Comparison between Causal and ATL-style logic approaches to modeling responsibility and blameworthiness.

Vincent Hoffmann - 6054226

July 2021

Abstract

The capability to model responsibility and blameworthiness is becoming more important as automated systems increase in complexity. So, in this literature review the causal model described by Halpern [7, 9] and the logical model based on ATL-style semantics [10, 11, 12, 13] are compared in their approach to modeling responsibility and blameworthiness. Further, the causal models' definitions are translated in logic and vise versa. First an example situation described, which the models will be based upon. Next the literature is explained and the models and their approaches to modeling responsibility and blameworthiness are defined. This is followed by comparing their modeling results and expressing the the causal models' definitions in logic. So we can see how a causal model can be expressed in logic when modeling responsibility and blameworthiness.

Contents

1 Introduction

Everyone is familiar with concepts such as responsibility and blame. You probably have been punished by your parents for doing something wrong or have been responsible for a task or event. Because the notions of responsibility and blame are such a common thing for us, their definitions are consistently reviewed[1, 2]. Furthermore, because the definitions are consistently reviewed, there is a lot of discussion about the definitions. Interestingly, Frankfurt noticed a dominant role in these discussions is what he calls the "Principle of Alternative Possibilities" which states that: "a person is morally responsible for what she has done only if she could have done otherwise"[3]. The principle of alternative possibilities, which is also referred to as "counterfactual possibility"[4], is also used when defining causality[5, 6].

The principle of alternative possibilities speaks about a person being responsible for their actions. However with us creating more capable AI, we also need to start looking at agents being responsible and blamable. For with the introduction of self-driving cars, instances where agents are responsible for their action start to appear. Take the following example for instance.

We take a t-split crossroad like the one in Figure 1, agent A is headed towards the crossroad and has full vision over it. At the same time, agent B is also headed towards the crossroad, but from the opposite direction of agent A. From the right of agent B is agent C nearing the crossroad. Agents B and C cannot see one another, for there is an obstruction preventing vision between the two. If both agents B and C keep heading towards the crossroad, they will collide. But if either or both brakes collision can be prevented.

In this case though, suppose agents B and C would collide it is easy to conclude both are partially responsible for the event, but neither of them is to blame. For neither did know and despite agent A knowing what was about to happen, it cannot intervene in any sense so is also not responsible nor to blame.

However, nowadays, communication is possible between self-driving cars. So let's assume all the agents are self-driving cars. From this assumption three new scenarios arise; the first scenario is where agent A communicates with neither, the second scenario is where agent A communicates with either of the agents and the third instance is where agent A communicates with both agents.

The **first** scenario of the above would mean that agent A communicated with neither of the other agents despite capable of doing so. However, if consequently a collision is to happen, agent A is neither responsible nor blamable for agent A can not prevent the actions that agents B and C will make. For, agent A could have influenced their actions, but not prevented them. Still, because an accident happened, agents B and C are responsible for the collision. However, not blamable, for neither knew the consequence of their actions would result into a collision.

In the **second** scenario, agent A sends a warning to either agent B or C. This makes it possible for the other agent to respond accordingly and prevent the collision. However, assume an instance where the receiving agent does not respond accordingly, suppose due to a system or brake failure, then agents B and

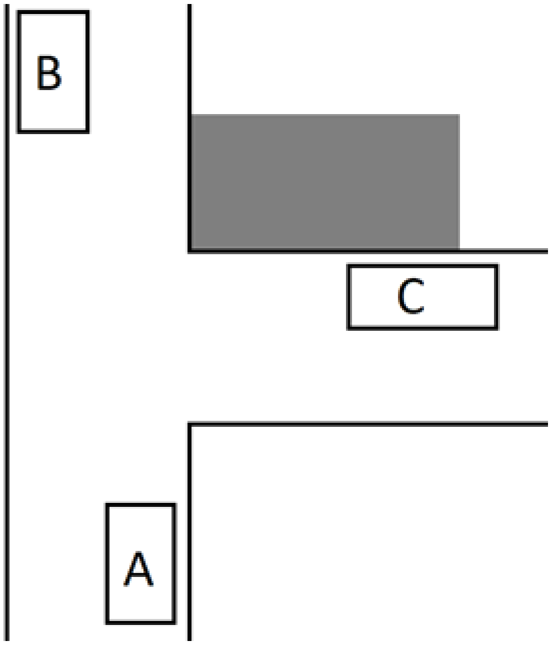

Figure 1: Example Case

C are responsible. For agent A can not prevent the collision, he cannot be held responsible nor to blame for the collision. But the signal receiving agent is to blame for the collision. Since, the receiving agent should be capable enough to prevent collision. The non-receiving agent is still responsible for it has collided but did not know and is therefore not to blame.

In the **third** scenario, agent A sends a warning to both agents. As a result, agent A cannot be called responsible or be blamed for the collision, since he acted optimally to prevent the accident and cannot intervene directly. But if a collision still happened, then agent B and C are both responsible and to blame, assuming they want to prevent the collision.

