when another agent can convince us to change our beliefs. In the case of the baker, what is important is that we *believe* the things that they say about bread are true - even if we initially did not believe them. Hence, a model of trust must specify how our beliefs change when we receive information from another agent. This requires two things:

- A formal model of knowledge or belief.
- A specification of how this model changes when new information is given.

In this paper, we use modal logics to represent knowledge or belief. Therefore, the second point requires a mechanism for mapping one Kripke structure to another. Due to the complexity of these structures, this can be hard. Our solution is to specify trust in terms of a set of rules that say "when the report is X, you should believe Y." In this paper, we show how a set of rules of this form actually defines a clear mapping on Kripke structures. This in turn defines a model of epistemic change that takes trust in to account.

2.2. **Notation and Conventions**

We assume the reader is familiar with propositional modal logic, as outlined in [3]. For definiteness, we briefly introduce our notation. A propositional formula is a combination of atoms, using the usual set $\{\neg, \wedge, \vee\}$ of propositional connectives. Given a propositional signature F, a literal is an element of F or an element of F prefixed with the negation symbol. The complement of a literal f will be denoted by $\bar{f}$.

Our focus here will be on logics with a single unary modal operator. A Kripke structure is a triple $\mathcal{M} = \langle M, R, \pi \rangle$, where M is a non-empty set, R is a binary accessibility relation on M and π associates a subset of M with every atomic formula. The satisfaction relation $\mathcal{M}, m \models \phi$ is given by the standard recursive definition; we omit the mention of $\mathcal{M}$ if it is clear from the context.

Let $\mathcal{L}$ be a modal logic given by a set of axiom schemata, and let Π be a set of Kripke structures. We say that $\mathcal{L}$ is *determined by* Π if the set of theorems of $\mathcal{L}$ is identical to the set of formulas valid in Π. In practice, we will refer to a modal logic either by a set of axioms or by a set of Kripke structures, depending on which presentation is more convenient for the task at hand.

Many important modal logics are determined by placing natural restrictions on the accessibility relation. We mention three examples: KT, KB, and $K4$. These modal logics are determined by the classes of structures in which the accessibility relation is reflexive, symmetric, and transitive, respectively. If we combine all three restrictions then the accessibility relation must be an equivalence relation, and we have the standard *epistemic logic* $S5$. The intuition is that two worlds are related by the accessibility relation if they are indistinguishable to the underlying agent. Standard *doxastic logic* is the logic $KD45$, in which the accessibility relation is serial, transitive and Euclidean. For a detailed discussion of modal logics of knowledge and belief, we refer the reader to [4].

2.3. Trust and Belief Change

The problem of *belief change* is concerned with modeling the way that an agent's beliefs change when they receive new information. The most influential model of belief change is the so-called AGM approach, in which the beliefs of an agent are represented by a set of propositional formulas [5]. However, in this paper, we are not explicitly concerned with AGM belief change, because we take a modal approach to representing beliefs.

We model belief change through transformations on Kripke structures. In other words, we assume that the initial beliefs of an agent are given by a Kripke structure and we demonstrate how this structure is mapped to a new one in response to new information. This is similar to the *model transformations* used to model belief change in Dynamic Epistemic Logic(DEL) [6, 7]. For simplicity, we restrict attention to logics with a single modal operator.

The transformations that we introduce will be defined with respect to a particular reporting agent. The trust that is held in that individual will be specified by a set of causal rules that explicitly describe the formulas over which they are trusted.

2.4. Overview

In the next section, we introduce a simple mechanism for describing the trust that one agent holds in another. We show how a model transformation representing belief change can be defined with respect to simple sets of causal rules. Although the present system is very simple and relatively limited in the transition relations that it can represent, the epistemic variant is sufficiently expressive to describe planning domains with sensing actions and incomplete information.

After introducing the basic formalism, we consider formal closure properties. In other words, we consider when a set of rules preserves the fundamental properties of the underlying accessibility relation. Finally, we show how our model of trust can be used to capture infallible sensing actions in a formal model for reasoning about action effects. We conclude with some directions future work.

