ÉTHIQUE ET CAUSALITÉ

Conférence invitée *CNIA 2023*

4 juillet 2023

Gauvain Bourgne

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

Ethique computationelle

Ethique computationnelle

Objet

- Intégrer des considérations éthiques aux (processus décisionnels des) IAs
- Utiliser l'IA pour formaliser/représenter/simuler des raisonnements éthiques

Un champs vaste

- Multi-disciplinaire : IA & Philosophie, mais aussi sciences sociales, psychologie...
- Nombreuses approches et outils
- Ici : approches prescriptive, avec représentations explicites à base de formalismes logiques

Ethique computationnelle

Objet

- Intégrer des considérations éthiques aux (processus décisionnels des) IAs
- Utiliser l'IA pour formaliser/représenter/simuler des raisonnements éthiques

Différentiation

- Ethique computationnelle : Pt de vue interne. Développement d'outil en IA pour l'éthique (-> outils et modèles).
- Cyber éthique : Pt de vue externe. Etude des impacts éthiques de l'IA (-> régulation, bonnes pratiques).

Raisonnement éthique

Tâche

Evaluer éthiquement une décision

Raisonnement éthique

Tâche

Evaluer éthiquement une décision

- pour juger de sa conformité éthique (a posteriori / vérification)
- pour orienter le processus décisionnel (prise de décision éthique)

Raisonnement éthique

Tâche

Evaluer éthiquement **une décision**

Décision (entrée)

- prise dans un contexte donné
- au sein d'un ensemble de choix possibles
- Nature de la décision
  - Action
  - Plan (séquentiel, contingent, stratégie)
  - Politique

Raisonnement éthique

Tâche

Evaluer éthiquement une décision

Evaluation (sortie)

- Choisir la meilleure option (-> une action)
- Filtrer les actions permissibles (-> un ensemble d'actions)
- Ordonner les actions de la plus éthique à la moins éthique (-> ordre ou pré-ordre sur les actions)
- Evaluer numériquement chaque actions (-> scores)

Raisonnement éthique

Tâche

Evaluer **éthiquement** une décision

Principe éthique (méthode/paramètre)

Définir la bonne façon de déterminer ce qui est juste est en philosophie tout l'objet de l'éthique normative.
Pas de consensus, plusieurs approches concurrentes définissant autant de principes

L'éthique en philosophie

Carte concise

- Méta-éthique
  - Statut et sens des concepts éthiques
- Ethique normative
  - Déterminer/comparer/expliquer ce qui constitue un comportement éthique (ou juste)
- Ethique appliquée
  - Définir/Appliquer les règles éthiques d'un environnement particulier
  - Nombreux sous-domaine : Bioéthique, éthique médicale, éthique des affaires, cyberéthique...

L'éthique en philosophie

Approches d'éthique normative

- Conséquentialisme (J.S. Mill, G.E. Moore)
  - fondé sur le **résultat** des actions
  - promeut le Bien
- Déontologie (W.D. Ross)
  - accorde une valeur morale intrinsèque à l'**action** (conforme à des règles de conduite)
  - promeut le Devoir
- Ethique de la vertu (Aristote, A. Comte-Sponville)
  - centré sur l'**agent** (raisons d'agir, état d'esprit, relations, expérience)
  - promeut la Réalisation de soi.

L'éthique en philosophie

SCIENCES SORBONNE UNIVERSITÉ

La dialectique des dilemmes

- Expériences de pensées
  - situations simplifiées
  - enjeux extrême
  - pas d'action
- Contraste les théories
  - construit pour attaquer une théorie ou justifier ou alternative
- Dilemmes classique
  - Polémique sur le droit de mentir (B. Constant et I. Kant)
  - Dilemme du tramway

L'éthique en philosophie

Le dilemme du Tramway

- Attaque de l'utilitarisme

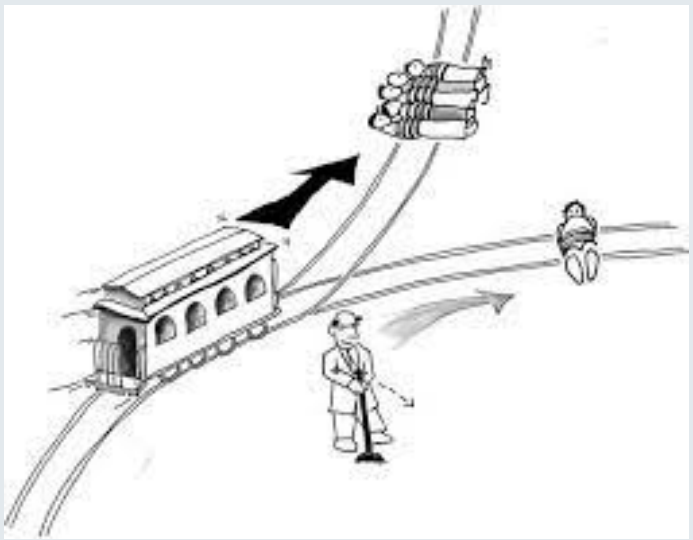


Figure – cas du 'switch'

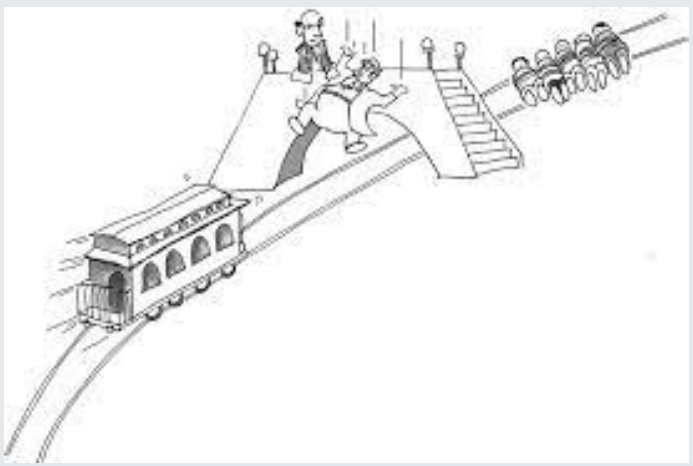


Figure – cas du 'push'

Raisonnement éthique

Tâche

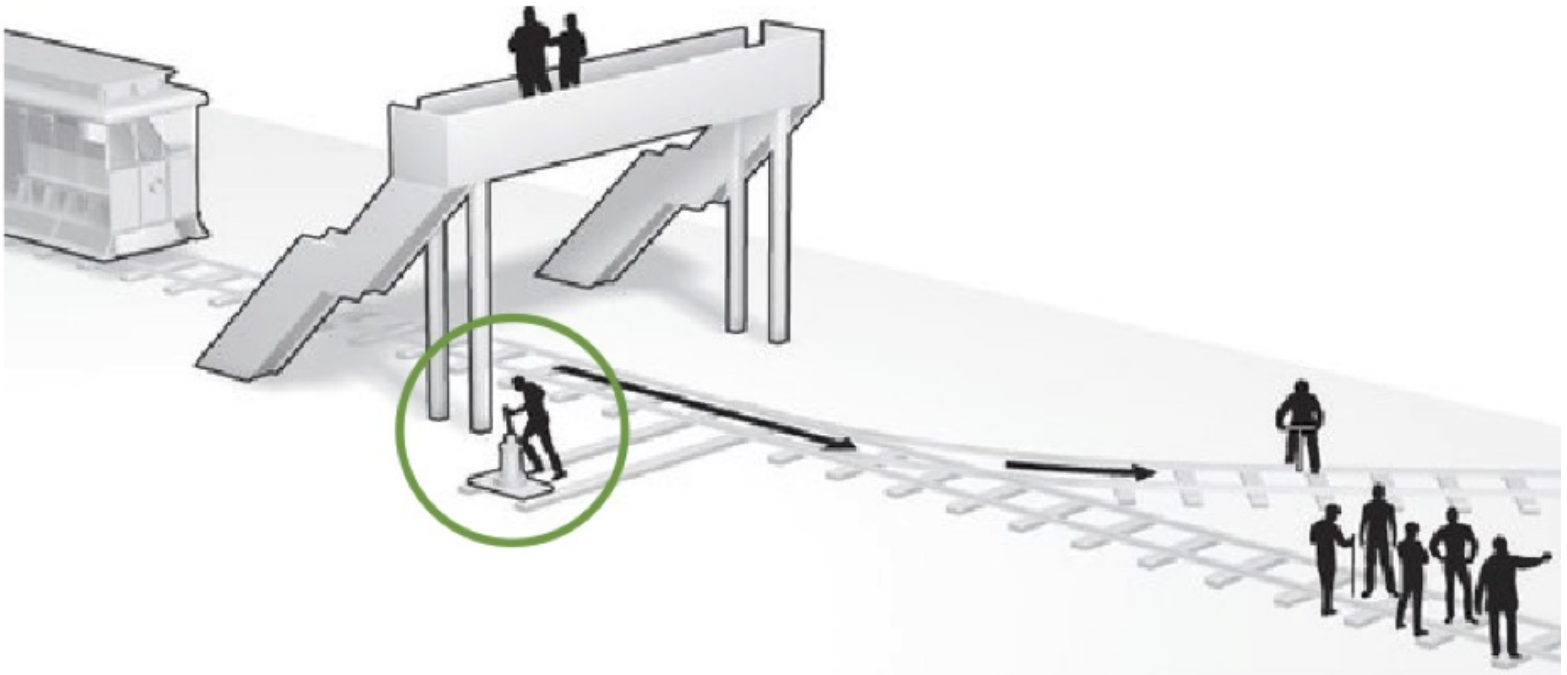
Evaluer une décision selon un principe éthique