Based on this example, agent A does not seem to be responsible for any of the results, which is correct. For this paper's intend is to look at responsibility and blame and not review positive actions. So, the positive effects of agent A's action are ignored.

The exact intent of this literature review, is to compare between causal models and logical models in their approaches to modeling responsibility and blameworthiness and further answer the question if it is possible to express a causal models' definitions in logic and vise versa. Thus, we want to compare the approaches to modeling responsibility and blameworthiness between the two models. Further, we want to translate a causal models' definitions in ATL-style logic and vise versa.

This will be achieved by comparing the causal model of Halpern [7, 9] with the logical approach based on ATL-style semantics [10, 11, 12, 13]. We start with explaining both structures in this paper. First the causal model, then followed by the ATL-style model. Which we then follow with comparing the responsibility and blameworthiness modeling results of both approaches and then finally translate the causal models' definitions in ATL-style logic and vise versa.

The causal model will be explained first in Section 2 followed by the explanation of the logical approach 3. In both Sections, the structures are applied to the example situation. The resulting models are used to make a comparison and to translate the causal models' result to the ATL-style model and vise versa in Section 4. Which will be followed by our results and discussion.

2 Causal Approach

In this section, we review the causal models we are going to use. The models used are the same models described by (Alechina, Halpern & Logan [9]), which is an extension of the model described by (Chockler & Halpern [7]) with a corrected definition of causality[8]. So, in this section we follow main definitions described in the corresponding section in (Alechina, Halpern & Logan [9]).

2.1 Causal Models

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature* that is, a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a finite set of *exogenous* variables, $\mathcal{V}$ is a finite set of *endogenous* variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a finite nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y *ranges*), and $\mathcal{F}$ is a function that associates with each endogenous variable $X \in \mathcal{V}$ a function denotes F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \to \mathcal{R}(X)$. That is, F_X describes how the value of the endogenous variable X is determined by the values of all other variables in $\mathcal{U} \cup \mathcal{V}$.

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, ..., Y_k \leftarrow y_k]\varphi$, where φ is a Boolean combination of primitive events, $Y_1, ..., Y_k$ distinct variables in $\mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$. Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, ..., Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, ..., k$.

Following from Halpern's[8] definition, we restrict attention to *acyclic* causal models, where there is a total ordering $\prec$ of the endogenous variables (the ones in $\mathcal{V}$) such that if $X \prec Y$, then X is independent of Y , that is, $F_X(\vec{x}, y, \vec{z}) = F_X(\vec{x}, y', \vec{z})$ for all $y, y' \in \mathcal{R}(Y)$. If $X \prec Y$, then the value of X may affect the value of Y, but the value of X cannot be affected by the value of Y. So, if M is an acyclic causal model, then given a *context*, which is a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, there is a unique solution for all the equations: we simply solve for the variables in the order given by $\prec$.

A pair $(M, \vec{u})$ consisting of a causal model and a context is called a causal setting. A causal formula ψ is true or false in a causal model, given a context. We write $(M, \vec{u})] \models \psi$ if the causal formula ψ is true in causal model M given context $\vec{u}$. The $\models$ relation is defined inductively. $(M, \vec{u})] \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$ (i.e., the unique vector of values for the exogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \vec{u})] \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y}=\vec{y}}, \vec{u}) \models \varphi$. Thus, $[\vec{Y} \leftarrow \vec{y}]\varphi$ is true in $(M, \vec{u})$ if φ is true in the model that results after setting the variables in $\vec{Y}$ to $\vec{y}$.

2.2 Definition of Causality

Using this as background, we can now give the definition of causality. Causes are conjunctions of primitive events, abbreviated as $\vec{X} = \vec{x}$. Arbitrary Boolean combinations of primitive events are what can be caused. So, in other words, $\vec{X} = \vec{x}$ is a cause of φ if, $\vec{X} = \vec{x}$ had not been the case, φ would not have happened. To deal with many well-known examples, the actual definition is somewhat more complicated.

Definition 2.1. (Cause): $\vec{X} = \vec{x}$ is an actual cause of φ in $(M, \vec{u})$ *if the following three conditions hold:*

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ *and* $(M, \vec{u}) \models \varphi$

AC2^m. *There is a set* $\vec{W}$ *of variables in* $\mathcal{V}$ *and settings* $\vec{x}'$ *of the variables in* $\vec{X}$ *and* $\vec{w}$ *of the variables in* $\vec{W}$ *such that* $(M, \vec{u}) \models \vec{W} = \vec{w}$*) and*

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$$

AC3. $\vec{X}$ *is minimal; no subset of* $\vec{X}$ *satisfies conditions AC1 and AC2^m.*

