EPiC Series in Computing

Volume 57, 2018, Pages 326–342

LPAR-22. 22nd International Conference on Logic for Programming, Artificial Intelligence and Reasoning

The Weak Completion Semantics and Equality

Emmanuelle-Anna Dietz Saldanha[1], Steffen Hölldobler[2], Sibylle Schwarz[3], and Lim Yohanes Stefanus[4*]

[1] International Center for Computational Logic, TU Dresden, Germany
emmanuelle.dietz@tu-dresden.de

[2] International Center for Computational Logic, TU Dresden, Germany
and North-Caucasus Federal University, Stavropol, Russian Federation
sh@iccl.tu-dresden.de

[3] HTWK Leipzig, Germany
sibylle.schwarz@htwk-leipzig.de

[4] University of Indonesia, Depok, Indonesia
yohanes@cs.ui.ac.id

Abstract

The weak completion semantics is an integrated and computational cognitive theory which is based on normal logic programs, three-valued Łukasiewicz logic, weak completion, and skeptical abduction. It has been successfully applied – among others – to the suppression task, the selection task, and to human syllogistic reasoning. In order to solve ethical decision problems like – for example – trolley problems, we need to extend the weak completion semantics to deal with actions and causality. To this end we consider normal logic programs and a set $\mathcal{E}$ of equations as in the fluent calculus. We formally show that normal logic programs with equality admit a least $\mathcal{E}$-model under the weak completion semantics and that this $\mathcal{E}$-model can be computed as the least fixed point of an associated semantic operator. We show that the operator is not continuous in general, but is continuous if the logic program is a propositional, a finite-ground, or a finite datalog program and the Herbrand $\mathcal{E}$-universe is finite. Finally, we show that the weak completion semantics with equality can solve a variety of ethical decision problems like the bystander case, the footbridge case, and the loop case by computing the least $\mathcal{E}$-model and reasoning with respect to this $\mathcal{E}$-model. The reasoning process involves counterfactuals which is necessary to model the different ethical dilemmas.

1 Introduction

The weak completion semantics (WCS) is a novel cognitive theory. Its original idea is based on the work of Stenning and van Lambalgen [28] who proposed to model human reasoning tasks by, firstly, reasoning towards a normal logic program to represent the reasoning task and, secondly, by reasoning with respect to the least model of the normal logic program. Unfortunately, Stenning and van Lambalgen's approach contained a technical bug which was corrected in [14].

[*] The authors are mentioned in alphabetical order.

The WCS is based on many techniques and methods from logic programming and computational logic. However, these techniques and methods are usually tweaked a little bit in order to model human reasoning tasks adequately. For example, programs are not completed in the sense of Clark [5], but only weakly completed. Instead of the semantic operator introduced by Fitting [9], a modified operator introduced in [28] is used. Instead of the the three-valued Kripke-Kleene logic used in [9], the three-valued Łukasiewicz logic [21] is used. Because of the latter, each normal logic program admits a least model and reasoning is performed with respect to this model (see [14]).

The approach has been applied to various human reasoning tasks like the suppression task [4,7], the selection task [8,31], and human syllogistic reasoning [19,24]. In fact, WCS performed better on the human syllogistic reasoning tasks than all 12 cognitive theories discussed in [19]. As all human reasoning tasks are solved within one framework, the WCS is an integrated and computational cognitive theory. We are unaware of any other theory of this kind and with such a wide variety of applications.

Recently, ethical decision making has received much attention as autonomous agents become part of our daily life. In particular, we were inspired by the work of Pereira and Saptawijaya [25], who studied computational models of machine ethics. Various ethical problems are implemented as logic programs and these programs can be queried for moral permissibility. Unfortunately, their approach does not provide a general method to account for ethical dilemmas and is not integrated into a cognitive theory about human reasoning.

The problems studied in [25] were trolley problems like the bystander, the footbridge, and the loop case. In each case, actions with direct and indirect effects must be considered. Hence, in order to model and reason about these problems within the WCS, the WCS must be extended to deal with these actions. There are several candidates for modeling actions and causality like the situation calculus [22,26], the event calculus [20], or the fluent calculus [15]. We opted for the fluent calculus because in this calculus the fluents are resources which can be consumed and produced. This property is shared with linear logic, the linear connection method [10] and Petri networks [13], the latter of which have also been used in computational models for human reasoning [2].