Eléments

- Entrées
  - Un ensemble d'actions (ou scénarios)
  - Un contexte (situation initiale / évolutions possibles)
- Paramètre
  - Un ou plusieurs principes éthiques
- Sortie
  - Ensemble des actions permissibles selon chaque principe

Cadre ACE

Retour sur le Tramway



Structure

- switch -> kill(1) & switch -> save(5)
- push -> kill(1) -> save(5)

Découle de chaînes d'actions et réactions qu'il faut modéliser

Cadre ACE

Tâche

Evaluer des actions selon différents principes éthiques

Formalisme unifié

- Formaliser et comparer différents principes éthiques en les représentant explicitement dans un même langage
- Distinctions importantes
  - Séparer le factuel de l'éthique
  - Séparer les faits/entrées spécifiques des axiomes généraux
- Modulaire et implémenté en ASP

3 niveaux : Action / Causalité / Ethique

Cadre ACE

ACE

Action - Causalité - Ethique

Modèle d'action

Description factuelle d'une situation dynamique

- Entrée : un ou plusieurs scénario
- Composants
  - Contexte
    - Situation initiale
    - Spécification des événements
  - Moteur d'action (mécanisme de projection temporelle)
- Sortie : une trace d'évènements (états et événements)

Cadre ACE

ACE

Action - **Causalité** - Ethique

Modèle causal

Articulation entre la couche d'action et la couche éthique.

- Entrée : les traces et spécifications issues de la couche d'action
- Sortie : une trace causale (relations causales entre les événements de la trace d'événement)

Nécessaire pour faire ressortir la structure

- Raisonnement sur fins et moyens (e.g. DDE)
- Lier une action à ses conséquences réelles (conséquentialisme, expression d'interdit du type 'ne pas provoquer la mort')

Cadre ACE

ACE

Action - Causalité - **Ethique**

Couche éthique

Evaluation de la permissibilité des actions

- Entrée : les traces d'événements et causales
- Composants
  - Théories du Juste : définition générale de chaque principe
  - Spécifications éthiques : éléments propres au domaine ou à la situation (définition du Bien ou du Devoir, attribution de valeurs...)
- Sortie : ensemble des actions permissibles selon chaque principe

Cadre ACE

ACE

Action - Causalité - Ethique

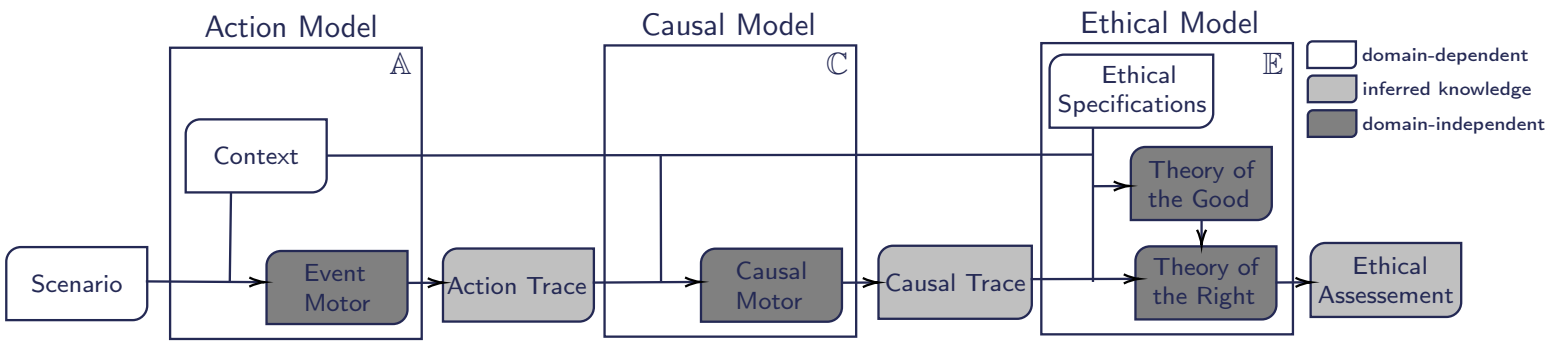


Figure – Architecture modulaire ACE

Principes éthiques

Approches conséquentialistes

Evaluer les actions selon leurs conséquences

Utilitarisme de l'acte

L'approche la plus classique.

- Principe d'optimisation
  1. Déterminer toutes les conséquences de l'action
  2. Associer une utilité à chacune (quantification du Bien)
  3. Sont permissibles les actions maximisant l'utilité totale (somme).

Principes éthiques

Approches conséquentialistes

Evaluer les actions selon leurs conséquences

Principes éthiques

Approches conséquentialistes

Evaluer les actions selon leurs conséquences

Théorie de la Valeur

- Utilitarisme
  - Utilitarisme des préférences (satisfaction des préférences individuelles)
  - Hédonisme (plaisir)
  - Utilitarisme idéal (liste objective de biens : beauté, amitié...)
- Pluralisme : plusieurs valeurs non reductibles
  - Conséquentialisme pluraliste

Sous-tend de la Théorie du Bien, quantification atomique des conséquences dans une situation donnée.

Principes éthiques

Approches conséquentialistes

Evaluer les actions selon leurs conséquences

Variantes de théories du Juste

- Conséquentialisme de l'acte ...
  - ... maximisant
  - ... satisfaisant (relatif ou absolu)
  - ... contraint
- Conséquentialisme de la règle
  Forme de justification utilitariste de règle déontologiques.

Principes éthiques

Approches déontologiques

Evaluer les actions selon leur conformité à des règles de conduite

Forme de règles

- Code de conduite comme série de contraintes dures
  - Règles logique pour prescrire le Devoir
  - Eventuellement quelques méta-règles
  - Gestion des conflits ?
- Priorité ou hiérarchie entre règles
  - prima facie duties
  - lois d'Aasimov

Principes éthiques

Approches déontologiques

Evaluer les actions selon leur conformité à des règles de conduite

Exemples

- Code déontologique (issu de l'éthique appliquée)
- Commmandements divins
- Doctrine du double effet :

  `imp(dde2,AcOcc) :- causes(AcOcc,BEvOcc`, bad(BEvOcc), causes(BEvOcc,GEvOcc), good(GEvOcc).
- 2nde formulation de l'impératif catégorique.

  Agis de façon telle que tu traites l'humanité, aussi bien dans ta personne que dans toute autre, toujours en même temps comme fin, et jamais simplement comme moyen. nécessite de rajouter des spécification sur les individus affectés par des évènements.