AC1 just says that for $\vec{X} = \vec{x}$ to be a cause of φ, both $\vec{X} = \vec{x}$ and φ have to be true. AC2^m (the "m" is for modified; the notation is taken from Halpern[8]) captures the counterfactual. The counterfactual is, the idea that A is a cause of B and if, had A not happened, B would not have happened[5]. This is the standard "but-for" test used in the law: but for A, B would not have occurred. The old model definition used by Chockler & Halpern[7] did satisfy the condition as well, but was brittle for more contingencies. So in Halpern[8]'s paper, he placed more stringent restrictions on the contingencies that can be considered. Resulting in the current AC2^m. Which says that if the value of $\vec{X}$ is changed from $\vec{x}$ to $\vec{x}'$, while possibly holding the set $\vec{W}$, which contains the values of the variables in some (possibly empty), fixed at their values in the current context, then φ becomes false. We say that $(\vec{W}, \vec{x}')$ is a witness to $\vec{X} = \vec{x}$ being a cause of φ in $(M, \vec{u})$. If $\vec{X} = \vec{x}$ is a cause of φ in $(M, \vec{u})$ and $X = x$ is a conjunct of $\vec{X} = \vec{x}$, then $X = x$ is part of a cause of φ in $(M, \vec{u})$. As for AC3, it is a minimality condition, which ensures that only the conjuncts of $\vec{X} = \vec{x}$ that are essential are parts of a cause. In general, there may be multiple causes for a given outcome.

2.3 Example Model

It is possible to make a causal model from the Example 1 and model we described. From the example it is clear there is only one end state which either can be true or false. We define it as $C = 1$ for collisions and $C = 0$ for no collision. Because C is dependent of either agent B or C braking, we have two states that imply C defined as B_B and B_C being 1 if the agents used the brakes and 0 if not. Because agent A can pass information to the other agents just before they act, we add a state called IT representing agent A transferring its information. We add two states to represent if information is given to the agents called T_B and T_C being 1 if information is received and 0 if not.

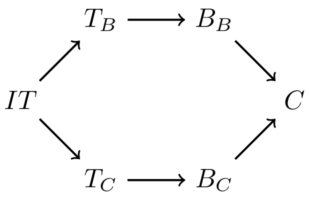

Figure 2: Causal Example Model

Following from the description, we get Figure 2. The model itself is not complex, but when we start assigning responsibility and blame we take each possible action for each case into account. The whole model can be considered complex.

2.4 Responsibility and Blame

To describe responsibility for causal models Chockler and Halpern[7] introduced the *notion of degree of responsibility.* It is intended to capture the intuition that if everybody is responsible for a failure, that each person is less responsible for the failure than if one person is responsible for a failure. So, the degree of responsibility $X = x$ for φ, in order to make make φ counter factually depend on $X = x$, measures the minimal number of changes and number of variables that must be held fixed. We use the formal definition by Halpern[6], which is appropriate for the modified definition of causality used here.

Definition 2.2. (Responsibility) *The* degree of responsibility of $X = x$ for φ in $(M, \vec{u})$, *denoted* $dr((M, \vec{u}), (X = x), \varphi)$, *is* 0 *if* $X = x$ *is not part of a cause of* φ *in* $(M, \vec{u})$*; it is* $1/k$ *if there exists a cause* $\vec{X} = \vec{x}$ *of* φ *and a witness* $(\vec{W}, \vec{x}')$

to $\vec{X} = \vec{x}$ being a cause of φ in $(M, \vec{u})$ such that (a) $X = x$ is a conjunct of $\vec{X} = \vec{x}$, (b) $|\vec{W}| + |\vec{X}| = k$, and (c) k is minimal, in that there is no cause $\vec{X}_1 = \vec{x}_1$ for *φ in $(M, \vec{u})$ and witness $(\vec{W}', \vec{x}_1')$ to $\vec{X}_1 = \vec{x}_1$ being a cause of φ in $(M, \vec{u})$ that includes $X = x$ as a conjunct with $|\vec{W}'| + |\vec{X}_1| < k$.*

We assume with this definition of responsibility that everything relevant for the model and how it works is known. Overall though, there are uncertainties both about the context as about the causal model. This is taken into account in the notion of blame. An agent's uncertainty is modeled by a pair $(\mathcal{K}, \Pr)$, where $\mathcal{K}$ is a set of causal settings, which are pairs of the form $(M, \vec{u})$, and Pr is a probability distribution over $\mathcal{K}$. We call such a pair an *epistemic state*. Note that with such a distribution, we can talk about the probability that $X \vec{=} \vec{x}$ is a cause of φ relative to $(\mathcal{K}, \Pr)$: it is just the probability of the set of pairs $(M, \vec{u})$ such that $\vec{X} = \vec{x}$ is a cause of φ in $(M, \vec{u})$. We also define the degree of blame of $X = x$ for φ to be the expected degree of responsibility:

Definition 2.3. (Blame) *The* degree of blame of *$X = x$ for φ relative to the epistemic state $(\mathcal{K}, \Pr)$ is*

$$\sum_{(M,\vec{u}) \in \mathcal{K}} dr((M, \vec{u}), X = x, \varphi) \Pr((M, \vec{u})).$$

2.5 Responsibility Example

We use the definitions of responsibility and blame for the causal model stated in Section 2.4 to assign responsibility and blame to our example causal model described in Section 2.3.