3. Trust Rules

3.1. Syntax

Assume a fixed propositional vocabulary $\mathbf{F}$, and let $\mathcal{L}$ be a fixed modal logic with a single unary modal operator $\Box$. Let $\mathbf{I}$ be a set of propositional formulas over $\mathbf{F}$, which we call the set of *reports*. Informally, $\mathbf{I}$ is he set of formulas that a particular agent might provide as an information update.

The following definition is the starting point for our formal approach.

Definition 1. *A* trust rule *is an expression of the form*

$$\alpha \textbf{ causes } \phi \textbf{ if } PRE$$

where $\alpha \in \mathbf{I}$, PRE is a formula, and ϕ is a formula of the form $\Box\psi$ for some ψ.

Notice that ψ need not be a literal; any formula can be the modal effect. A *trust description* is a set of trust rules.

The format of trust rules is clearly borrowed from the action language $\mathcal{A}$ [8, 9]. The main difference is that $\mathcal{A}$ is concerned with the effects of *actions*; these are events that change the state of the world. In our setting, there are no actions; instead there are reports represented by formulas. Hence, trust rules do not specify any changes to the actual state of the world; reports only modify the modal accessibility relation . Nevertheless, we will see later that the are clear connections here with action languages with sensing effects.

Let $\mathcal{L}_K$ denote the logic $S5$ with unary modal operator K; we think of K as a modal knowledge operator. In this context, it is natural to think of propositions of the form

$$\phi \textbf{ causes } K\phi \textbf{ if } PRE$$

as descriptions of sensing action effects.

3.2. Semantics

We associate a transition function with each set of trust rules. To be precise, we need transition functions between Kripke structures. With each trust description TD, we define a transition function Φ_{TD}. With this notation, $\Phi_{TD}(\mathcal{M}, \alpha)$ denotes the Kripke structure that results when α is reported in the structure $\mathcal{M}$.

The following definition associates a transition function Φ_{TD} with a trust description TD.

Definition 2. *Let TD be a trust desription. Let $\mathcal{M} = \langle M, R, \pi \rangle$ be a Kripke structure for $\mathcal{L}$ and let $\alpha \in \mathbf{I}$. The Kripke structure $\Phi_{TD}(\mathcal{M}, \alpha) = \langle M^*, R^*, \pi^* \rangle$ is defined as follows.*

(1) $M^* = M$
(2) $m \in \pi^*(F)$ iff $m \in \pi(f)$
(3) $R^*(m_1, m_2)$ holds iff the following both hold

- $(m_1, m_2) \in R$
- *there is no rule in TD of the form*

$$\alpha \textbf{ causes } \Box\phi \textbf{ if } PRE$$

where $\mathcal{M}, m_1 \models PRE$ and $\mathcal{M}, m_2 \models \neg\phi$

Hence, given a structure $\mathcal{M}$ and a report α, we construct $\Phi_{TD}(\mathcal{M}, \alpha)$ as follows:

(1) The fluent values of each world are unchanged.
(2) For each modal effect $\Box\phi$ and each world m, remove all edges Rm_1m_2 where $\mathcal{M}, m_1 \models PRE$ and $\mathcal{M}, m_2 \models \neg\phi$.

It is easy to see that this procedure gives the correct result. Note that preconditions are always evaluated in the initial Kripke structure, rather than the successor structure. The accessibility relation is changed in the successor structure.

We remark that, in the transition between structures, edges are never added. Hence, in the case of knowledge, we can think of reports as *refinements* to the agent's knowledge. Certainly it would be interesting to consider actions that reduce an agent's knowledge as well; such actions could be represented by action effects of the form $\neg\Box\phi$. We leave this problem for future work, and restrict our attention to simple refinements for the present paper.

The following example illustrates how to apply the basic definitions.

Example We represent a situation with a single agent inside a room with no window. The agent receives messages from a friend, indicating whether or not it is raining outside. Let TD denote the trust description containing the following propositions:

$$\mathit{Rain} \textbf{ causes } K(\mathit{Rain}) \textbf{ if } \mathit{Rain}$$
$$\mathit{Rain} \textbf{ causes } K(\neg \mathit{Rain}) \textbf{ if } \neg \mathit{Rain}.$$

Informally, the first proposition says that a report of rain causes the agent to know it is raining, provided that it is in fact raining. The second proposition makes the parallel assertion for non-raining reports. Together these reports essentially assert that the reporting agent is *trusted* to know if it is raining or not.