Principes éthiques

Approches déontologiques

Evaluer les actions selon leur conformité à des règles de conduite

Construction des maximes

- 1ère formulation de l'impératif catégorique de Kant

  Agis uniquement d'après la maxime qui fait que tu puisses vouloir en même temps qu'elle devienne une loi universelle.

- Utilitarisme de la règle
  Règle évalué par l'utilité de son application dans l'ensemble des cas.

Notion d'universalisation

Principes éthiques

Incertitude morale

Considérer plusieurs théories morales comme crédibles

Principe

- Ensemble de théorie avec des poids.
  - Choiceworthiness : on considère que chaque théorie évalue quantitativement les alternatives
  - Approche ordinale : chaque théorie se contente de classer les alternatives
- Méthode : formes d'agrégation ou de compromis (e.g. maximal choiceworthiness).

Causalité réelle

Modèle causal

Objet du modèle causal

Proposer une définition de la causalité réelle (actual causality) fondée sur le langage d'action du modèle d'action.

- Qu'est-ce que la causalité réelle ?
- Pourquoi est-ce important ?
- Quels sont les limitations de l'état de l'art ?

Modèle causal

Objet du modèle causal

Proposer une définition de la **causalité réelle** (actual causality) fondée sur le langage d'action du modèle d'action.

- Qu'est-ce que la causalité réelle ?
- Pourquoi est-ce important ?
- Quels sont les limitations de l'état de l'art ?

Modèle causal

Objet du modèle causal

Proposer une définition de la causalité réelle (actual causality) fondée sur le langage d'action du modèle d'action.

- Qu'est-ce que la causalité réelle ?
- Pourquoi est-ce important ?
- Quels sont les limitations de l'état de l'art ?

Qu'est-ce que la causalité Réelle ?

Causalité réelle

Determiner les éléments qui ont fait qu'un événement réel spécifique soit arrivé (ou qu'un certain état du monde ait été atteint)

Au sein de la causalité

Diffère de la *causalité générique* (type causality)

- Causalité générique : *Speeding cause accidents* (analyse prédictive a priori)
- Causalité réelle : *The fact that Caitlyn sped caused her to have an accident today* (explication a posteriori)

La causalité générique est un énorme enjeu en causalité, notamment en apprentissage (apprendre des corrélations ne permet pas d'agir)

Qu'est-ce que la causalité Réelle ?

Causalité réelle

Determiner les éléments qui ont fait qu'un événement réel spécifique soit arrivé (ou qu'un certain état du monde ait été atteint)

Example 1 : Suzy and Billy

Suzy and Billy throws a rock at a bottle. Both aim correctly, but Suzy's rock hit first and shatters the bottle. What caused the bottle to shatter ?

Qu'est-ce que la causalité Réelle ?

Causalité réelle

Determiner les éléments qui ont fait qu'un événement réel spécifique soit arrivé (ou qu'un certain état du monde ait été atteint)

Example 1 : Suzy and Billy

Suzy and Billy throws a rock at a bottle. Both aim correctly, but Suzy's rock hit first and shatters the bottle. What caused the bottle to shatter ?

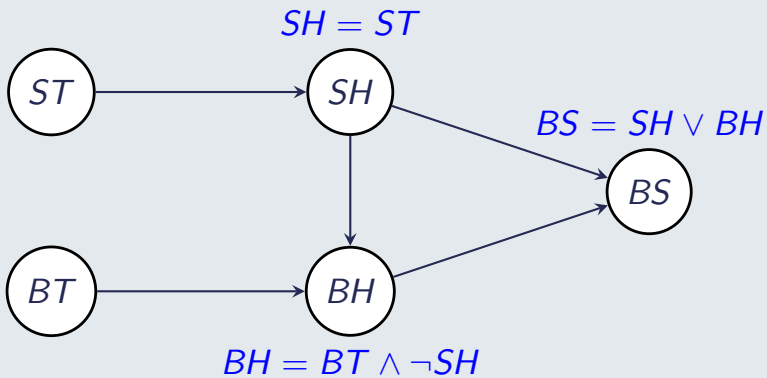


Figure – Suzy and Billy example.

Qu'est-ce que la causalité Réelle ?

Causalité réelle

Determiner les éléments qui ont fait qu'un événement réel spécifique soit arrivé (ou qu'un certain état du monde ait été atteint)

Example 2 : Pollution

Three different factories dump waste in the river. Each of them put 60L. The river become toxic when more than 100L are poured into it. What caused the river to become toxic ?

Qu'est-ce que la causalité Réelle ?

Causalité réelle

Determiner les éléments qui ont fait qu'un événement réel spécifique soit arrivé (ou qu'un certain état du monde ait été atteint)

Example 3 : Electrical circuit

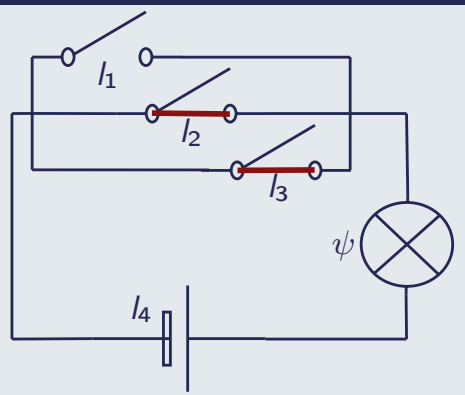


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

Modèle causal

Objet du modèle causal

Proposer une définition de la causalité réelle (actual causality) fondée sur le langage d'action du modèle d'action.

- Qu'est-ce que la causalité réelle ?
- Pourquoi est-ce important ?
- Quels sont les limitations de l'état de l'art ?

Why is Causality Important ?

Domains Where Causal Reasoning is Important

In legal reasoning for :

- Legal responsibility.

In ethical reasoning for :

- Reasoning about means and ends ;
- Determining direct/indirect consequences of an action.

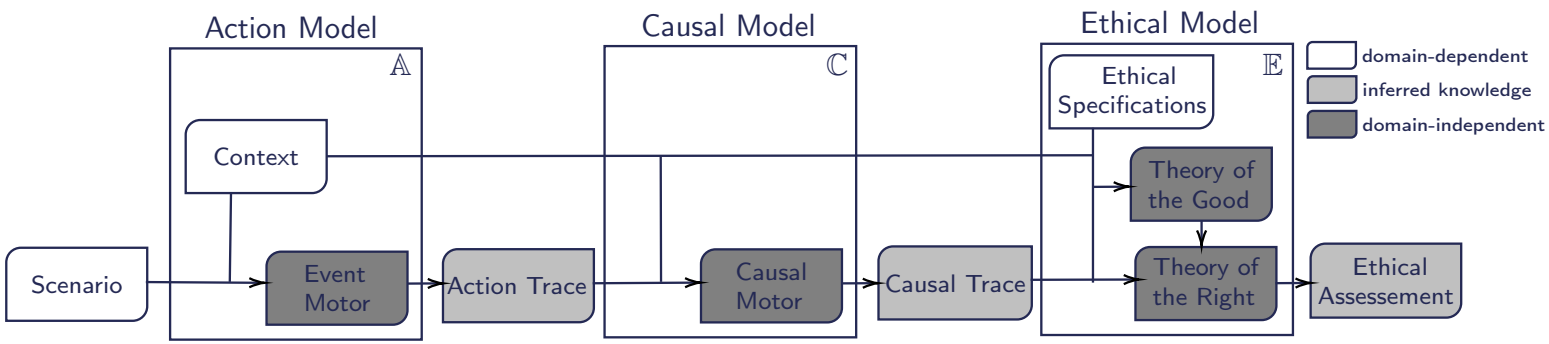


Figure – ACE modular framework for representing and applying ethical principles.[?]

. [1] Sarmiento, C., Bourgne, G., Cavalli, D., Inoue, K., Ganascia, J.G. : Action Languages Based Actual Causality in Ethical Decision Making Contexts. In arXiv[cs.AI] (2022).

Modèle causal

Objet du modèle causal

Proposer une définition de la causalité réelle (actual causality) fondée sur le langage d'action du modèle d'action.

- Qu'est-ce que la causalité réelle ?
- Pourquoi est-ce important ?
- Quels sont les limitations de l'état de l'art ?