2.5.1 Examples Responsibility

We can assign responsibility to the model by simply using Definition 2.2. Since the Definition 2.2 of responsibility is dependent of Definition 2.1 of causality, we define which states are the cause of the collision. We cannot assign responsibility for an event that did not happen, so in all instances when $C = 0$ there is no one responsible. Based on Definition 2.1, we discover that only states B_B and B_C are the cause of state $C = 1$. From which we conclude that agent B and C are 1/2 responsible for the collision and agent A is not responsible in any case.

2.5.2 Examples Blame

To assign blame using Definition 2.3 to the model, we need to create an epistemic state for it. To create an epistemic state, we need to assign a probability distribution over the model, we do not need to change the models structure. For our example we use the following probability relations:

- $IT = 1 \rightarrow T_{B \vee C} = 0.9$ and $IT = 0 \rightarrow T_{B\&C} = 0$
- $T_{B\&C} = 1 \rightarrow B_{B\&C} = 0.1$ and $T_{B\&C} = 0 \rightarrow B_{B\&C} = 0.8$

So, if agent A transfers information to another agent there is a one in ten chance that the information is not received and when agent B or C has received the information the chance of breaking increases from twenty percent to eighty.

Using this as our probability distribution and the responsibility definition given in the previous section, the degree of blame of agents B and C is equal to 94%. However, with the probability distribution it came apparent that of the 94%, 64% was the cause of agent A not providing information. But since agent A is not responsible, it cannot be blamed for it.

3 Logical approach

In this section, we take a look at the logical approach we are going to use. We will be using *alternating-time temporal logic*(ATL)[10] as the core of our logical approach. For ATL is interpreted over *concurrent game structures*(CGS), we will be describing both. We start by describing the syntax of ATL followed by the semantics, which is dependent on the concurrent game structures. In this section we follow main definitions described in Ågotnes, et al.[11]

3.1 Alternating-Time Temporal Logic

As described at the start of this section we will first describe the syntax of ATL, which is then followed by concurrent game structures. Which is used to describe the semantics of ATL

3.1.1 Syntax

The language of Alternating-Time Temporal Logic (ATL)[10] is built from the following components: $Agt = a_1, ..., a_n$ a set of n agents and Π a set of propositions. Formulas of the language $\mathcal{L}_{ATL}$are defined by the following syntax:

$$\varphi, \psi ::= p | \neg\varphi | \varphi \wedge \psi | \langle\langle C \rangle\rangle \bigcirc \varphi | \langle\langle C \rangle\rangle \varphi \mathcal{U} \psi | \langle\langle C \rangle\rangle \square \varphi$$

where $p \in \Pi$ is a proposition, and $C \subseteq Agt$ is a coalition of agents. Informally, ATL operators be interpreted as follows:

- $\langle\langle C \rangle\rangle \bigcirc \varphi$: means that the coalition C has a collective strategy to ensure that the next state satisfies φ
- $\langle\langle C \rangle\rangle \varphi \mathcal{U} \psi$: means that the coalition C has a collective strategy to ensure satisfying ψ while maintaining the truth of φ
- $\langle\langle C \rangle\rangle \square \varphi$: means that the coalition C has a collective strategy to ensure that φ is always true.

Now we define the CGS our logical structure works upon.

3.1.2 Concurrent Game Structures

The model which ATL is interpreted over are Concurrent Game Models (CGM)[10]. An CGM is a Concurrent Game Structure (CGS) with assigned variable values, therefore we first describe what a CGS is.

Concurrent Game Structures: Formally, a *concurrent game structure* (CGS) is a tuple $\mathcal{S} = \langle Agt, Q, Act, d, o \rangle$ where:

- $Agt = a1, ..., an$ is a finite, non-empty set of agents; subsets of Agt are called *coalitions* C
- Q is a finite, non-empty set of states
- Act is a finite set of atomic actions
- *action manager* function $d : Agt \times Q \mapsto \mathcal{P}(Act)$ specifies the sets of actions available to agents at each state. An *action profile* is a tuple of actions $\alpha = \langle \alpha_1, ..., \alpha_i \rangle \in Act^k$.
- o is a *transition function* that assigns the outcome state $q' = o(q, \alpha_1, ..., \alpha_n)$ to state q and a tuple of actions $\alpha_i \in d(i, q)$ that can be executed by Agt in q.

Having defined what a structure is, we extend the structure (CGS) to a *Concurrent Game Model* (CGM) with a labeling function $L : Q \rightarrow \mathcal{P}(\Pi)$, such that the states of Q are labeled by sets of atomic propositions from a fixed set Π. The labeling describes which atomic propositions are true at a given state.

We use the following auxiliary notions to represent and start reasoning about strategies and outcomes. (References to elements of $\mathcal{M}$ are to elements of a CEGS $\mathcal{M}$ modeling a given multiagent system, e.g., we write Q instead of Q in $\mathcal{M}$.)