Suppose that $\mathcal{M} = \langle M, R, \pi \rangle$ is a structure where the accessibility relation is universal. We construct $\Phi_{TD}(\mathcal{M}, Rain)$.

According to Definition 2, the set of worlds M and the function π both remain unchanged. Hence, all that changes is the accessibility relation. Let $m \in M$ and suppose $m \in \pi(Rain)$. Due to the first proposition in TD, we need to remove all edges from m to worlds where it is not raining. So, in $\Phi_{TD}(\mathcal{M}, Rain)$, the world m will be related to a world m' if and only if $m' \in \pi(Rain)$. Similarly, by the second proposition in TD, we remove all edges from non-raining worlds to raining worlds. The resulting accessibility relation is an equivalence relation with two equivalence classes that partition the worlds based on the value of $Rain$. This result is consistent with the intuitive interpretation of the accessibility relation as an indistinguishability relation; after receiving a report of rain, the agent is able to distinguish raining worlds from non-raining worlds.

The preceding example highlights an interesting issue. In particular, one might observe that the effects of both rules in TD are obtained by adding a $\Box$ to the preconditions. By constrast, one might be interested in the interpretation of a trust description TD' containing the single proposition

$$Rain \ \textbf{causes} \ K(Rain).$$

This proposition asserts that a report of rain causes the agent to know it is raining, whether or not it is actually raining. Suppose that $\mathcal{M}$ is an $S5$ structure containing a world m where it is not raining. Let $\mathcal{M}' = \Phi_{TD'}(\mathcal{M}, Rain)$. Applying Definition 2, it is clear that m is a world in $\mathcal{M}'$ but m is not related to itself in $\mathcal{M}'$. Informally, the edge (m, m) is removed in the transition between structures. Therefore $\mathcal{M}'$ is not an $S5$ structure, because the accessibility relation is not reflexive.

Clearly this is a problem. The transition function between Kripke structures is intended to describe epistemic change in a static world. Presumably, however, the fundamental nature of knowledge should not be changed when a report is received. Does this mean that trust descriptions like TD' are pathological? We suggest that the status of TD' depends on the modal logic of interest. For example, if we are interested an $S5$ modality, then we would like to assure that the transition functions defined by trust descriptions always map equivalence relations to equivalence relations. Hence, for epistemic logic, we want to say that TD is an admissible trust description, but TD' is not admissible because it does not preserve reflexivity. For some other modal logics, however, TD' may be perfectly acceptable. For example, in a modal logic of belief, we may allow trust descriptions like TD' because preserving reflexivity would not be important.

In a general modal setting, we would like to ensure that trust descriptions preserve all of the important structural characteristics of the modality under consideration. Preservation properties of this sort are the topic of the next section.

3.3. **Standard Modal Logics**

Let Π be a class of Kripke structures, and let TD be a trust description. We say that TD *preserves* Π if $\Phi_{TD}(\mathcal{M}, \alpha) \in \Pi$ whenever $\mathcal{M} \in \Pi$.

Definition 3. *Let $\mathcal{L}$ be a modal logic determined by a class of structures Π. A trust description TD is* admissible *for $\mathcal{L}$ just in case TD preserves Π.*

We now provide restricted classes of trust descriptions that preserve some natural systems of modal logic. We start with the modal logic KT, which is the modal logic where the accessibility relation is reflexive.

Proposition 1. *Let TD be a set of trust rules such that $PRE \models \phi$ for every rule in TD of the form*

$$A \textbf{ causes } \Box\phi \textbf{ if } PRE.$$

Then TD is admissible for teh modal logic KT.

Proof. Let TD be a set of trust rules satisfying the premise, let α be a report, let $\mathcal{M} = \langle M, R, \pi \rangle$ with R reflexive, and let $\Phi_{TD}(\mathcal{M}, \alpha) = \langle M, R^*, \pi^* \rangle$. Suppose that $(m, m) \notin R^*$ for some $m \in M$. Since R is reflexive, we know that $(m, m) \in R$. Informally, this means that the edge (m, m) is removed in the transition between structures. Hence, there must be some rule

$$\alpha \textbf{ causes } \Box\phi \textbf{ if } PRE$$

in TD such that $\mathcal{M}, m \models PRE$ and $\mathcal{M}, m \models \neg\phi$. This contradicts our assumption that ϕ is a logical consequence of PRE. □

We now consider the modal logic KB, which consists of Kripke structures where the accessibility relation is symmetric.