Main Remaining Limitations of Recent Works

i) Conflating Causality and Responsibility [?]

A causal inquiry is necessary but not sufficient to determine responsibility. Causal inquiry needs to be **factual and independent of policy choices**.

ii) Unsatisfactory handling of the Cases of Overdetermination

Simple Counterfactual test fails to recognize causes in such cases.
HP definition solves the problem by forcing some variable to keep their original values ('miracle') which is difficult to justify semantically.

iii) Lack of expressivity of underlying language

No distinction between states and events, nor explicit time.

[Structural model's] limited expressivity render them less than ideal for some of the more delicate causal queries, like actual causation[?]

. [2] Wright, R.W. : Causation in Tort Law. California Law Review 73(6), 1735–1828 (1985).

. [3] Hopkins, M., Pearl, J. : Causality and Counterfactuals in the Situation Calculus. Journal of Logic and Computation 17(5), 939–953 (2007).

Main Remaining Limitations of Recent Works

i) Conflating Causality and Responsibility [?]

A causal inquiry is necessary but not sufficient to determine responsibility. Causal inquiry needs to be **factual and independent of policy choices**.

ii) Unsatisfactory handling of the Cases of Overdetermination

Simple Counterfactual test fails to recognize causes in such cases.
HP definition solves the problem by forcing some variable to keep their original values ('miracle') which is difficult to justify semantically.

iii) Lack of expressivity of underlying language

No distinction between states and events, nor explicit time.

[Structural model's] limited expressivity render them less than ideal for some of the more delicate causal queries, like actual causation[?]

. [2] Wright, R.W. : Causation in Tort Law. California Law Review 73(6), 1735–1828 (1985).

. [3] Hopkins, M., Pearl, J. : Causality and Counterfactuals in the Situation Calculus. Journal of Logic and Computation 17(5), 939–953 (2007).

Main Remaining Limitations of Recent Works

i) Conflating Causality and Responsibility [?]

A causal inquiry is necessary but not sufficient to determine responsibility. Causal inquiry needs to be **factual and independent of policy choices**.

ii) Unsatisfactory handling of the Cases of Overdetermination

Simple Counterfactual test fails to recognize causes in such cases.
HP definition solves the problem by forcing some variable to keep their original values ('miracle') which is difficult to justify semantically.

iii) Lack of expressivity of underlying language

No distinction between states and events, nor explicit time.

[Structural model's] limited expressivity render them less than ideal for some of the more delicate causal queries, like actual causation[?]

. [2] Wright, R.W. : Causation in Tort Law. California Law Review 73(6), 1735–1828 (1985).

. [3] Hopkins, M., Pearl, J. : Causality and Counterfactuals in the Situation Calculus. Journal of Logic and Computation 17(5), 939–953 (2007).

Why action languages ?

'These queries [actual causation] ***require a language*** *that is suited for dealing with* ***complex, dynamically changing situations****.' [?].*

. [3] Hopkins, M., Pearl, J. : Causality and Counterfactuals in the Situation Calculus. Journal of Logic and Computation 17(5), 939–953 (2007).

. [4] Batusov, V., Soutchanski, M. : Situation Calculus Semantics for Actual Causality. Thirty-Second AAAI Conference on Artificial Intelligence. (2018).

Why action languages ?

*'These queries [actual causation] **require a language** that is suited for dealing with **complex, dynamically changing situations**.' [?].*

*'Some of those examples [popular examples of actual causation formulated in philosophical literature] sound deceptively simple, but faithful modelling of them requires time, **concurrency** and **natural actions**.' [?].*

	Concurrency of Events	**Natural Actions**	**Non Deterministic or Durative Actions**
$\mathcal{A}$,$\mathcal{B}$,**PDDL**			
$\mathcal{A}_c$	X		
$\mathcal{C}$,**PDDL+**	X	X	X
Ours	X	X	

Table – Comparison of existing action languages.

. [3] Hopkins, M., Pearl, J. : Causality and Counterfactuals in the Situation Calculus. Journal of Logic and Computation 17(5), 939–953 (2007).

. [4] Batusov, V., Soutchanski, M. : Situation Calculus Semantics for Actual Causality. Thirty-Second AAAI Conference on Artificial Intelligence. (2018).

Basic Elements of the Action Language

Reasoning Objective of our Action Language

Represent and determine the evolution of the world given a set of actions corresponding to deliberate choices of the agent. This task is the simplest kind of temporal reasoning—temporal projection.

$\mathbb{F}$: Fluents (time-varying properties)

A set $L \subseteq Lit_{\mathbb{F}}$ is a **state** iff it is :

- Coherent : $\forall l \in L, \bar{l} \notin L$;
- Complete : $\forall f \in \mathbb{F}, f \in L$ or $\neg f \in L$.

Figure – Transition system.

$\mathbb{E}$: Events (describe transitions)

Each event is characterised by :

- $\mathcal{P} ::= l | \psi_1 \wedge \psi_2 | \psi_1 \vee \psi_2$
- $\mathcal{E} ::= [\psi] l | \varphi_1 \wedge \varphi_2$

$\mathbb{E}$ is divided into two sub-sets :

- $\mathbb{A}$ actions (volition) ;
- $\mathbb{U}$ exogenous events (agentless)

Basic Elements of the Action Language

Reasoning Objective of our Action Language

Represent and determine the evolution of the world given a set of actions corresponding to deliberate choices of the agent. This task is the simplest kind of temporal reasoning—temporal projection.

$\mathbb{F}$: Fluents (time-varying properties)

A set $L \subseteq Lit_{\mathbb{F}}$ is a **state** iff it is :

- Coherent : $\forall l \in L, \bar{l} \notin L$;
- Complete : $\forall f \in \mathbb{F}, f \in L$ or $\neg f \in L$.

Figure – Transition system.

$\mathbb{E}$: Events (describe transitions)

Each event is characterised by :

- $\mathcal{P} ::= l | \psi_1 \wedge \psi_2 | \psi_1 \vee \psi_2$
- $\mathcal{E} ::= [\psi] l | \varphi_1 \wedge \varphi_2$

$\mathbb{E}$ is divided into two sub-sets :

- $\mathbb{A}$ actions (volition) ;
- $\mathbb{U}$ exogenous events (agentless)

Basic Elements of the Action Language

Reasoning Objective of our Action Language

Represent and determine the evolution of the world given a set of actions corresponding to deliberate choices of the agent. This task is the simplest kind of temporal reasoning—temporal projection.

$\mathbb{F}$: Fluents (time-varying properties)

A set $L \subseteq Lit_{\mathbb{F}}$ is a **state** iff it is :

- Coherent : $\forall l \in L, \bar{l} \notin L$;
- Complete : $\forall f \in \mathbb{F}, f \in L$ or $\neg f \in L$.

Figure – Transition system.

$\mathbb{E}$: Events (describe transitions)

Each event is characterised by :

- $\mathcal{P} ::= l | \psi_1 \wedge \psi_2 | \psi_1 \vee \psi_2$
- $\mathcal{E} ::= [\psi] l | \varphi_1 \wedge \varphi_2$

$\mathbb{E}$ is divided into two sub-sets :

- $\mathbb{A}$ actions (volition) ;
- $\mathbb{U}$ exogenous events (agentless)

Temporal Projection

Keep Track of :

- States of the world resulting from event occurrence ;
- Occurrence of exogenous events [1].

1. An action can have more consequences than expected.

Simulate the Evolution of the World

Basic Causal and Directional Information

- $actualEff(E, L)$: actual effects of $E \in \mathbb{E}$ if occurring in the state L (or partial state).

$$\begin{aligned} actualEff(E, L) &= \bigcup_{e \in E} actualEff(\{e\}, L) \\ &= \bigcup_{e \in E} \{l_j, [\psi_j] l_j \in eff(e),\ \ L \vDash \psi_j,\ \text{and}\ l_j \notin L^2\} \end{aligned}$$

- $S_i \triangleright E$: resulting state when $E \in \mathbb{E}$ occurs at S_i.

$$S_i \setminus \overline{actualEff(E, S_i)} \cup actualEff(E, S_i)$$

Figure – Evolution of the world from S_0 with $S_{i+1} = S_i \triangleright E_i$.