Successors and Computations: For two states q and q', we say q' is a *successor* of q if there exist actions $\alpha_i \in d(i, q)$ for $i \in \{1, ..., n\}$ in q such that $q' = o(q, \alpha_1, ..., \alpha_n)$, i.e., agents in Agt can collectively guarantee in q that q' will be the next system state. A *computation* of a CEGS $\mathcal{M}$ is an infinite sequence of states $\lambda = q_0, q_1, ...$ such that, for all $i > 0$, we have that q_i is a successor of q_{i-1}. We refer to a computation that starts in q as a $q-computation.$ For $i \in \{0, 1, ...\}$, we denote the i'th state in λ by $\lambda[i]$, and $\lambda[0, 1]$ and $\lambda[i, \infty]$ respectively denote the finite prefix $q_0, ..., q_i$ and infinite suffix $q_i, q_{i+1}, ...$ of λ. We refer to any two arbitrary states q_i and q_{i+1} as two *consecutive* states in $\lambda[i, \infty]$. Finally, we say a finite sequence of states $q_0, ..., q_n$ is a $q - history$ if $q_n = q, n \geq 1$, and for all $0 \leq i < n$ we have that q_{i+1} is a successor of q_i. We denote a q-history that starts in q_i and has n steps with $\lambda[q_i, n]$

Strategies and Outcomes: A *positional (aka. memoryless) strategy* in $\mathcal{S}$ for an agent $a \in Agt$ is a function s_a: $Q \mapsto Act$, such that $s_a(q) \in d(q, a)$.

For a coalition $C \subseteq Agt$, a *collective strategy* $Z_C = \{s_a | \ a \in C\}$ is an indexed set of strategies, one for every $a \in C$. Then, $out(q, Z_C)$ is defined as the set of potential q-computations that agents in C can enforce by following their corresponding strategies in Z_C. We extend the notion to sets of states $\omega \subseteq Q$ in the straightforward way: $out(\omega, Z_C) = \cup_{q' \in \omega} out(q', Z_C)$.

3.1.3 Semantics

We described CGS for the semantics of ATL since it is defined relative to a CGM $\mathcal{M}$ and state q, where the truth of ATL-formulae at a state $q \in Q$ is defined inductively as follows:

- $\mathcal{M}, q \vDash p$ iff $q \in L(p)$
- $\mathcal{M}, q \vDash \neg\varphi$ iff $\mathcal{M}, q \nvDash \varphi$
- $\mathcal{M}, q \vDash \varphi \wedge \psi$ iff $\mathcal{M}, q \vDash \varphi$ and $\mathcal{M}, q \vDash \psi$
- $\mathcal{M}, q \vDash \langle\langle C \rangle\rangle \bigcirc \varphi$ iff there exists a stratagy Z_C such that for all computations $\lambda \in out(q, Z_C)$, $\mathcal{M}, \lambda[1] \vDash \varphi$
- $\mathcal{M}, q \vDash \langle\langle C \rangle\rangle \varphi \mathcal{U} \psi$ iff exists a strategy Z_C such that for all computations $\lambda \in out(q, Z_C)$, for some $i, \mathcal{M}, \lambda[i] \vDash \psi$, and for all $j < i, \mathcal{M}, \lambda[j] \vDash \varphi$
- $\mathcal{M}, q \vDash \langle\langle C \rangle\rangle \Box \varphi$ iff exists a strategy Z_C such that for all computations $\lambda \in out(q, Z_C)$, for all $i, \mathcal{M}, \lambda[i] \vDash \varphi$

3.2 Concurrent Example Model

We make a CGM out of Example 1's logic. There are three agents: a, b and c. From the example we conclude the set of states Q consists four states: state S_0 which is the initial state, splitting into state S_1 where collision happened and state S_2 where collision did not happen. For S_0 is the initial state and there is no follow states after S_1 and S_2, we only need to describe the available actions from S_0. We specify the actions of the agents as follows: doing nothing is N, transferring information to one agent is T_1, to both is T_2, and breaking is represented by B. Resulting the set of actions based of the example is: $a : \{N, T_1, T_2\}, \ b : \{N, B\}, \ c : \{N, B\}$. Following this we get the transition function o:

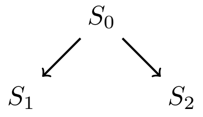

Figure 3: Concurrent Example Model

- $o(S_0, (N, N, N)) = S_1, \ o(S_0, (T_1, N, N)) = S_1, \ o(S_0, (T_2, N, N)) = S_1$

- $o(S_0,(N,B,N)) = S_2,\ o(S_0,(N,N,B)) = S_2,\ o(S_0,(N,B,B)) = S_2,$
  $o(S_0,(T_1,B,N)) = S_2,\ o(S_0,(T_1,N,B)) = S_2,\ o(S_0,(T_1,B,B)) = S_2,$
  $o(S_0,(T_2,B,N)) = S_2,\ o(S_0,(T_2,N,B)) = S_2,\ o(S_0,(T_2,B,B)) = S_2,$

With this we have described the CGM fitted for the example.

3.3 Responsible ATL

In order to assign responsibility and blame to the ATL structure, we extend the structure with the formulae described by Naumov & Tao and Yazdanpanah, et al. [12, 13] From their definition the following syntax gets added:

$$\varphi ::= \langle\!\langle C\rangle\!\rangle R_C\varphi | \langle\!\langle C\rangle\!\rangle B_C\varphi$$

Formula $\langle\!\langle C\rangle\!\rangle R\varphi$ means "coalition C can enforce that it is responsible for φ" and formula $\langle\!\langle C\rangle\!\rangle B\varphi$ means "coalition C can enforce that it is blamable for φ". The addition of this syntax allows us to start speaking about responsibility and blame in ATL. We do not use the notions of Forward and Backward responsibility described by Yazdanpanah, et al.[13] because the definitions given by Naumov & Tao define responsibility sufficiently.

