Human Reasoning, Computational Logic, and Ethical Decision Making

Dominic Deckert[1] Emmanuelle-Anna Dietz Saldanha[1]
Steffen Hölldobler[1,2] **Sibylle Schwarz**[3]

[1]International Center for Computational Logic, TU Dresden, Germany,
[2]North-Caucasus Federal University, Stavropol, Russian Federation,
[3]HWTK Leipzig, Germany

Poznań Reasoning Week
25. August 2019

Inspiration

Luís Moniz Pereira and Ari Saptawijaya [2016]:
Programming Machine Ethics

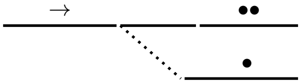
- Computational models of machine ethics
- Various ethical problems are implemented as logic programs
- Query for moral permissability
- However, the approach
  - does not provide a general method to account for ethical dilemmas
  - is not integrated into a cognitive theory about human reasoning

We do not aim at suggesting a moral theory!

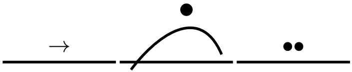
The attempt of implementing a machine ethics, will help us understand human ethics and address the ambiguities that have not been sorted out so far. (Wallach and Allen, 2008)

Trolley Problem (Foot [1967])

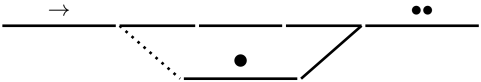
- Bystander Case



- Footbridge Case

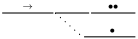
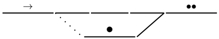


- Loop Case



Which action is morally permissable?

Ethical Decision Principles in Trolley Problems


	Bystander Case	Loop Case	Footbridge Case
Doctrine of double effect	change	-	-
Doctrine of triple effect	change	change	-
Maximize humans saved	change	change	throw down
action permissible say	85%	56%	12%

Maximize the number of humans saved (Utilitarism)

Could I save more humans by my action than humans that would be killed?

Doctrine of double effect: Killing is not permissible as *a means to save others*

If there were no human on the side track and I changed the switch
then I would still save humans on the main track?

Doctrine of triple effect: *Intentional and direct* kill is not permissible

Could I avoid to intentionally and directly kill someone
in order to save the others?

(Hauser, Cushman, Young, Kang-Xing Jin, Mikhail [2007]:
A Dissociation Between Moral Judgments and Justifications)

Ethical Decision Making

Basic assumption
Humans construct models and reason with respect to them

An integrated computational cognitive theory must be able to consider

- actions with direct and indirect effects
- ethical principles
- conditional reasoning
  If I change the switch then I will save the humans on the main track
- counterfactual or prefactual reasoning
  Is a killing a side effect?
  If there were no human on the side track and I changed the switch then I would still save the humans on the main track

This is ongoing work

Towards an Integrated Computational Cognitive Theory

- Stenning, van Lambalgen [2009]
  Human Reasoning and Cognitive Science
- Hölldobler, Kencana Ramli [2009]
  Logic Programs under Three-Valued Łukasiewicz's Semantics

Normal logic programs $\mathcal{P}$ are finite sets of

$$\begin{array}{rll} \text{Facts} & e \leftarrow \top & \\ \text{Rules} & s \leftarrow e \wedge \neg ab_1 & s \leftarrow t \wedge \neg ab_2 \\ \text{Assumptions} & ab_1 \leftarrow \bot & ab_2 \leftarrow \bot \end{array}$$

Weak completion $wc\mathcal{P}$ of program $\mathcal{P}$

$$\{e \leftrightarrow \top, s \leftrightarrow (e \wedge \neg ab_1) \vee (t \wedge \neg ab_2), ab_1 \leftrightarrow \bot, ab_2 \leftrightarrow \bot\}$$

Least models under three-valued Łukasiewicz logic

$$\langle \{e, s\}, \{ab_1, ab_2\} \rangle$$

Three-Valued Łukasiewicz Logic

truth values $\{0, 1/2, 1\}$ (syntactically represented by $\{\top, \mathsf{U}, \bot\}$)

negation $\neg x \mapsto 1 - x$

(weak) disjunction $x \vee y \mapsto \max(x, y)$

(weak) conjunction $x \wedge y \mapsto \min(x, y)$

implication $x \rightarrow y \mapsto \min(1, 1 - x + y)$

equivalence $x \leftrightarrow y \mapsto 1 - |x - y|$

$\rightarrow$	0	1/2	1
0	1	1	1
1/2	1/2	1	1
1	0	1/2	1

$\leftrightarrow$	0	1/2	1
0	1	1/2	0
1/2	1/2	1	1/2
1	0	1/2	1

truth ordering $0 <_t 1/2 <_t 1$ (total)

information ordering $1/2 <_i 0$ and $1/2 <_i 1$ (partial)