2. An action can have less consequences than expected.

Example : Suzy and Billy

Event specification

- Action $\texttt{throws}(P)$ (P is either $\texttt{suzy}$ or $\texttt{billy}$)
  - No precondition (true)
  - Effect : $\texttt{rockThrown}(P)$
- Event $\texttt{hits}(P)$
  - Triggering cond : $\texttt{rockThrown}(P) \wedge \neg \texttt{bottleShattered}$
  - Effect : $\texttt{bottleShattered}$
  - Priority : $\texttt{hits(suzy)} > \texttt{hits(billy)}$

Example : Suzy and Billy

Event specification

- Action $\texttt{throws}(P)$ (P is either `suzy` or `billy`)
  - No precondition (true)
  - Effect : $\texttt{rockThrown}(P)$
- Event $\texttt{hits}(P)$
  - Triggering cond : $\texttt{rockThrown}(P) \wedge \neg \texttt{bottleShattered}$
  - Effect : `bottleShattered`
  - Priority : $\texttt{hits(suzy)} > \texttt{hits(billy)}$

Trace

- `throws(suzy)` and `throws(billy)` occur at time 0
- State becomes : {`rockThrown(billy)`,`rockThrown(suzy)`}
- `hits(suzy)` occurs first (as it overtakes `hits(billy)`)
- State becomes : {`rockThrown(billy)`,`rockThrown(suzy)`, `bottleShattered`}. `hits(billy)` can no longer occur.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

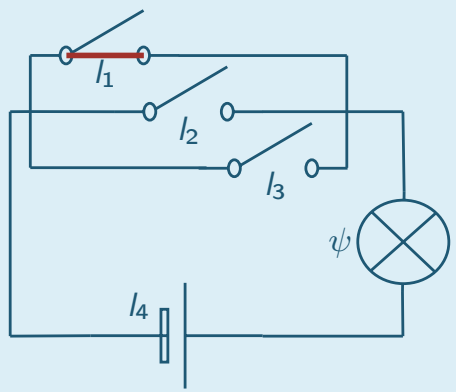


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

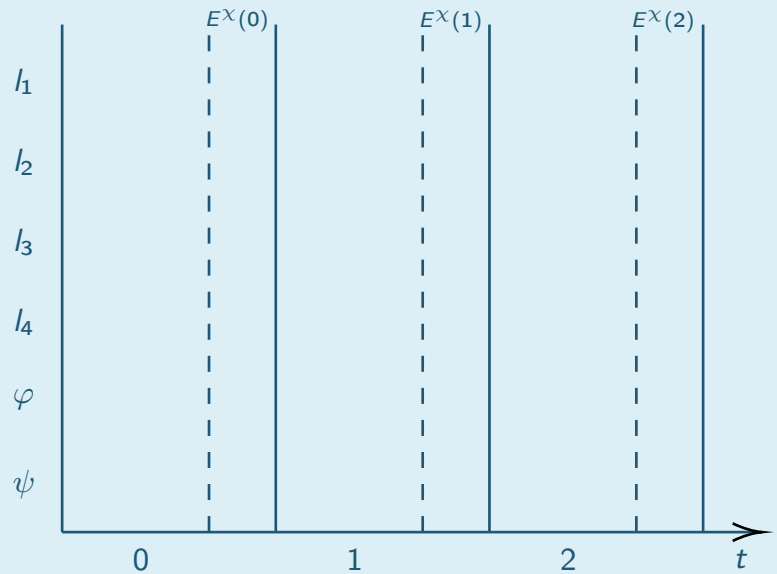


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

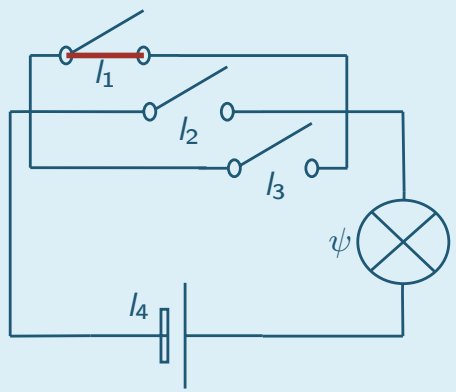


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

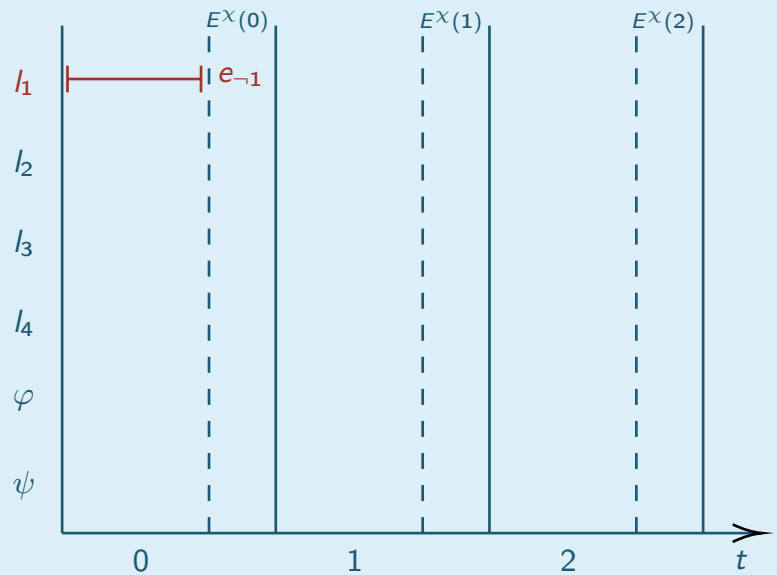


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

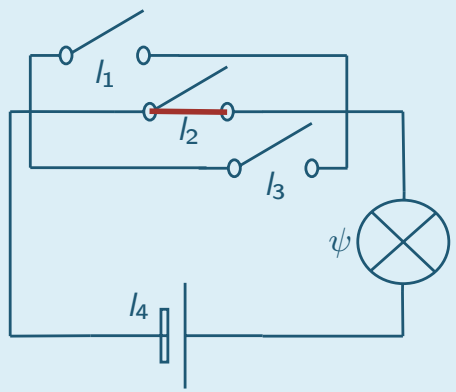


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

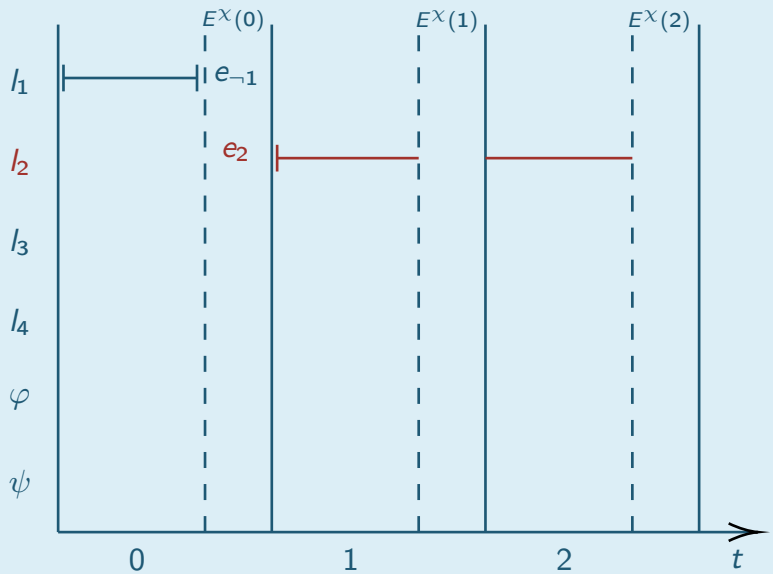


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = I_1 \vee I_2 \vee I_3$
$\psi = \varphi \wedge I_4$
$\quad = (I_1 \wedge I_4) \vee (I_2 \wedge I_4) \vee (I_3 \wedge I_4)$