We also extend the model to an epistemic model. For epistemic uncertainty must be considered when analyzing responsibility. As the ability of agents to execute a strategy depends on their knowledge of the environment. So, the CGS we defined gets extended into a Concurrent Epistemic Game Structure(CEGS)[11] by adding an *epistemic indistinguishability relation* $\sim_a \subseteq Q \times Q$ for each agent $a \in Agt$. We assume that $\sim_a$ is an equivalence relation, where $q \sim_a q'$ indicates that states q and q' are indistinguishable to a.

As a result of this addition, our strategy definition needs to be updated as well to handle the equivalence effect of $\sim$. Thus, the definition of our strategy s_a in $\mathcal{S}$ for an agent $a \in Agt$ is turned into: for all $q \in Q$ (1) $s_a(q) \in d(q,a)$, and (2)$q \sim_a q' \rightarrow s_a(q) = s_a(q')$. When referring collective strategies we also use the new strategy definition in the collective. Furthermore we include an extra strategy definition that will be used:

Uniform Strategies: A uniform strategy is a strategy in which agents select the same actions in all states where they have the same information available to them. In particular, if agent $a \in Agt$ is uncertain whether the current state is q or q', then a should select the same action in q and in q'. A strategy s_a for agent $a \in Agt$ is called uniform if for any pair of states q, q' such that $q \sim_a q', s_a(q) = s_a(q')$. A strategy Z_C is uniform if it is uniform for every $a \in C \subseteq Agt$. Realistic modeling of strategic ability under imperfect information requires restricting attention to uniform strategies only.

With this the semantic meaning of the syntax that was just added can be explained. For this is an extension on the original CGM, all other semantics statements also hold. So, the semantics of ATL defined relative to a CEGM $\mathcal{M}$ and state q, where of ATL-formulae at a state $q \in Q$ is defined inductively as follows:

- $\mathcal{M}, q \vDash \langle\langle C \rangle\rangle R\varphi$ iff the following conditions hold
  1. $\mathcal{M}, q \vDash \langle\langle C \rangle\rangle \bigcirc \varphi$
  2. there exists a strategy Z_C and state q such that there exists a computation $\lambda \in out(q, Z_C), \mathcal{M}, \lambda[1] \nvDash \varphi$
  3. for each proper subset $D \subset C$ and each strategy Z_D there exists a computation $\lambda \in out(q, Z_D), \mathcal{M}, \lambda[1] \vDash \varphi$
- $\mathcal{M}, q \vDash \langle\langle C \rangle\rangle B\varphi$ iff the following conditions hold
  1. $\mathcal{M}, q \vDash \langle\langle C \rangle\rangle \bigcirc \varphi$
  2. for all states q' there exists a uniform strategy Z_C such that $q \sim_C q'$ there exists a computation $\lambda \in out(q', Z_C), \mathcal{M}, \lambda[1] \nvDash \varphi$
  3. for each proper subset $D \subset C$ and each strategy Z_D there exists a computation $\lambda \in out(q, Z_D), \mathcal{M}, \lambda[1] \vDash \varphi$

The first item of both definitions state that they can enforce to be responsible or blamable. The second item of the first definitions stands for: if the coalition had an other option then they are responsible. The second definitions stands for: if the coalition had an other option and knew about it then they are blamable. As for the third item guaranties that every member should have an action that contributes to the event.

3.4 Examples Logical Responsibility

Based of the examples CGM and the additional definitions for ATL, we can now start assigning responsibility and blame to the model. For starters there is no indistinguishable state between any of the states in our old model. Thus, we need to extend by adding possibly indistinguishable states, which we have dependent on the information given by agent A. The initial state S_0 can be dived into four states for each scenario, S_{0N}, S_{0PB}, S_{0PC} and S_{0F} which refer to: S_{0N} is the initial state where agent A did not provide information to the agents, S_{0PB} is where agent A provided information to agent B, S_{0PC} is where agent A provided information to agent C, and S_{0F} is where agent A provided information to both agents. So our renewed model looks as follows:

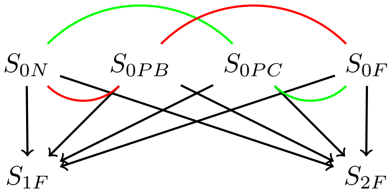

Figure 4: Concurrent Epistemic Example Model

Because of agents B and C do not know of one another in S_{0N} and only partially know in states S_{0PB} and S_{0PC}, are the states S_{0N} and S_{0PB} indistinguishable for agent B and the states S_{0N} and S_{0PC} indistinguishable for agent C. However, since agents B and C know the scenario in S_{0F} and partially know in states S_{0PB} and S_{0PC}, are also the states S_{0PB} and S_{0F} indistinguishable for agent B and the states S_{0PC} and S_{0F} indistinguishable for agent C. However, the states S_{0PB} and S_{0F} are also indistinguishable for agent B and agent C likewise cannot distinguish the states S_{0PC} and S_{0F}. Since, since agents B and C know the scenario in S_{0F} and partially know in states S_{0PB} and S_{0PC}. The indistinguishable states for agent B are connected with red and the indistinguishable states for agent C are connected with green.