In the fluent calculus [15] states are represented as multisets of fluent. Multisets are represented with the help of a binary function symbol $\circ$ written infix and a constant 1 such that $(X \circ Y) \circ Z \approx X \circ (Y \circ Z)$ (associativity), $X \circ Y \approx Y \circ X$ (commutativity), and $X \circ 1 \approx X$ (unit element) hold, where all variables are assumed to be universally quantified and $\approx$ denotes equality. In other words, $\circ$ is an AC1-function symbol.

In order to deal with such function symbols in the WCS, we need to extend WCS to handle equality, and this is the main task tackled in this paper. In very much the same way as Jaffar, Lassez, and Maher extended definite logic programs to deal with equality [16], we will extend the WCS to deal with equality. In particular, we will prove that normal logic programs with equality admit a least model under Łukasiwiecz logic in Section 3 and that this model can be computed as least fixed point of an appropriate semantic operator in Section 4. Finally, we will show in Section 5 how the bystander, the footbridge, and the loop case can be modeled under the extended WCS. Of particular interest in these models is that in order to solve ethical dilemmas we need to reason about counterfactuals [23]. We assume the reader to be familiar with logic programming but discuss some basics in Section 2. A discussion and an outlook will complete the paper in Section 6.

2 Basics

A set $\mathcal{E}$ of equations together with the axioms of equality defines a finest congruence relation $\equiv$ on the set of ground terms (see e.g. [12]). In the fluent calculus, $\equiv=\equiv_{AC1}$. Let t be a ground term. $[t]$ denotes the congruence class defined by $\equiv$ and containing t. We abbreviate $p([t_1], \ldots, [t_n])$ by $[p(t_1, \ldots, t_n)]$. Furthermore, $[p(t_1, \ldots, t_n)] = [q(s_1, \ldots, s_m)]$ if and only if $p = q$, $n = m$, and $[t_i] = [s_i]$ for all $1 \leq i \leq n$. In the fluent calculus, $[d \circ t_2] = [t_2 \circ d]$, $[d \circ t_1 \circ d] = [t_1 \circ d \circ d \circ 1]$ and $[p(d \circ t_2, d \circ t_1 \circ d)] = [p(t_2 \circ d, t_1 \circ d \circ d \circ 1)]$, where d, t_1, and t_2 are constants and p is a binary relation symbol.

The *Herbrand $\mathcal{E}$-universe* is the quotient of the set of ground terms modulo $\equiv$. The *Herbrand $\mathcal{E}$-base* is the set of all expressions of the form $[p(t_1, \ldots, t_n)]$, where p is an n-ary function symbol and $[t_i]$, $1 \leq i \leq n$, are elements of the Herbrand $\mathcal{E}$-universe.

We will consider three-valued interpretations over the Herbrand $\mathcal{E}$-universe. Such an *$\mathcal{E}$-interpretation* I is represented by $\langle I^\top, I^\bot \rangle$, where $I^\top$ and $I^\bot$ are disjoint subsets of the Herbrand $\mathcal{E}$-base such that $[A] \in I^\top$ iff $I(A) = \top$, $[A] \in I^\bot$ iff $I(A) = \bot$, and $[A] \notin I^\top \cup I^\bot$ iff $I(A) = \mathrm{U}$, where A is a ground atom and $\top$, $\bot$, and U mean true, false, and unknown, respectively. The truth ordering on $\{\bot, \mathrm{U}, \top\}$ is defined by $\bot <_t \mathrm{U} <_t \top$. Complex formulas are interpreted as usual under Łukasiewicz logic [21]. An $\mathcal{E}$-interpretation I is an *$\mathcal{E}$-model* for a formula F, in symbols $I \models F$, iff $I(F) = \top$.