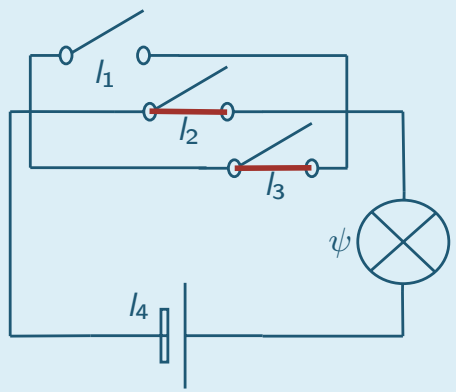


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

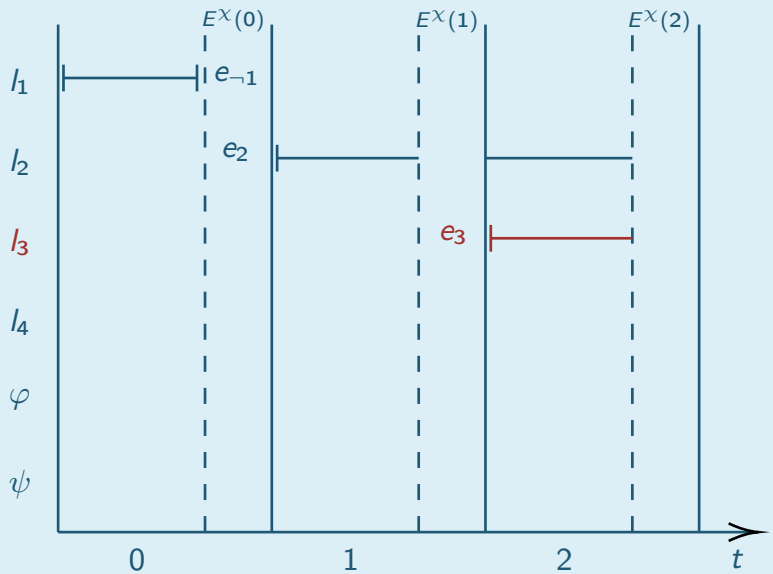


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

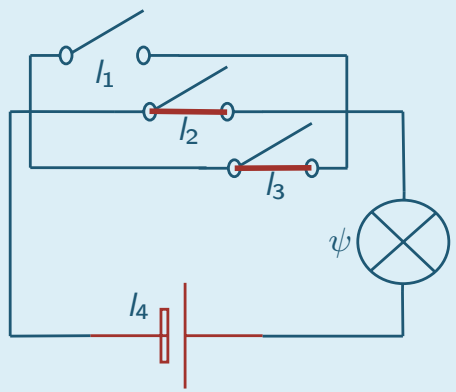


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

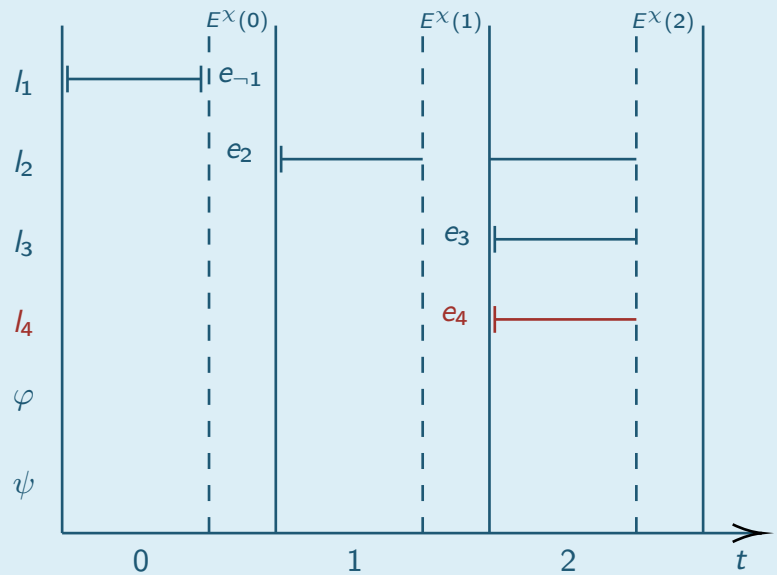


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

Causes of $(l_2, 2)$?

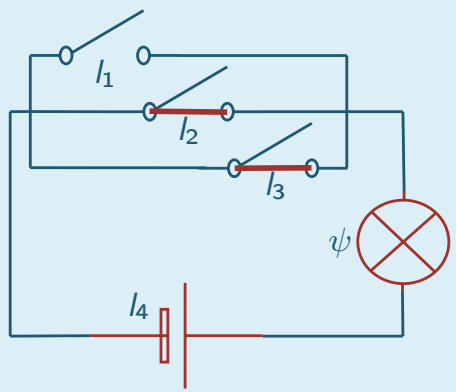


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

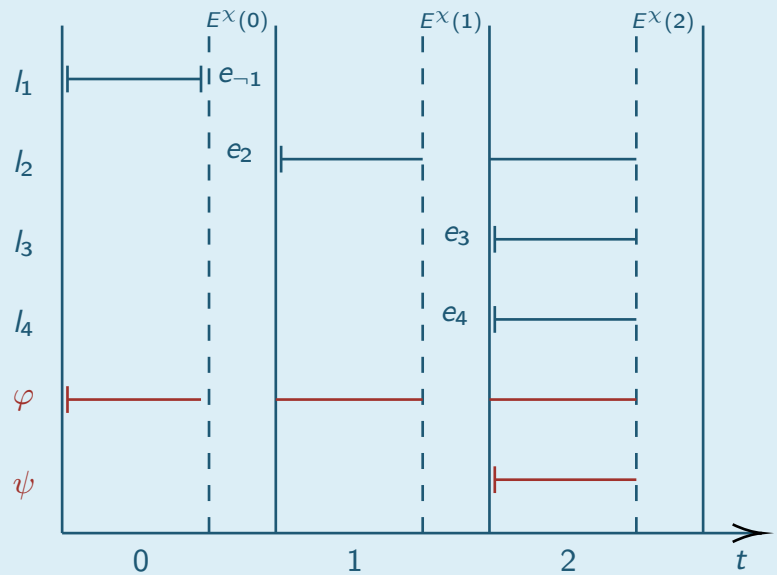


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

Causes of $(l_2, 2)$? $(e_2, 0)$

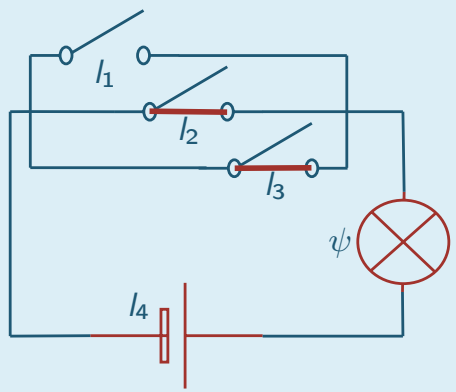


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

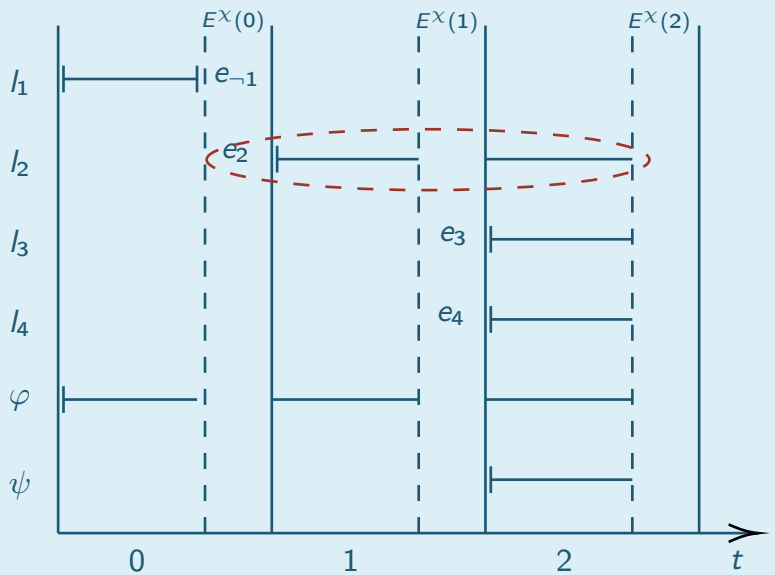


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

Causes of $(\varphi, 2)$?