Proposition 2. *Let TD be a set of trust rules such that, for every rule in TD of the form*

$$\alpha \textbf{ causes } \Box\phi \textbf{ if } PRE,$$

TD also contains a rule of the form

$$\alpha \textbf{ causes } \Box\neg PRE \textbf{ if } \neg\phi.$$

Then TD is admissible for the modal logic KB.

Proof. Let TD be a set of trust rules satisfying the premise, let α be a report, let $\mathcal{M} = \langle M, R, \pi \rangle$ with R symmetric, and let $\Phi_{TD}(\mathcal{M}, \alpha) = \langle M, R^*, \pi^* \rangle$. Suppose that $(m, n) \in R^*$ and $(n, m) \notin R^*$ for some m, n. Since $(n, m) \notin R^*$, there must be some rule of the form

$$A \textbf{ causes } \Box\phi \textbf{ if } PRE$$

in TD, where $\mathcal{M}, n \models PRE$ and $\mathcal{M}, m \models \neg\phi$. By assumption, TD also contains

$$A \textbf{ causes } \Box\neg PRE \textbf{ if } \neg\phi.$$

Since $\mathcal{M}, m \models \neg\phi$ and $\mathcal{M}, n \models PRE$, it follows from Definition 2 that $(m, n) \notin R^*$, which is a contradiction. Hence R^* is symmetric. □

Finally, we look at the modal logic $K4$ where the accessibility relation is transitive.

Proposition 3. *Let TD be a set of trust rules such that, for every rule in TD of the form*

$$\alpha \textbf{ causes } \Box\phi \textbf{ if } PRE,$$

TD also contains a rule of the form

$$\alpha \textbf{ causes } \Box\Box\phi \textbf{ if } PRE.$$

Then TD is admissible for the modal logic $K4$.

Proof. Let TD be a set of trust rules satisfying the premise, let α be a report, let $\mathcal{M} = \langle M, R, \pi \rangle$ with R transitive, and let $\Phi_{TD}(\mathcal{M}, \alpha) = \langle M, R^*, \pi^* \rangle$. Assume that $(m, n) \in R^*$ and $(n, p) \in R^*$. Now suppose $(m, p) \notin R^*$, so there is some rule

$$\alpha \textbf{ causes } \Box\phi \textbf{ if } PRE$$

in TD such that $\mathcal{M}, m \models PRE$ and $\mathcal{M}, p \models \neg\phi$. By assumption, TD also contains

$$\alpha \textbf{ causes } \Box\Box\phi \textbf{ if } PRE.$$

But then, since $\mathcal{M}, m \models PRE$ and $(m,n) \in R^*$, it follows that $\mathcal{M}, n \models \Box\phi$. Then, since $(n,p) \in R^*$, we must have $\mathcal{M}, p \models \phi$. This is a contradiction, hence R^* is transitive. □

I follows that a trust description TD that satisfies the preconditions of each proposition will be admissible for $S5$. Note, however, that the conditions in Propositions 1-3 are sufficient, but not necessary. So there are also many $S5$-admissible trust descriptions that do not satisfy the given conditions. For example, there are certainly finite action descriptions that preserve $S5$, but every trust description satisfying the condition in Proposition 3 must be infinite.

Giving a constructive definition of all admissible trust descriptions for any interesting modal logic $\mathcal{L}$ is a non-trivial problem. For some natural modal logics, it is clear that no simple syntactic characterization can be given. For example, specifying a useful class of descriptions that preserve seriality is difficult, due to the fact that we only allow refinements. As a result, the current framework has somewhat limited applicability to logics determined by non-reflexive, serial structures.

4. **Related Formalisms**

4.1. **Sensing Actions**

The intention of our formalism is to provide a way to describe the trust held in a source, and then use that model of trust to represent epistemic change due to reports. However, there is a special case that has previously been studied in the literature: *sensing actions*. We can view sensing actions as infallible reports that come from an agent's own senses. In this section, we look at a related formalism in which the effects of sensing actions are defined by rules.