Weak Completion Semantics of logic programs (WCS)

(Hölldobler and Kencana Ramli [2009])

Semantic Operator $\Phi_{\mathcal{P}}(I) = \langle J^{\top}, J^{\perp} \rangle$ of ground program $\mathcal{P}$, where

$$\begin{array}{lcll} J^{\top} & = & \{A \mid & A \leftarrow Body \in \mathcal{P} \text{ and } I(Body) = \top\} \\ J^{\perp} & = & \{A \mid & A \leftarrow Body \in \mathcal{P} \text{ and} \\ & & & \text{for all } A \leftarrow Body \in \mathcal{P} \text{ we find } I(Body) = \perp\} \end{array}$$

Least model of weakly completed program $\mathcal{P}$ = least fixed point of $\Phi_{\mathcal{P}}$

$$\{e \leftarrow \top, s \leftarrow e \wedge \neg ab_1, s \leftarrow t \wedge \neg ab_2, ab_1 \leftarrow \perp, ab_2 \leftarrow \perp\}$$

$$\begin{array}{lclccc} & & & \top & & \perp \\ \hline I & = & \langle & \emptyset & , & \emptyset & \rangle \\ \Phi_{\mathcal{P}}(I) & = & \langle & \{e\} & , & \{ab_1, ab_2\} & \rangle \\ \Phi_{\mathcal{P}}(\Phi_{\mathcal{P}}(I)) & = & \langle & \{e, s\} & , & \{ab_1, ab_2\} & \rangle \end{array}$$

$\Phi_{\mathcal{P}}(\Phi_{\mathcal{P}}(I))$ is a fixed point of $\Phi_{\mathcal{P}}$

Resoning under Weak Completion Semantics

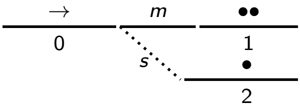
- Under WCS
  - represent a scenario as a logic program
  - compute the least model of the weak completion of the program
  - reason with respect to the least model
  - add skeptical abduction if necessary

- WCS is an integrated computational cognitive theory
  - suppression task
  - selection task
  - belief bias effect
  - syllogistic reasoning
  - spatial reasoning

How can we add actions and causality to WCS?

Fluent Calculus (Hölldobler and Schneeberger [1990])

- states are represented as multisets of fluents
- states are changed by the execution of actions
- actions are specified by its preconditions and direct effects
- actions might have indirect effects, which can be computed by ramifications



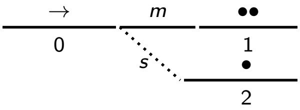
$$\dot{\{} t_0, c_0, m, h_1, h_1, h_2 \dot{\}} \xrightarrow{change} \dot{\{} t_0, c_0, s, h_1, h_1, h_2 \dot{\}}$$

Fluents t_0 t_1 t_2 m s c_0 h_1 h_2 d

Fluent terms t_0 $t_0 \circ c_0$ $t_0 \circ c_0 \circ 1$ 1 (unit)
where $\circ$ is an AC1-function symbol written infix

States multisets of ethically irrelevant / relevant fluents
$(t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)$

Actions



Agent

$$action(1, 1, donothing, 1, 1) \leftarrow \top$$
$$action(m, 1, change, s, 1) \leftarrow \top$$

Trolley

$$action(t_0 \circ c_0 \circ m, 1, downhill, t_1 \circ c_0 \circ m, 1) \leftarrow \top$$
$$action(t_0 \circ c_0 \circ s, 1, downhill, t_2 \circ c_0 \circ s, 1) \leftarrow \top$$

$$action(t_1 \circ h_1, 1, kill, t_1, d) \leftarrow \top$$
$$action(t_2 \circ h_2, 1, kill, t_2, d) \leftarrow \top$$