Since by the definition of responsibility, agent A does not have an option to prevent collision as for agents B and C are the only cause of it. Because there is a way for agents B and C to enforce a collision by both not using the brakes and being able to prevent the collision by not braking, is $\langle\!\langle \text{B,C} \rangle\!\rangle R_{\text{B,C}} S_1$. Agents B and C are all the agents responsible for the case collision happens, because no knowledge is required to be responsible. Further we assume that there is no obvious solution to the collision such that the agents know that braking is the solution to the collision. The information that agent A can provide however, makes agents B and C additionally to responsible when a collision happens also blamable. For when an agent knows how to prevent the collision, the formula $\langle\!\langle Co \rangle\!\rangle B_{Co} S_1$ becomes true. In other words, if a collision happens but agent A provided information, then that agent with knowledge will also become blamable as well as being responsible. So, agents B and C are responsible for the collision from state S_{0N} but not blamable for could not distinguish the states resulting into $\langle\!\langle \varnothing \rangle\!\rangle B_{\varnothing} S_1$. From S_{0PB} however, agent B is knows how to prevent the collision so $\langle\!\langle \text{B} \rangle\!\rangle B_{\text{B}} S_1$ is true. Same goes for agent C in S_{0PC} from which follows that $\langle\!\langle \text{C} \rangle\!\rangle B_{\text{C}} S_1$ is true. As in state S_{0F} the formula $\langle\!\langle \text{B,C} \rangle\!\rangle B_{\text{B,C}} S_1$ holds and are therefore both agents B and C blamable.

4 Causal Responsibility in Logic

In this section, we compare the models and their results from Sections 2 and 3 and try to logically express the causal definition with the ATL-style logical approach and vise versa. We start by comparing the results of the models, such that we have an understanding what the similarities and differences are between the models. So, when we logically express the causal definition with the logical approach and vise versa. We use the comparison results for the translations in definitions.

4.1 Comparing Models

Here we compare the results of models described in Sections 2.1 and 3.4. We start by looking at the similarities and followed by looking at the differences of the models.

First of the similarities, we noticed that both structures are used to describe the order of actions. It might seem trivial since both describe actions in time, but it still relevant enough when we look at the model structures. Another similarity is that both models agree about the interpretation that agent A is not involved enough in the collision for it to be responsible for the accident. Also, both agree about the fact that both agents B and C are always partially responsible for the collision, if we take into account that every member in a coalition is responsible as defined in Section 3.3.

There is, however, the difference in defining knowledge, structure, and interpreting blame. Well, the difference in defining knowledge is also a cause to why the structures differ. For the causal model uses states to represent knowledge of variable and instances and the states before that current state tell what knowledge is known and what not. As for in an ATL model, the states represents the knowledge that is known. Therefore, resulting into different structures. More importantly, there is the difference in interpreting blame. Since, our ATL model models to full blame or shared blame for a collision, whereas the causal model represents blame with probability values. This is partially the consequence of causal models needing the previous states to obtain the knowledge of the final state. Still, the value of blame differs from the fact that the causal model uses a probability distribution to assign its blame and the ATL model uses its interpreted blame.

4.2 Defining Cause in Logic

For both structures used by the models are so dissimilar, it is necessary to manually do the translation between them. Since, there is an interesting absence of results about translations between two approaches which model responsibility and blame. Still, since we only attempt to express the definition of the causal model in ATL-style logic, it is not necessary to look at the structures of the models itself. But if we were to look at the structures, the biggest difference between the models is that the causal model has the action of agent A integrated in the actions, as for the ATL model has the actions of agent A integrated as initial states.

Still, we only intend to translate the causal model's definitions of responsibility and blame into logic and vise versa. Well, since the definition of responsibility is fairly similar between the approaches, we do not need to modify anything. We do need to, however, add the statement that responsibility in a coalition is equally shared between members of the coalition.

As for approaching the definition of blame there is the effect in the causal model that more blame is assigned when the estimation suggests a more probable collision. So if we are to assign the same probabilistic distribution to the ATL model, then the results still differ. So, it is necessary to inverse the probability distribution of the causal model for the ATL model to approach the probability result. Since the ATL model assumes that only with knowledge can somebody be blamable and for the causal model states that based of the probabilistic estimation somebody can be blamable.

As the result of adding the inverse probability distribution there is a blame assignment in the case where both agents B and C have no knowledge, this blame assignment is however a 3% of the total 94% which is ignorable. As for the rest of the probabilities, those align with both the ATL model in blame is assigned to those blamable and the percentages with those blames. And with this we have defined the causal definition of responsibility and blame from a ATL-style logic approach.