Consider the epistemic action languge $\mathcal{A}_L$ [10], which we briefly introduce presently. In this framework, we assume a set of action symbols that are partitioned into sensing actions and non-sensing actions. In $\mathcal{A}_L$, there are two kinds of propositions. First of all, if A is a non-sensing action, f is a literal, and PRE is a conjunction of literals then

$$A \textbf{ causes } f \textbf{ if } PRE$$

is a proposition of $\mathcal{A}_L$. If A is a sensing action, f is a fluent symbol, and PRE is a conjunction of literals, then

$$A \textbf{ causes to know } f \textbf{ if } PRE$$

is a proposition. A set of propositions is called an *action description*.

Note that, in the rules for sensing actions, the intended interpretation is that the execution of A causes the agent to know the truth value of f. This contrasts with our trust rules, which assert that a report causes a certain modal formula to be true. Informally, $\mathcal{A}_L$ is making a higher level assertion about the property f rather than a direct assertion about the truth value of a formula.

We remark that non-deterministic action effects can also be represented in $\mathcal{A}_L$ through a third propositional form. However, we will not consider non-deterministic effects in this paper.

Given an action description AD, we say that f is a *potential sensing effect* of A if AD contains a proposition of the form

$$A \textbf{ causes to know } f \textbf{ if } PRE.$$

The *knowledge precondition* of a fluent symbol f with respect to a sensing action A is the disjunction of all of the preconditions appearing in propositions involving the action A and the sensing effect f.

The semantics of $\mathcal{A}_L$ uses the notion of a *situation*. A situation is a set of *states* and a state is an interpretation of the set of fluent symbols. A fluent f is true in a situation Σ if it is true in every state in Σ, it is false if it is false in every state in Σ and it is unknown otherwise. Truth or falsity in $\mathcal{A}_L$ is understood to reflect the knowledge of an agent, and knowledge is understood to be correct but not necessarily complete.

The semantics of $\mathcal{A}_L$ associates a transition relation Φ_{AD} with every action description AD. We give the definition for the special case where each action has at most one potential sensing effect f. Let Σ be a situation and let A be an action symbol. The triple (Σ, A, Σ^*) is in Φ_{AD} if and only if the following hold.

(1) If A is non-sensing, then the interpretation associated with each world in Σ^* is the interpretation obtained by updating the worlds of Σ in accordance with the $\mathcal{A}$ propositions in AD.
(2) If A is sensing, and f is unknown with precondition P, then Σ^* satisfies one of the following three conditions
   (a) Σ^* is the set of states in Σ where P and f hold
   (b) Σ^* is the set of states in Σ where P and $\neg f$ hold
   (c) Σ^* is the set of states in Σ where $\neg P$ holds

Hence, given a pair (Σ, A) where A has a single potential sensing effect, there will generally be three possible successor situations. A set of situations is called an *epistemic state*. Hence, the semantics of $\mathcal{A}_L$ actually maps a situation and an action to an epistemic state.

We illustrate the intuition behind the the effects of sensing actions with an example.

Example Consider the proposition

$$\mathit{Listen}\ \textbf{causes to know}\ \mathit{MusicOn}\ \textbf{if}\ \neg \mathit{EarPlugs}.$$

If an agent executes the action *Listen*, there are 3 possible outcomes.

(1) The agent learns that *MusicOn* is true.
(2) The agent learns that *MusicOn* is false.
(3) The agent does not learn the value of *MusicOn*.

The only way the third possibility can happen is if the agent is wearing ear plugs. Hence, if the agent listens and still does not know the value of *MusicOn*, then the agent is justified in concluding that *EarPlugs* is true.

In general, each action may have several potential sensing effects. We briefly outline how the definition above can be extended to handle multiple sensing effects. We say that a situation is (f, P)-admissible with respect to an action A if it satisfies the definition given above. Now suppose that A has n potential sensing effects $f_1, \ldots, f_n$ with corresponding knowledge preconditions $P_1, \ldots, P_n$. In this case, $\Phi_{AD}(\Sigma, A, \Sigma^*)$ holds if and only if Σ^* is the intersection of n situations $\Sigma_1, \ldots, \Sigma_n$ where each Σ_i is (f_i, P_i)-admissible with respect to A. We refer the reader to [10] for the details.