Causality

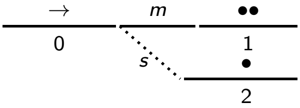
Original fluent calculus

- *plan*(X, P, Y) or *causes*(X, A, Y)
- the execution of plan P transforms state X into state Y
  - where a plan P is a sequence of actions
- *causes* can be defined recursively on plans

Problems:

- If a program $\mathcal{P}$ contains recursive structures like lists or natural numbers then $\Phi_{\mathcal{P}}$ is generally not continuous anymore
  Avoid recursive structures or restrict them to finite subsets
- There are infinitely many ground instances of *causes*(X, P, X)
  - Consider as base case only finite scenarios
  - Consider only the states obtained by executing the actions of the agent
  - Compute successor states as ramifications wrt the actions of the trolley

Weak Completion Semantics and Causality



Base cases

$$\begin{array}{ll} causes(\textcolor{red}{donothing}, t_0 \circ c_0 \circ \textcolor{red}{m} \circ h_1 \circ h_1 \circ h_2, 1) & \leftarrow \top \\ causes(\textcolor{red}{change}, t_0 \circ c_0 \circ \textcolor{red}{s} \circ h_1 \circ h_1 \circ h_2, 1) & \leftarrow \top \end{array}$$

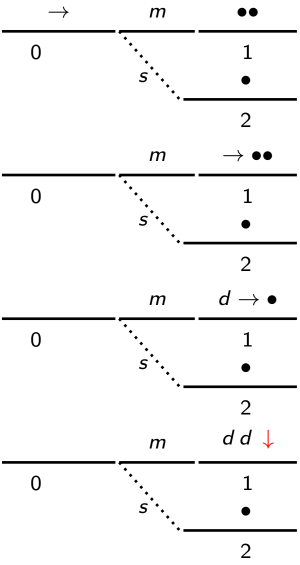
Recursive case

$$\begin{array}{rl} causes(\textcolor{red}{A}, E_1 \circ Z_1, E_2 \circ Z_2) \leftarrow & action(P_1, P_2, A', E_1, E_2) \\ & \wedge\ causes(\textcolor{red}{A}, P_1 \circ Z_1, P_2 \circ Z_2) \\ & \wedge\ \neg ab(A') \end{array}$$

Abnormalities

$$ab(downhill) \leftarrow \bot \qquad ab(kill) \leftarrow \bot$$

The Bystander Doing Nothing



$causes(donothing, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)$

$\Downarrow$ *downhill*

$causes(donothing, t_1 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)$

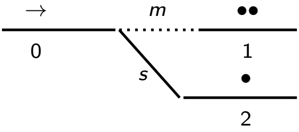
$\Downarrow$ *kill*

$causes(donothing, t_1 \circ c_0 \circ m \circ h_1 \circ h_2, d)$

$\Downarrow$ *kill*

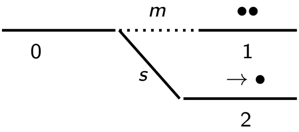
$causes(donothing, t_1 \circ c_0 \circ m \circ h_2, d \circ d)$

The Bystander Changing the Switch



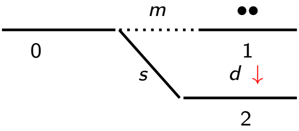
$causes(change, t_0 \circ c_0 \circ {\color{red}s} \circ h_1 \circ h_1 \circ h_2, 1)$

$\Downarrow$ *downhill*



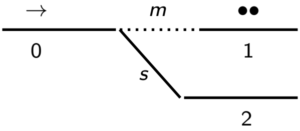
$causes(change, {\color{red}t_2} \circ c_0 \circ s \circ h_1 \circ h_1 \circ h_2, 1)$

$\Downarrow$ *kill*



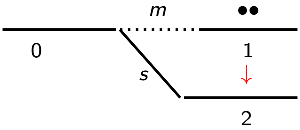
${\color{red}causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1, d)}$

The Bystander Changing Switch while Assuming Empty Side Track



$$causes(change, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1)$$

$\Downarrow$ *downhill*



$$causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1)$$

Equational Theories

(Jaffar, Lassez, Maher [1984]:
A Theory of Complete Logic Programs with Equality)

$\mathcal{P}$ a (ground) normal logic program not containing the equality symbol

$\mathcal{E}$ a set of equations

$\equiv_{\mathcal{E}}$ finest congruence relation on the set of ground terms defined by $\mathcal{E}$