A *(normal logic) program* $\mathcal{P}$ is a finite set of clauses of the form $A \leftarrow \mathit{Body}$, where the *head* A is an atom different from $\approx$ and *Body* is either a non-empty conjunction of literals, $\top$ (denoting truth), or $\bot$ (denoting falsehood). Clauses of the form $A \leftarrow \top$ and $A \leftarrow \bot$ are called *(positive) facts* and *(negative) assumptions*, respectively. An atom A is *defined* in $\mathcal{P}$ iff $\mathcal{P}$ contains a clause of the form $A \leftarrow \mathit{Body}$; otherwise A is said to be *undefined*. The set of all atoms that are defined in $\mathcal{P}$ is denoted by $\mathit{def}(\mathcal{P})$. The set of all ground instances of all clauses occurring in $\mathcal{P}$ is denoted by $\mathtt{g}\mathcal{P}$.

Consider the following transformation for a given ground program $\mathcal{P}$: (1) For all $A \in \mathit{def}(\mathcal{P})$, replace all clauses of the form $A \leftarrow \mathit{Body}_1, \ \ldots, \ A \leftarrow \mathit{Body}_n$ occurring in $\mathcal{P}$ by $A \leftarrow \mathit{Body}_1 \vee \ldots \vee \mathit{Body}_n$. (2) Replace all occurrences of $\leftarrow$ by $\leftrightarrow$. The resulting set of equivalences is called the *weak completion* of $\mathcal{P}$, denoted by $\mathit{wc}\mathcal{P}$ [14]. One should observe that the weak completion differs from the completion defined by Clark [5] in that undefined atoms are not identified with falsehood.

3 Weakly Completed Programs and Equality

In [14,18] it was shown that logic programs as well as their weak completions admit a least model under Łukasiewicz logic. We are going to extend this result for the weak completion semantics with equality (WCSE). Throughout this section we consider a given equational theory $\mathcal{E}$ with finest congruence relation $\equiv$ and a given logic program $\mathcal{P}$.

Proposition 1. *Let I be an $\mathcal{E}$-interpretation. If $I = \langle I^\top, I^\bot \rangle \models \mathcal{P}$ then $I' = \langle I^\top, \emptyset \rangle \models \mathcal{P}$.*

Proof. Let $\mathcal{P}$ be a logic program and $I = \langle I^\top, I^\bot \rangle \models \mathcal{P}$, i.e. for all rules $A \leftarrow \mathit{Body} \in \mathtt{g}\mathcal{P}$: $I(A \leftarrow \mathit{Body}) = \top$. By definition of the Łukasiewicz implication [21], we have $I(A \leftarrow \mathit{Body}) = \top$ iff $I(A) \geq_t I(\mathit{Body})$ with respect to the truth ordering $\bot <_t \mathrm{U} <_t \top$. We consider all possible cases for $I(A)$ and show $I' \models A \leftarrow \mathit{Body}$ by $I'(A) \geq_t I'(\mathit{Body})$:

1. $I(A) = \top = I'(A)$: $I'(A) \geq_t I'(\mathit{Body})$ holds for any *Body* because $I'(\mathit{Body}) \in \{\bot, \mathrm{U}, \top\}$ and $\top$ is the truth-maximal element in $\{\bot, \mathrm{U}, \top\}$.

2. $I(A) = \mathrm{U} = I'(A)$: From $I \models A \leftarrow \mathit{Body}$, i.e. $I(A \leftarrow \mathit{Body}) = \top$, we learn $I(\mathit{Body}) \leq_t I(A) = \mathrm{U}$ and, therefore, $I(\mathit{Body}) \in \{\bot, \mathrm{U}\}$. Hence, there is a conjunct L in *Body* such that $I(L) = \min_t\{I(L) \mid L \text{ is a conjunct in } \mathit{Body}\} = I(\mathit{Body}) \in \{\bot, \mathrm{U}\}$, where $\min_t$ denotes the minimal element with respect to the truth ordering $\leq_t$. We consider two cases for the literal L:

   (a) $L = B$ for an atom B: If $I(L) = I(B) \in \{\bot, \mathrm{U}\}$, we have $I'(\mathit{Body}) = I'(L) = I'(B) = \mathrm{U} = I'(A)$ by the definition of $I'^{\bot}$, i.e. $I'^{\bot} = \emptyset$.

   (b) $L = \neg B$ for an atom B: If $I(L) = I(\neg B) \in \{\bot, \mathrm{U}\}$ then $I(B) \in \{\mathrm{U}, \top\}$ and $I'(B) \in \{\mathrm{U}, \top\}$. Hence $I'(\neg B) \in \{\bot, \mathrm{U}\}$.