To defined the ATL-style logic definition of responsibility and blame from a causal approach the probabilities of the distributions and use the results of the previous states to define blame. So agent B is blamable when T_B of Figure 2 is 1 and agent C when $T_C = 1$. With this we have defined the ATL-style logic definition of responsibility and blame from a causal approach.

5 Results

As the intent of this literature review is to compare between ATL-style logical and causal models approaches to modeling responsibility and blameworthiness and to answer the question if it is possible to express a causal models' definition in ATL-style logic and vise versa.

We started by comparing the approaches on modeling responsibility and blameworthiness. The model descriptions in Sections 2.1 and 3.4 are what came out of the example with the literature definitions. After that, the models were compared in Section 4, which discovered that responsibility is assigned mostly in a similar way but blame differently.

Using the comparison results, we translated the definition of the causal model in ATL-style logic. By adding that responsibility is shared between members of a coalition and an inverse probability distribution for blame, was the causal model described in ATL. For the translation of the definition of the ATL-style logic from the causal model, the probabilistic distribution had to be discarded and tanking previous states as knowledge had to implemented.

6 Discussion

Showing that a causal definition can be translated from a logical approach and vise versa is the result of this literature review. However, this can be the result of the example being too simple and therefore not approach any of the edge cases, which could prove otherwise. Such as the fact that ATL can be cyclic, but since the example was so simple this possible conflict did not happen. Since our chosen causal model was the model defined by Halpern[7, 8, 9], in the case of using a different model a different result could have been observed. Also, in the logical approach the definitions of responsibility and blame Naumov & Tao were taken instead of those defined by Yazdanpanah, et al. [13]. Possibly giving us a different result. It would even be possible to remove the probability difference with using ATL which uses probabilistic terms when defining knowledge.[14]

Furthermore, the logic for our logical approach in our paper is ATL. We could have taken a different type of logic, which possibly would have resulted into a different result as well.

References

[1] Talbert, M. (2019). ”Moral Responsibility”, The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2019/entries/moral-responsibility/

[2] Tognazzini, N., & Coates, D. J. (2021) ”Blame”, The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), forthcoming URL = https://plato.stanford.edu/archives/sum2021/entries/blame/

[3] Frankfurt, H. (1969). Alternate Possibilities and Moral Responsibility. The Journal of Philosophy, 66(23), 829-839. doi:10.2307/2023833

[4] Cushman, F. (2015). Deconstructing intent to reconstruct morality. Current Opinion in Psychology. Volume 6. Pages 97-103. ISSN 2352-250X. https://doi.org/10.1016/j.copsyc.2015.06.003.

[5] D. Lewis. Causation. Journal of Philosophy, 70:556–567, 1973. Reprinted with added “Postscripts” in D. Lewis, Philosophical Papers, Volume II, Oxford University Press, 1986, pp. 159–213.

[6] J. Y. Halpern. Actual Causality. MIT Press, Cambridge, MA, 2016. https://mitpress.mit.edu/books/actual-causality

[7] H. Chockler, J. Y. Halpern. (2004) Responsibility and Blame: A Structural-Model Approach. Journal of Artificial Intelligence Research 22 (2004) 93-115 https://www.aaai.org/Papers/JAIR/Vol22/JAIR-2204.pdf

[8] Halpern, J.Y. (2015) A modification of the Halpern-Pearl definition of causality. In Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), pages 3022–3033, 2015. https://www.cs.cornell.edu/home/halpern/papers/modified-HPdef.pdf

[9] Alechina, N., Halpern, J. Y., & Logan, B. (2020). Causality, responsibility and blame in team plans. arXiv preprint arXiv:2005.10297. https://arxiv.org/pdf/2005.10297.pdf

[10] Rajeev Alur, Thomas A. Henzinger, & Orna Kupferman. (2002). Alternating-time temporal logic. J. ACM 49, 5 (2002), 672–713. https://doi.org/10.1145/585265.585270

[11] Ågotnes, T., Goranko, V., Jamroga, W., & Wooldridge, M. (2015). Knowledge and ability. In Handbook of Epistemic Logic, Hans van Ditmarsch, Joseph Halpern, Wiebe van der Hoek, and

Barteld Kooi (Eds.). College Publications, 543–589. https://www.diva-portal.org/smash/get/diva2:874163/FULLTEXT01.pdf

[12] Naumov, P., & Tao, J. (2020). An epistemic logic of blameworthiness. Artificial Intelligence. Volume 283. 103269. ISSN 0004-3702. https://doi.org/10.1016/j.artint.2020.103269.

[13] Yazdanpanah, V., Dastani, M., Alechina, N., Logan, B., & Jamroga, W. (2019). Strategic responsibility under imperfect information. In Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems AAMAS 2019 (pp. 592-600). IFAAMAS. http://www.ifaamas.org/Proceedings/aamas2019/pdfs/p592.pdf

[14] Huang, X., Su, K., & Zhang, C. (2012). Probabilistic Alternating-Time Temporal Logic of Incomplete Information and Synchronous Perfect Recall. Proceedings of the AAAI Conference on Artificial Intelligence, 26(1). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/8214