$[t]$ congruence class defined by the ground term t

Herbrand $\mathcal{E}$-universe quotient of the set of ground terms modulo $\equiv_{\mathcal{E}}$

$[p(t_1, \ldots, t_n)]$ abbreviation for $p([t_1], \ldots, [t_n])$

$[p(t_1, \ldots, t_n)] = [q(s_1, \ldots, s_m)]$ iff $p = q$, $n = m$, and $[t_i] = [s_i]$ for all i

Theorem The weak completion of $\mathcal{P}$ has a least $\mathcal{E}$-model under the three-valued Łukasiewicz logic

Computing Least $\mathcal{E}$-Models

Semantic Operator $\quad \Phi_{\mathcal{E},\mathcal{P}}(I) = \langle J^\top, J^\bot \rangle$, where

$$\begin{array}{rcl} J^\top & = & \{[A] \mid A \leftarrow Body \in \mathcal{P} \text{ and } I(Body) = \top\} \\ J^\bot & = & \{[A] \mid A \leftarrow Body \in \mathcal{P} \text{ and for all } A' \text{ where} \\ & & \qquad A' \leftarrow Body \in \mathcal{P} \text{ with } [A] = [A'] \text{ we find } I(Body) = \bot\} \end{array}$$

Theorem $\quad \Phi_{\mathcal{E},\mathcal{P}}$ is monotonic.

It has a least fixed point. (by Knaster-Tarski Fixed Point Theorem)

Note that $\Phi_{\mathcal{E},\mathcal{P}}$ is not continuous in general.

$$q(1) \leftarrow \top \qquad q(a \circ X) \leftarrow q(X) \qquad r(1) \leftarrow \neg q(X)$$

Fixed point is reached after $\omega + 1$ step, where ω is the first limit ordinal.

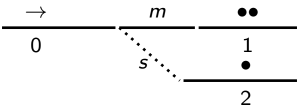
More results under the restriction to programs $\mathcal{P}$ that are

- propositional,
- finite ground,
- or finite datalog programs with finite Herbrand $\mathcal{E}$-universe.

Theorem $\quad \Phi_{\mathcal{E},\mathcal{P}}$ is continuous.

Theorem $\quad$ The least $\mathcal{E}$-model of the weak completion of $\mathcal{P}$ is the least fixed point of $\Phi_{\mathcal{E},\mathcal{P}}$ and vice versa.

Ethical Decision Making – The Bystander Case (1)



Background Knowledge $\mathcal{P}_B$

$$action(t_0 \circ c_0 \circ m, 1, downhill, t_1 \circ c_0 \circ m, 1) \leftarrow \top$$
$$action(t_0 \circ c_0 \circ s, 1, downhill, t_2 \circ c_0 \circ s, 1) \leftarrow \top$$

$$action(t_1 \circ h_1, 1, kill, t_1, d) \leftarrow \top$$
$$action(t_2 \circ h_2, 1, kill, t_2, d) \leftarrow \top$$

$$ab(downhill) \leftarrow \bot$$
$$ab(kill) \leftarrow \bot$$

$$\begin{aligned} causes(A, E_1 \circ Z_1, E_2 \circ Z_2) \leftarrow\ & action(P_1, P_2, A', E_1, E_2) \\ & \wedge\ causes(A, P_1 \circ Z_1, P_2 \circ Z_2) \\ & \wedge\ \neg ab(A') \end{aligned}$$

Ethical Decision Making – The Bystander Case (2)

If I do nothing then the humans on the main track will be killed. Yes

$$\mathcal{P}_B$$
$$causes(donothing, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1) \leftarrow \top$$

- Its least $\mathcal{E}$-model maps $causes(donothing, t_1 \circ c_0 \circ m \circ h_2, d \circ d)$ to $\top$

If I change the switch then the humans on the main track will be saved. Yes

If I change the switch then the human on the side track will be killed. Yes

$$\mathcal{P}_B$$
$$causes(change, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ h_2, 1) \leftarrow \top$$

- Its least $\mathcal{E}$-model maps $causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1, d)$ to $\top$

Ethical Decision Making – The Bystander Case (3)

Changing the switch is preferable to do nothing as it will kill fewer humans. *Yes*

$$\begin{array}{l}\mathcal{P}_B \\ causes(donothing, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1) \leftarrow \top \\ causes(change, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ h_2, 1) \leftarrow \top\end{array}$$