4.2. **Translation from $\mathcal{A}_L$**

In this section, we translate $\mathcal{A}_L$ into our framework. To begin, we present the translation and give an intuitive explanation.

Definition 4. *Let AD be an action description in $\mathcal{A}_L$. The trust description $\tau(AD)$ is obtained from AD as follows.*

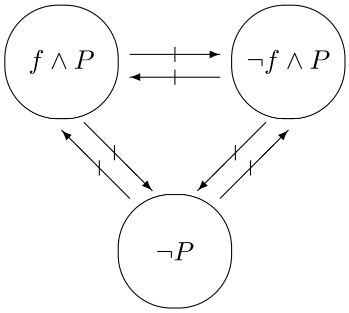

Figure 1. Partitioning Accessibility

(1) For each action A with potential sensing effect f and knowledge precondition P, $\tau(AD)$ contains the following propositions:

$$f \textbf{ causes } K(f \wedge P) \textbf{ if } f \wedge P$$
$$f \textbf{ causes } K(\neg f \wedge P) \textbf{ if } \neg f \wedge P$$
$$f \textbf{ causes } K\neg P \textbf{ if } \neg P.$$

Suppose that AD is an action description involving an action A with a single potential sensing effect f with knowledge precondition P. If A is executed, then $\Phi_{\tau(AD)}$ maps a Kripke structure $\mathcal{M}$ to a new structure $\mathcal{M}'$ in which the accessibility relation is refined as illustrated in Figure 1. Each circled region represents the set of worlds in which the indicated formula is true. The edges of $\mathcal{M}$ that go between the circled regions are removed in $\mathcal{M}'$. Clearly, the three circled regions together form a partition of the universe. This observation suggests that action descriptions in the image of τ will preserve equivalence relations. We formalize this claim in the following proposition.

Proposition 4. *Let AD be a set of $\mathcal{A}_L$ propositions, where each action has at most one potential sensing effect. The trust description $\tau(AD)$ is admissible for the modal logic $S5$.*

Proof. Let $\mathcal{M} = \langle M, R, \pi \rangle$. Let A be a sensing action with potential sensing effect f and knowledge precondition P. Let $\Phi_{\tau(AD)}(\mathcal{M}, f) = \langle M, R^*, \pi \rangle$. We remark that π has remained unchanged because A is a sensing action. We prove that R^* is an equivalence.

By Proposition 1, R^* is reflexive. Moreover, it is straightforward to modify the proof of Proposition 2 to prove that R^* is symmetric. All that remains is to show that R^* is transitive. Suppose that R^*mn and R^*np, but not R^*mp. There are three possible cases to consider.

(1) $m \models f \wedge P$ and $p \models \neg(f \wedge P)$
(2) $m \models \neg f \wedge P$ and $p \models \neg(\neg f \wedge P)$
(3) $m \models \neg P$ and $p \models P$

Suppose the first case holds. Since R^*mn, it must be the case that $n \not\models \neg(f \wedge P)$. Since R^*np, it must be the case that $n \models \neg(f \wedge P)$. Hence, the first case is not possible. The other two cases lead to similar contradictions. Therefore R^* is transitive. □

Next, we will prove that $\tau(AD)$ defines the same transformation on models that is defined by AD. First, we illustrate that there is a natural way to turn an epistemic state E into a Kripke structure $\mathcal{M}_E$. For the moment, assume that the collection of situations in E are pairwise disjoint. We discuss this assumption below. Define $\mathcal{M}_E = \langle M, R, \pi \rangle$ as follows.

(1) $M = \bigcup E$

(2) $R(m_1, m_2)$ iff there is $\Sigma \in E$ such that $m_1, m_2 \in \Sigma$
(3) for any fluent f, $m \in \pi(f)$ iff $f \in m$

Clearly R is an equivalence relation and, moreover, each $\Sigma \in E$ corresponds to the equivalence class $[s]$ generated by $s \in \Sigma$. If Σ is a situation, we write $\mathcal{M}_\Sigma$ as an abbreviation for $\mathcal{M}_{\{\Sigma\}}$.