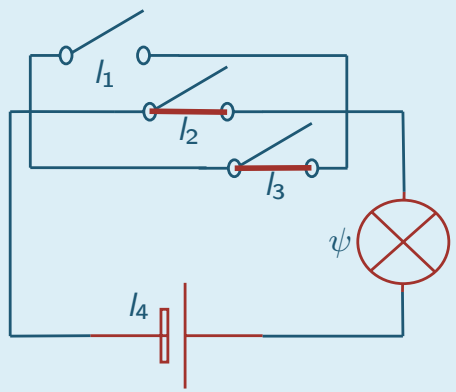


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

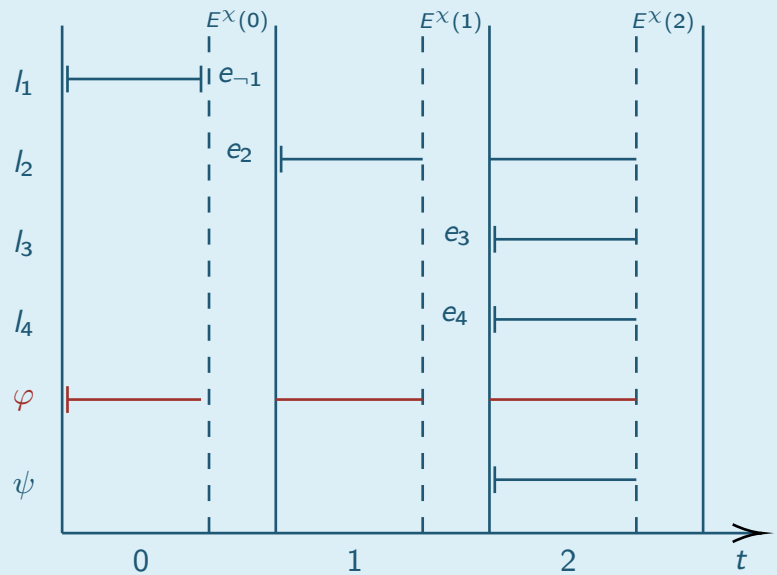


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

Causes of $(\varphi, 2)$? $(e_2, 0)$ ou $(e_3, 1)$

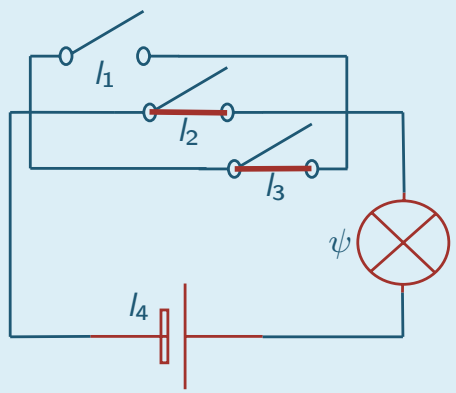


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

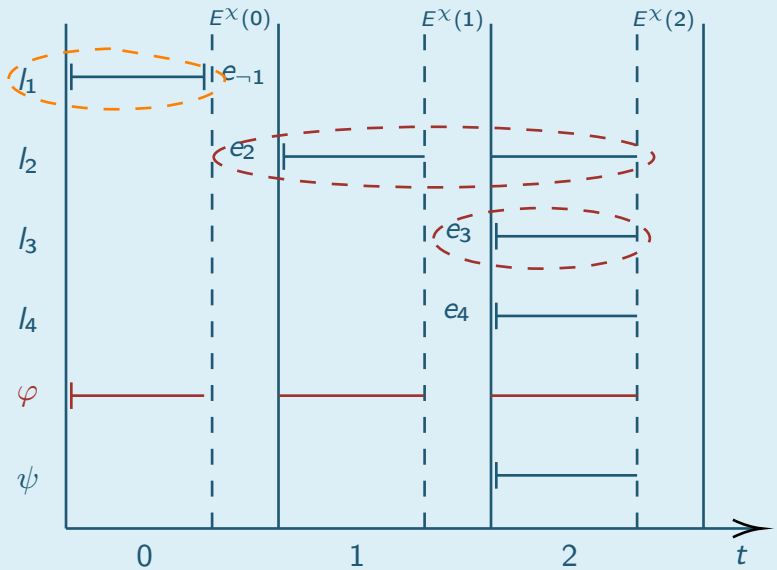


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

Causes of $(\psi, 2)$?

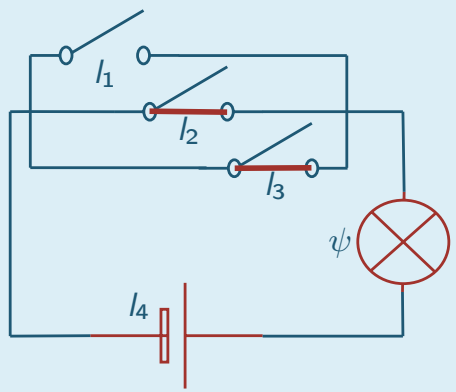


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

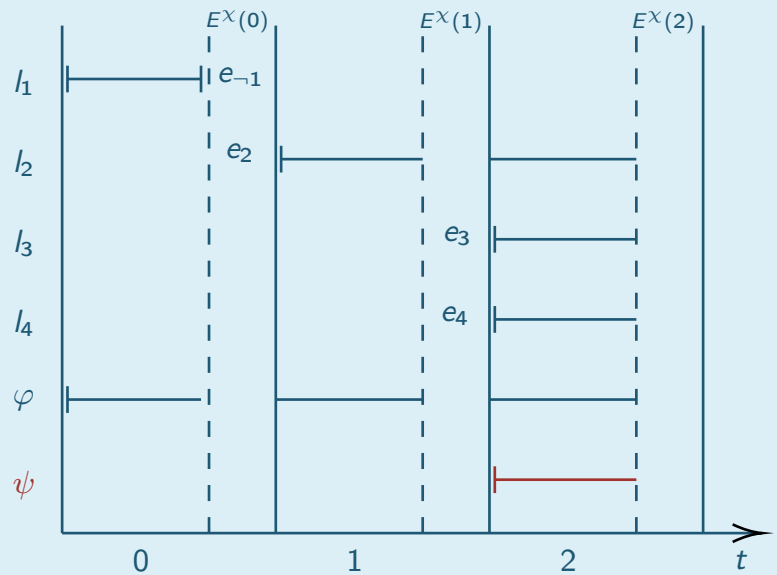


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Electrical Circuit Example

$\varphi = l_1 \vee l_2 \vee l_3$
$\psi = \varphi \wedge l_4$
$\quad = (l_1 \wedge l_4) \vee (l_2 \wedge l_4) \vee (l_3 \wedge l_4)$

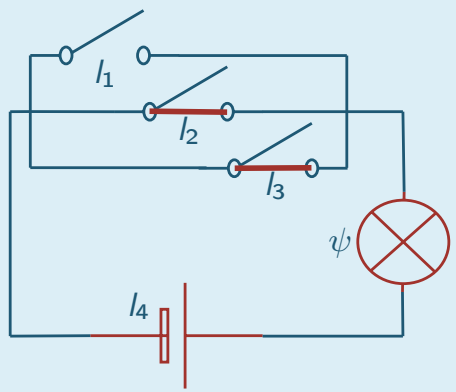


Figure – Electrical circuit consisting of a voltage source, three switches, and an individual connected to electrodes.