- Its least $\mathcal{E}$-model maps the following atoms to $\top$

$$\begin{array}{l}causes(donothing, t_1 \circ c_0 \circ m \circ h_2, d \circ d) \\ causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1, d)\end{array}$$

- Using

$$\begin{array}{lcl}prefer(A_1, A_2) & \leftarrow & causes(A_1, Z_1, D_1) \\ & & \wedge\ causes(A_2, Z_2, D_1 \circ d \circ D_2) \\ & & \wedge\ \neg ab_{prefer}(A_1) \\ ab_{prefer}(change) & \leftarrow & \bot \\ ab_{prefer}(donothing) & \leftarrow & \bot\end{array}$$

In the least model the following atoms are mapped to $\top$

$$\begin{array}{l}causes(donothing, t_1 \circ c_0 \circ m \circ h_2, d \circ d) \\ causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1, d)\end{array}$$

The number of humans killed is minimized by changing the switch.

Utilitarianism

Ethical Decision Making – The Bystander Case (4)

- *If there were no human on the side track and I changed the switch then I would still save the humans on the main track.* *Yes*

$$\mathcal{P}_B$$
$$causes(change, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1) \leftarrow \top$$

  - Its least $\mathcal{E}$-model maps $causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1)$ to $\top$
- Using

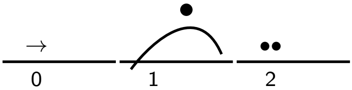
$$\begin{aligned} permissible(change) \leftarrow\ & prefer(change, donothing) \\ & \wedge\ causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1) \\ & \wedge\ \neg ab_{permissible}(change) \\ ab_{permissible}(change) \leftarrow\ & \bot \end{aligned}$$

allows to conclude that changing the switch is permissible

Doctrine of Double Effect

(Killing is permissible as a side effect but not as a means to save others)

Ethical Decision Making – The Footbridge Case



- Base cases

$$causes(donothing, t_0 \circ c_0 \circ c_1 \circ b_1 \circ h_2 \circ h_2, 1) \leftarrow \top$$
$$causes(throw, t_0 \circ c_0 \circ h_2 \circ h_2, d) \leftarrow \top$$

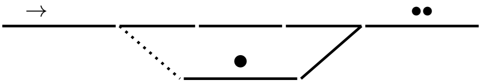
- *Is throwing the person from the bridge preferable to do nothing?* *No*

$$\begin{aligned} prefer(A_1, A_2) \leftarrow\ & causes(A_1, Z_1, D_1) \\ & \wedge\ causes(A_2, Z_2, D_1 \circ d \circ D_2) \\ & \wedge\ \neg ab_{prefer}(A_1) \\ ab_{prefer}(throw) \leftarrow\ & intentional\_direct\_kill(throw) \\ intentional\_direct\_kill(throw) \leftarrow\ & \top \end{aligned}$$

Pushing the person from the bridge is not permissible by

Doctrine of Double Effect

Ethical Decision Making – The Loop Case

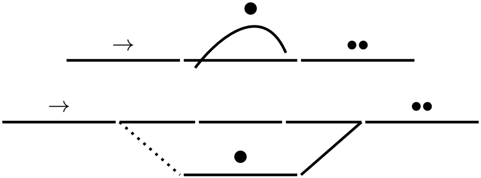


- *If I do nothing then the humans on the main track will be killed.* Yes
- *If I change the switch then the humans on the main track will be saved.* Yes
  If I change the switch then the human on the side track will be killed. Yes
- *If there were no human on the side track and I changed the switch then I would still save the humans on the main track.* No

Changing the switch is not permissible by

Doctrine of Double Effect

Ethical Decision Making: Loop versus Footbridge Case



- Humans seem to distinguish the cases
- Throwing the person from the bridge is not permissible
- However, changing the switch is acceptable
- Direct versus indirect intentional kill

Could I avoid to intentionally and directly kill someone to save others?

Doctrine of Triple Effect

(Intentional and direct kill is not permissible.)

Conclusion

- This is ongoing work
- We can solve all examples discussed in (Pereira, Saptawijaya 2017) uniformly in WCS with equality
- We are aiming at more general ethical rules
  - *If an action does something good and nothing abnormal is known then it is permissable.*
  - *A direct intentional kill is an abnormality.*
- Extension of WCS to more than three-valued Łukasiewicz logic