The assumption that the elements of E are pairwise disjoint is a simplifying assumption to ensure that each state in each situation in E corresponds to a unique element in the universe of $\mathcal{M}_E$. Without this assumption, we can still define a natural structure representing E by using a universe of ordered pairs where one component is an interpretation s and the other component is a situation $\Sigma \in E$ containing s. However, for our purposes, it is sufficient to consider the restricted case described above.

The following result demonstrates the close relationship between $\mathcal{A}_L$ and the semantics of trust descriptions.

Theorem 1. *Let AD be an $\mathcal{A}_L$ action description, let Σ, Σ^* be non-empty situations, and let A be a sensing action in AD with potential sensing effect f. Then $\Phi_{AD}(\Sigma, A, \Sigma^*)$ if and only if Σ^* is an equivalence class in $\Phi_{\tau(AD)}(\mathcal{M}_\Sigma, f)$.*

Proof. Let A be a sensing action with potential sensing effect f with knowledge preconditions P.

Note that $\mathcal{M}_\Sigma = \langle \Sigma, R, \pi \rangle$, where R is universal and $p \in \pi(s)$ iff $p \in s$. Let $\Phi_{\tau(AD)}(\mathcal{M}_\Sigma, f) = \langle \Sigma, R^*, \pi \rangle$.

The relation R^* is obtained by making the following changes to R, for each $i \leq n$:

(1) remove edges $(m, n) \in R$ where $m \models (f \wedge P)$ and $n \models \neg(f \wedge P)$
(2) remove edges $(m, n) \in R$ where $m \models (\neg f \wedge P)$ and $n \models \neg(\neg f \wedge P)$
(3) remove edges $(m, n) \in R$ where $m \models \neg P$ and $n \models P$

Hence $R^* mn$ if and only if one of these conditions holds:

(1) $m, n \models f \wedge P$
(2) $m, n \models \neg f \wedge P$
(3) $m, n \models \neg P$

By definition, this holds if and only if $\Phi_{AD}(\Sigma, A, \Sigma^*)$. □

Intuitively, Theorem 1 says that, under a natural translation between situations and structures, Φ_{AD} and $\Phi_{\tau(AD)}$ represent the same transition relation. Hence, we can intuitively capture complete trust in sensing actions through trust descriptions.

It is worth noting that there is another very similar action language $\mathcal{A}_B$ for representing sensing effects [11]. Just as we have seen for $\mathcal{A}_L$, we can also translate the representation of sensing actions in $\mathcal{A}_B$ into trust descriptions.

There is also a well-known approach to modelling changes in knowledge and belief in the Situation Calculus [12]. The Situation Calculus is essentially a variant of first-order logic, in which a history of actions executed is formally represented by a *situation*. In the epistemic Situation Calculus, there is an accessibility relation on situations that captures the notion of knowledge (or belief). This accessibility relation can be modified by performing sensing actions. The work presented here differs from this approach in two respects. First of all, we are not concerned with infallible sensing actions; we are concerned with reports from other agents. This is the context where trust plays an important role. Second, our work is based on a much simpler model of action effects. Nevertheless, the fundamental ideas are quite similar. We suggest that the work presented here could be translated to the epistemic Situation Calculus to capture that same kinds of problems. We leave this translation for future work.

4.3. Trust Formalisms

The interaction between trust and belief change has been explored in several recent works. One notable example is the work on trust-sensitive revision operators [13]. A trust-sensitive revision operator is based on an underlying AGM revision operator * along with a partition Π over possible states that indicates which states are indistinguishable to the reporting agent. Informally, trust-sensitive revision works by expanding every reported set of states to the set of states that the reporting agent can not distinguish. After this expansion, belief revision is carried out in the usual way.

The basic intuition behind trust-sensitive revision can be captured by sets of trust rules of the form:

$$\alpha \textbf{ causes } K\psi \textbf{ if } PRE$$

where α-states are seen as indistinguishable from ψ-states. Through rules of this form, we can define a trust description in which a report of ϕ will be seen as equivalent to the disjunction of all formulas ψ in the conclusion of a trust rule. Of course, our approach is still more expressive in that we are able to capture nested beliefs, since our approach is based on modal logic.