Causes of $(\psi, 2)$? $\{(e_2, 0), (e_4, 1)\}$ ou $\{(e_3, 1), (e_4, 1)\}$

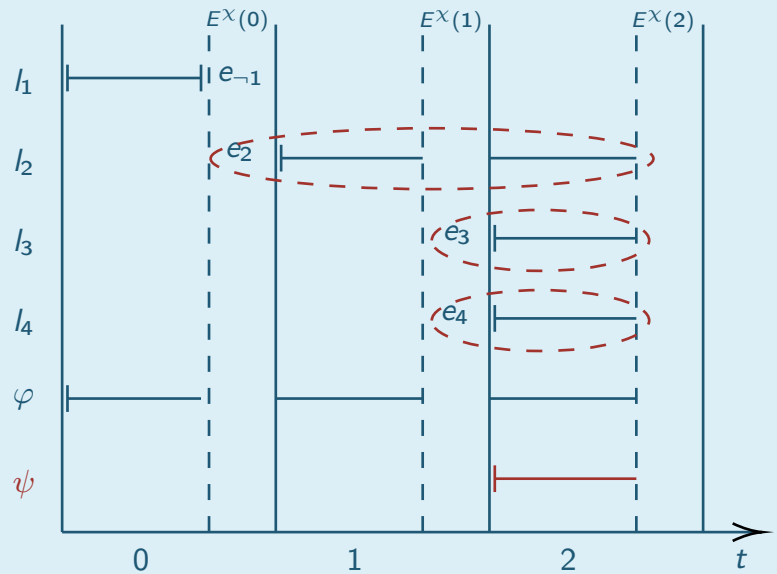


Figure – Evolution of electrical circuit elements state.

Direct NESS-Cause Intuition

Method

Given (ψ, t)

- Find a minimal backing W such that $W \models \psi$ and W true at time t
- Decompose W by considering since when each of its literal has been true, getting a sequence of $(W_i, t_i + 1)$
- For each step t_i determine the minimal set of events $C(t_i)$ happening at t_i that contributed by its effect to make part of W_i true
- $C = \bigcup C(t_i)$ is a sufficient set of direct-NESS causes of (ψ, t) (via backing W) : denoted $C \underset{W}{\rightsquigarrow} (\psi, t_\psi)$.
- Each $c \in C$ is a direct-NESS of (ψ, t) (via backing W)

Causal Setting χ

Definition

The action language causal setting denoted χ is the couple (σ, κ) with σ a scenario and κ a context [3].

Trace given a causal setting χ

- A state $S^{\chi}(t)$ is associated to each time point t of $\mathbb{T} = \{-1, 0, \ldots, N\}$.
- $E^{\chi}(t)$ is the set of all events which occur at a time point t.

3. Initial state (problem knowledge) and events specifications (domain knowledge).

Direct NESS-Cause

Definition[?]

Given a causal setting χ, the occurrence of events set $C = \{(e,t),\ e \in E^{\chi}(t),\ t \in \mathbb{T}\}$ is a sufficient set of direct NESS-causes of the truthfulness of the formula ψ at t_ψ, denoted $C \underset{W}{\rightsquigarrow} (\psi, t_\psi)$, iff there exists a partial state $W \subseteq Lit_{\mathbb{F}}$ that we call backing such that :

- Causal sufficiency and minimality of W : $W \vDash \psi$ and $\forall W' \subset W,\ W' \nvDash \psi$.

  There is a decreasing sequence $t_1, \ldots, t_k$ and a partition $W_1, \ldots, W_k$ of W such that $\forall i \in \{1, \ldots, k\}$, given $C(t_i) = C \cap E^{\chi}(t_i)$:

  - Weak necessity and minimality of C at t_i : $S^{\chi}(t_i) \triangleright C(t_i) \vDash W_i$ and $\forall C' \subset C(t_i),\ S^{\chi}(t_i) \triangleright C' \nvDash W_i$.
  - Persistency of necessity : $\forall t,\ t_i < t \leq t_\psi,\ S^{\chi}(t) \vDash W_i$.
- Minimality of C : $C = \bigcup_{i \in \{1,\ldots,k\}} C(t_i)$.

(e,t) is a direct NESS-cause of (ψ, t_ψ) iff $\exists C \subseteq \mathbb{E} \times \mathbb{T}$ such that $(e,t) \in C$, and $C \underset{W}{\rightsquigarrow} (\psi, t_\psi)$.

. [5] Sarmiento, C., Bourgne, G., Inoue, K., Ganascia, J.G. : Action Languages Based Actual Causality in Decision Making Contexts. In PRIMA (2022).

NESS-Cause

Figure – Example of an agent throwing a rock.

Definition

Establish a causal chain by going back in time and looking :

- at the events that caused exogenous events to be triggered ;
- at the events that caused events to have their actual effects.

Actual Cause

Definition

Given a causal setting χ and an event $e \in E^{\chi}(t_{\psi})$, the actual causes of (e, t_{ψ}) are the NESS-causes of $(tri(e), t_{\psi})$ [4].

4. $tri(e) \in \mathcal{P}$.

Actual Cause

Definition

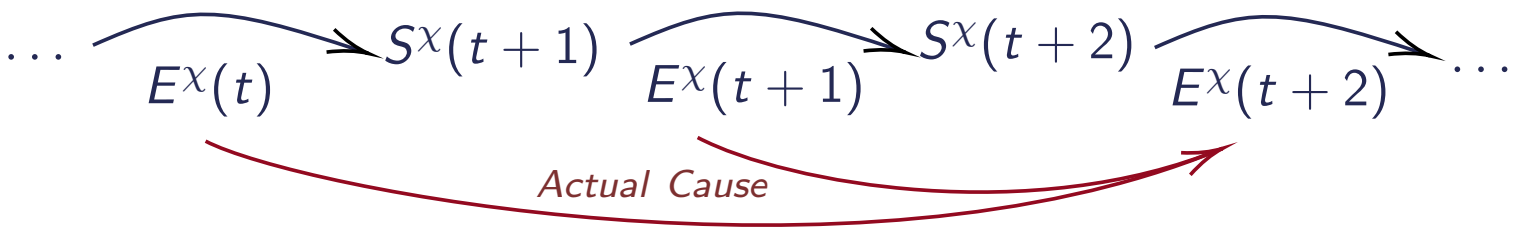
Given a causal setting χ and an event $e \in E^{\chi}(t_{\psi})$, the actual causes of (e, t_{ψ}) are the NESS-causes of $(tri(e), t_{\psi})$.

$$\cdots \xrightarrow{E^{\chi}(t)} S^{\chi}(t+1) \xrightarrow{E^{\chi}(t+1)} S^{\chi}(t+2) \xrightarrow{E^{\chi}(t+2)} \cdots$$

Actual Cause

Definition

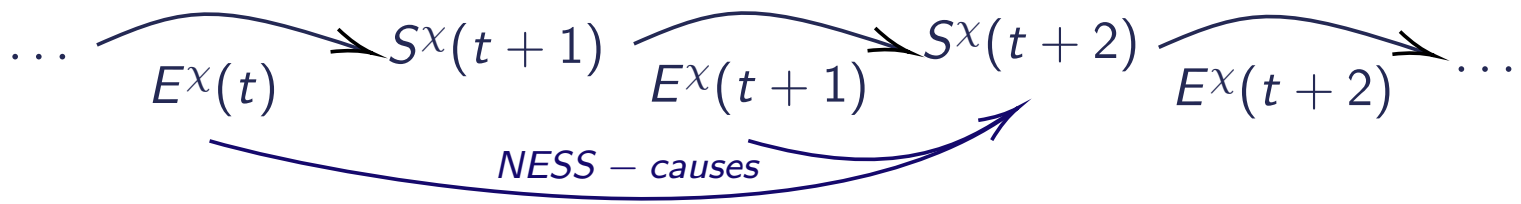
Given a causal setting χ and an event $e \in E^{\chi}(t_{\psi})$, the actual causes of (e, t_{ψ}) are the NESS-causes of $(tri(e), t_{\psi})$.



Actual Cause

Definition

Given a causal setting χ and an event $e \in E^{\chi}(t_{\psi})$, the actual causes of (e, t_{ψ}) are the NESS-causes of $(tri(e), t_{\psi})$.



Travaux futurs

Causalité négative

.

- Non-occurence d'évènement
  - comme conséquence : prevents
  - comme causes : notation d'omisson
- Evénements composés / complexes

Autres extensions

- Outil de comparaison : traductions depuis PDDL
- Théorie décisionnelle de la causalité
  - Relations causales 'enables' et 'excludes'
  - prise en compte des connaissances et de l'intentionalité : états épistémiques, manipulations....

ÉTHIQUE ET CAUSALITÉ

Conférence invitée *CNIA 2023*

4 juillet 2023

Gauvain Bourgne

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France