Multi-Agent Systems

Reasoning about Actual Causality

Albert-Ludwigs-Universität Freiburg

Bernhard Nebel, Felix Lindner, and Thorsten Engesser

April 17, 2018

Responsibility and Blame

Motivating Example I: Responsibility

Example (Suzy and Billy throwing rocks again)

Suzy and Billy both throw rocks at a bottle, but Suzy's hits the bottle, and Billy's doesn't (although it would have hit had Suzy's not hit first). The bottle shatters.

- To give an argument for why Suzy is a cause for the bottle's shattering (and Billy is not), we had to make adaptions to our model of the situation (viz., witness $(\vec{W} = \{BH\}, \vec{w} = 0, \vec{x}' = 0)$ in modified HP).
- Intuitively, the more adaptions we have to make to prove s.th. a cause for an effect (the bigger $\vec{W}$), the less responsibility we are ready to attribute to the cause.

Motivating Example II: Responsibility

Example (Disjunctive Forest Fire again)

Forest fire breaks out in case there is lightning or a matched lit. As a matter of fact, there was lightning and a matched lit.

- Using but-for cause or the modified HP definition, neither L nor MD is a cause, but both L and MD are part of the cause $L \vee MD$.
- Intuitively, the bigger the cause, the less responsibility we are ready to attribute to the parts of the cause.

Definition: Responsibility

Definition (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 $(M,\vec{u})$;
- **1/k** if there exists a cause $\vec{X} = \vec{x}$ of φ and a witness $(\vec{W},\vec{w},\vec{x}')$ to $\vec{X} = \vec{x}$ to $\vec{X} = \vec{x}$ being a cause of φ in $(M,\vec{u})$ such that
  - (a) $X = x$ is part 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 $(\vec{W}',\vec{w}',\vec{x}_1')$ to being a cause of φ in $(M,\vec{u})$ that includes $X = x$ with $|\vec{W}'| + |\vec{X}_1'| < k$.

Application: Rock Throwing

- Rock Throwing, $(M, (1,1))$
  - But-For Cause: Both $ST = 1, BT = 1$ are part of the cause $ST = 1 \vee BT = 1$.
    - $dr((M,(1,1)),(ST = 1),(BS = 1)) = \frac{1}{|\emptyset|+|\{ST=1,BS=1\}|} = 1/2$
    - $dr((M,(1,1)),(BT = 1),(BS = 1)) = \frac{1}{|\emptyset|+|\{ST=1,BS=1\}|} = 1/2$
  - HP definitions: Only $ST = 1$ is a cause, but we have to make at least one change to the model to prove that.
    - $dr((M,(1,1)),(ST = 1),(BS = 1)) = \frac{1}{|\{BH=0\}|+|\{ST=1\}|} = 1/2$
    - $dr((M,(1,1)),(BT = 1),(BS = 1)) = 0$

Application: Disjunctive Forest Fire

- Disjunctive Forest Fire, $(M, (1, 1))$
  - But-for cause and modified HP definition: $L = 1, MD = 1$ are part of the cause $L = 1 \lor MD = 1$.
    - $dr((M, (1, 1)), (L = 1), (FF = 1)) = \frac{1}{|\emptyset| + |\{L=1, MD=1\}|} = 1/2$
    - $dr((M, (1, 1)), (MD = 1), (FF = 1)) = \frac{1}{|\emptyset| + |\{L=1, MD=1\}|} = 1/2$
  - original and updated HP definition: $L = 1$ and $MD = 1$ are seperate causes with witnesses $(\{MD\}, 0, 0)$ and $(\{L\}, 0, 0)$, respectively.
    - $dr((M, (1, 1)), (L = 1), (FF = 1)) = \frac{1}{|\{MD=0\}| + |\{L=1\}|} = 1/2$
    - $dr((M, (1, 1)), (MD = 1), (FF = 1)) = \frac{1}{|\{L=0\}| + |\{MD=1\}|} = 1/2$

Voting Case

Example (Voting)

Ben is a republican and thus always votes for the republican candidate. Ralf always votes for the democrate. Jonas is a swinger, and this time votes for the republican. Who is responsible for the republican candidate to win the election?

Epistemic States: Motivation

- The attribution of blame (rather than responsibility) requires to take some agent's epistemic state before the actual situation occured into account.
- A responsible agent might have been uncertain about the actual outcome, and therefore deserves less blame.
- Two sources of uncertainty:
  - What values the (exogeneous) variables have, i.e., uncertainty about $\vec{u}$.
    - E.g., in the conjunctive Forest Fire, you consider possible that there was no lightning.
  - How the world works, i.e., uncertainty about M.
    - E.g., you consider possible that only lightnings cause fire but not lit matches.