There is also related work on the impact of trust on belief change in a modal setting [14]. In this work, each agent is essentially associated with a set of formulas over which they are trusted. This can be captured in our framework by having "sensing style" rules as in Example 1, but only for the set of formulas over which a particular agent is trusted.

5. Discussion

5.1. Future Work

There are several directions in which we would like to extend the present work. First of all, we would like to formally address rules that add edges to the accessibility relation. Such rules are useful for describing complex trust relationships where agents might make false reports. Moreover, some combination of adding and removing edges is required to give natural descriptions that preserve seriality. This is an important concern for the representation of some natural modal logics. For example, if we would like to represent change in the context of deontic logic, we need to be able to preserve seriality.

The second extension we would like to consider would allow multiple agents, each with their own associated trust descriptions. This would allow us to consider belief change with different trust constraints simultaneously. This introduces new challenges, because we then need to deal with situations where we have conflicting reports from multiple agents. At present, our formalism does not have any notion of *strength of trust*, so resolving this kind of conflict is difficult. We would like to be able to give a compact treatment for comparing these kinds of trust relationships.

Third, we would like to be able to implement a planner for a less restricted class of modal action languages. In particular, we would like to allow some limited nesting of modal operators in our descriptions. Such nesting is required to address the representation of simple knowledge games, and it is also required for the verification of communication protocols. We would be interested in demonstrating the practical utility of modal action languages by solving some realistic verification problems.

5.2. Conclusion

There are many cases where a notion of trust impacts modal change. By combining a modal logic with simple causal rules, we can create a simple tool for representing and reasoning about change in such an environment. In this paper, we have provided a simple model for describing trust relationships and for reasoning about transitions between Kripke structures. The paradigmatic example has been the representation of changes in the knowledge of a single agent. We have seen that the modal approach naturally subsumes existing approaches to reasoning about epistemic action effects, and it requires little formal machinery on top of elementary modal logic and causal rules.

References

[1] C. Baral and M. Gelfond. “Reasoning About Effects of Concurrent Actions”. In: *Journal of Logic Programming* 31.1-3 (1997), pp. 85–117.

[2] C. Baral, M. Gelfond, and A. Provetti. “Representing Actions: Laws, Observations and Hypothesis”. In: *Journal of Logic Programming* 31.1-3 (1997), pp. 201–243.

[3] B. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, 1980.

[4] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning About Knowledge.* MIT Press, 1995.

[5] C. Alchourrón, P. Gärdenfors, and D. Makinson. “On The Logic of Theory Change: Partial Meet Functions for Contraction and Revision”. In: *Journal of Symbolic Logic* 50.2 (1985), pp. 510–530.

[6] A. Baltag and S. Smets. “Dynamic belief revision over multi-agent plausibility models”. In: *Proceedings of LOFT.* Vol. 6. 2006, pp. 11–24.

[7] J. van Benthem. *Logical Dynamics of Information and Interaction.* New York, NY, USA: Cambridge University Press, 2014.

[8] M. Gelfond and V. Lifschitz. “Representing Action and Change By Logic Programs”. In: *Journal of Logic Programming* 17 (1993), pp. 301–321.

[9] E. Giunchiglia and V. Lifschitz. “Action Languages, Temporal Action Logics and the Situation Calculus”. In: *Linköping Electronic Articles in Computer and Information Science* 4.40 (1999), pp. 1–19.

[10] J Lobo, G. Mendez, and S. Taylor. “Knowledge and the Action Description Language $\mathcal{A}$”. In: *Theory and Practice of Logic Programming* 1.2 (2001), pp. 129–184.

[11] T. Son and C. Baral. “Formalizing sensing actions: A transition function based approach”. In: *Artificial Intelligence* 125.1-2 (2001), pp. 19–91.

[12] S. Shapiro, M. Pagnucco, Y. Lespérance, and H. J. Levesque. “Iterated belief change in the situation calculus”. In: *Artificial Intelligence* 175.1 (2011), pp. 165–192.

[13] R. Booth and A. Hunter. “Trust as a Precursor to Belief Revision”. In: *Journal of Artificial Intelligence Research* 61 (2018), pp. 699–722.

[14] E. Lorini, G. Jiang, and L. Perrussel. “Trust-based belief change”. In: *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI).* 2014, pp. 549–554.