   In both cases, we have $I'(A \leftarrow \mathit{Body}) = \top$ by $\mathrm{U} = I'(A) \geq_t I'(\mathit{Body})$.

3. $I(A) = \bot$: Then, $I'(A) = \mathrm{U}$ by the definition $I'^{\bot} = \emptyset$. From $I \models A \leftarrow \mathit{Body}$, i.e. $I(A \leftarrow \mathit{Body}) = \top$, we know $I(\mathit{Body}) = \min_t\{I(L) \mid L \text{ is a conjunct in } \mathit{Body}\} = \bot$. Hence, there is a conjunct L in *Body* such that $I(L) = \min_t\{I(L) \mid L \text{ is a conjunct in } \mathit{Body}\} = I(\mathit{Body}) = \bot$. We consider two cases for the literal L:

   (a) $L = B$ for an atom B: By $I(L) = I(B) = \bot$, we have $I'(\mathit{Body}) = I'(L) = \mathrm{U} \leq_t I'(A)$.

   (b) $L = \neg B$ for an atom B: By $I(L) = I(\neg B) = \bot$, we have $I(B) = \top = I'(B)$ and $I'(\mathit{Body}) = I'(\neg B) = \bot \leq_t I'(A)$.

   In both cases, we have $I'(A \leftarrow \mathit{Body}) = \top$ by $\mathrm{U} = I'(A) \geq_t I'(\mathit{Body})$. □

Proposition 2. *If $\langle I_1^{\top}, \emptyset\rangle \models \mathcal{P}$ and $\langle I_2^{\top}, \emptyset\rangle \models \mathcal{P}$, then $\langle I_1^{\top} \cap I_2^{\top}, \emptyset\rangle \models \mathcal{P}$.*

Proof. Let $\mathcal{P}$ be logic program, $\langle I_1^{\top}, \emptyset\rangle \models \mathcal{P}$ and $\langle I_2^{\top}, \emptyset\rangle \models \mathcal{P}$, i.e. for all rules $A \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P}$ we have $I_1(A) \geq_t I_1(\mathit{Body})$ and $I_2(A) \geq_t I_2(\mathit{Body})$. Let $I = \langle I_1^{\top} \cap I_2^{\top}, \emptyset\rangle$. By $I^{\bot} = \emptyset$ we have $I(A) = \min_t(I_1(A), I_2(A)) \in \{\mathrm{U}, \top\}$ for all ground atoms A.

We show that $I \models A \leftarrow \mathit{Body}$ for every rule $A \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P}$. We consider all possible cases for $I(A)$ and show $I(A \leftarrow \mathit{Body}) = \top$ by $I(A) \geq_t I(\mathit{Body})$:

1. $I(A) = \top$: For any *Body* we have $I(A) \geq_t I(\mathit{Body})$ because $I(A) = \top$ is the truth-maximal element in $\{\bot, \mathrm{U}, \top\}$.

2. For $I(A) = \min_t(I_1(A), I_2(A)) = \mathrm{U}$, we show that $I(\mathit{Body}) \leq_t \mathrm{U}$. We have two cases:

   (a) If $I_1(A)) = \min_t(I_1(A), I_2(A)) = \mathrm{U}$, we have $I_1(\mathit{Body}) \leq_t \mathrm{U}$ by $I_1(A \leftarrow \mathit{Body}) = \top$ and, therefore, $I(\mathit{Body}) = \min_t(I_1(\mathit{Body}), I_2(\mathit{Body})) \leq_t \mathrm{U}$.

   (b) If $I_2(A) = \min_t(I_1(A), I_2(A)) = \mathrm{U}$, we have $I_2(\mathit{Body}) \leq_t \mathrm{U}$ by $I_2(A \leftarrow \mathit{Body}) = \top$ and, therefore, $I(\mathit{Body}) = \min_t(I_1(\mathit{Body}), I_2(\mathit{Body})) \leq_t \mathrm{U}$.

   In both cases, we have $I(\mathit{Body}) = \min_t(I_1(\mathit{Body}), I_2(\mathit{Body})) \leq_t \mathrm{U} = I(A)$ and, therefore, $I(A \leftarrow \mathit{Body}) = \top$.