Epistemic States: Definition

Definition (Epistemic State)

An agent's epistemic state is given by a pair $(\mathcal{K}, Pr)$, where $\mathcal{K}$ is a set of situations $(M, \vec{u})$, and Pr is a probability distribution over $\mathcal{K}$.

- Additional assumption: In case this definition is used to compute a degree of blame to $X = x$, it is assumed that $(M, \vec{u}) \models X = x$ for all $(M, \vec{u}) \in \mathcal{K}$ holds.
- Justifications for the assumption: If we ask for the degree of blame to $X = x$, we take the occurence of $X = x$ as given.

Definition: Blame

Definition (Blame)

The degree of blame of $X = x$ for φ relative to epistemic state $(\mathcal{K}, Pr)$, denoted $db(\mathcal{K}, Pr, X = x, \varphi)$ is

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

Example: Disjunctive Forest Fire

- Consider the following situations:
  - $(M_1,(1,1))$: Fire breaks out if $L = 1$ or $MD = 1$, both of which hold.
  - $(M_2,(1,1))$: Fire breaks out if $L = 1$, which is the case. $MD = 1$ also holds, but does not cause fire.
- How much blame does the lit match deserve for $FF = 1$, if:
  - $\mathcal{K} = \{(M_1,(1,1))\}, Pr((M,\vec{u})) = 1$?
    - $1/2 \cdot 1 = 1/2$
  - $\mathcal{K} = \{(M_2,(1,1))\}, Pr((M,\vec{u})) = 1$?
    - $0 \cdot 1 = 0$
  - $\mathcal{K} = \{(M_1,(1,1)),(M_2,(1,1))\}, Pr((M,\vec{u})) = 1/2$?
    - $(1/2 \cdot 1/2) + (0 \cdot 1/2) = 1/4$

Note: Obliged Epistemic State

Example (Doctor)

A doctor treats a patient with a particular drug. The doctor does not know the drug would have a side effect which kills the patient.

- Especially in legal contexts, to determine blame, it may be more relevant to represent what should have been known (probably along with a representation of what actually was known).

Discussion: Case I

Example (Voodoo Case)

People who believe in voodoo think that by sticking pins in a doll they are thereby bringing about the agonizing death of their enemy. Suppose a voodoo believer sticks a pin in a doll with the intention of killing her enemy. Is the sticking of a pin in a doll impermissible? (Scanlon 2009: 46)

Discussion: Case II

Example (Alfred's Case)

Alfred, whose wife is dying, and whose death he wishes to hasten, buys a stuff-thought-to-be-poison and gives it to his wife with the intention of thereby hastening her death. Unbeknown to him, the stuff is the only cure for his wife. Is giving the cure-thought-to-be-poison to the wife impermissible? (Thomson 1991: 293)

Discussion: Case III

Example (Tennis Player's Case)

A debt collector in danger of dying is rescued by an agent who hates him and would be happy to see him die. The agent, who happens to be a tennis player, saves him because she believes that the debt collector will break Maria's finger before their match. The agent's belief is false. The debt collector has no harmful intention toward her opponent, Maria. Is saving the victim impermissible? (suggested by Scanlon 2009)

Psychology of Counterfactual Reasoning

- Modeling various types of counterfactual thinking
  - Additive Upward: “If I started studying three days ago, instead of last night, I could have done better on my test.”
  - Subtractive Upward: “I should have never started drinking, then life would be much easier.”
  - Additive Downward: “If I went drinking last night as well, I would have done even worse.”
  - Subtractive Downward: “If I didn’t start studying two days ago, I would have done much worse.”

Possible Topics for Projects and Theses

- Models of Relief & Regret: Robot expresses relief and regret, understands human's relief and regret. Tells human things could have turned out worse to make them feel better.
- Learning from failure: Robot understands when it did wrong and adapts behavior accordingly. Tells humans how they could have done better.
- Means and Side effects: In various ethical theories, this distinction is essential to moral permissibility judments.
- Explanations and Justifications
  - Justifications: Robot justifies a decision it has made, or tells human how to justify his/her decision.
  - Explanation: Takes the epistemic state of the addressee into account, viz., if I ask the robot to explain some phenomenon to me, I might not want it to tell me things I already know.
- Special topics: Thankworthiness, Volition & Blame

- In the counterfactual world, where no one of you attended the MAS lecture, the lecture would not have been a success. Thanks for attending and

  Good luck for the exams :-)

Literature

Literature I

Pearl, J., Mackenzie, D.
The Book of WHY – The New Science of Cause and Effect,
Basic Books, 2018.

Halpern, J. Y.
Actual Causality,
MIT Press, 2016.