3. $I(A) = \bot$ is impossible because $I^{\bot} = \emptyset$.

Because the interpretation $I = \langle I_1^{\top} \cap I_2^{\top}, \emptyset\rangle$ is a model for each rule in $\mathsf{g}\mathcal{P}$, we have $I \models \mathcal{P}$. □

Theorem 1. *The $\mathcal{E}$-model intersection property holds for $\mathcal{P}$, i.e. $\cap\{I \mid I \models \mathcal{P}\} \models \mathcal{P}$.*

Proof. The claim follows immediately from Propositions 1 and 2. □

Theorem 2. *The $\mathcal{E}$-model intersection property holds for $wc\mathcal{P}$ as well.*

We will prove this result in Section 4.

4 Computing Least Models under Equality

In [14, 18] it was shown that the semantic operator defined by Stenning and van Lambalgen in [28] computes the least model of the weak completion of a program. We are going to extend this result for the WCSE.

Let $\mathcal{P}$ be a logic program and I an interpretation. We define $\Phi^{\mathcal{E}}_{\mathcal{P}}(I) = \langle J^\top, J^\bot\rangle$, where

$$\begin{array}{rcl} J^\top & = & \{[A] \mid \text{there exists } A \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P} \text{ and } I(\mathit{Body}) = \top\}, \\ J^\bot & = & \{[A] \mid \text{there exists } A \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P} \\ & & \quad \text{and for all } A' \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P} \text{ with } [A] = [A'] \text{ we find } I(\mathit{Body}) = \bot\}. \end{array}$$

Let X be a set of (three-valued) $\mathcal{E}$-interpretations. Let $X^\top = \{I^\top \mid \langle I^\top, I^\bot\rangle \in X\}$ and $X^\bot = \{I^\bot \mid \langle I^\top, I^\bot\rangle \in X\}$.

Proposition 3. *Let X be a directed set of $\mathcal{E}$-interpretations. Then, the $\mathcal{E}$-interpretation $I = \langle \bigcup X^\top, \bigcup X^\bot\rangle$ is the least upper bound of X.*

Proof.

1. First we show that $\langle \bigcup X^\top, \bigcup X^\bot\rangle$ is an $\mathcal{E}$-interpretation. i.e. $\bigcup X^\top \cap \bigcup X^\bot = \emptyset$.

   Assume we find $[A]$ with $[A] \in \bigcup X^\top \cap \bigcup X^\bot$. Then, there exist $\mathcal{E}$-interpretations $J_1 \in X$ and $J_2 \in X$ such that $[A] \in J_1^\top$ and $[A] \in J_2^\bot$. Because the set X is directed, it contains a common upper bound K of J_1 and J_2, where $[A] \in K^\top$ and $[A] \in K^\bot$. Then, $K^\top \cap K^\bot \neq \emptyset$ and K is not an $\mathcal{E}$-interpretation. This contradicts the precondition that X is a directed set of $\mathcal{E}$-interpretations. Hence, the assumption $[A] \in \bigcup X^\top \cap \bigcup X^\bot$ is false and $\langle \bigcup X^\top, \bigcup X^\bot\rangle$ is an $\mathcal{E}$-interpretation.

2. Next we show that $\langle \bigcup X^\top, \bigcup X^\bot\rangle$ is an upper bound of X, i.e. for all $J \in X$ we find $J^\top \subseteq \bigcup X^\top$ and $J^\bot \subseteq \bigcup X^\bot$.

   (a) $J^\top \subseteq \bigcup X^\top$ because for all $[A] \in J^\top$ we find $[A] \in \bigcup\{J^\top \mid J \in X\} = X^\top$.

   (b) $J^\bot \subseteq \bigcup X^\bot$ because for all $[A] \in J^\bot$ we find $[A] \in \bigcup\{J^\bot \mid J \in X\} = X^\bot$.

3. It remains to show that $\langle \bigcup X^\top, \bigcup X^\bot\rangle$ is the least upper bound of X, i.e. for every upper bound J of X, we have $\bigcup X^\top \subseteq J^\top$ and $\bigcup X^\bot \subseteq J^\bot$.

   Assume that X has an upper bound J where $\bigcup X^\top \not\subseteq J^\top$ or $\bigcup X^\bot \not\subseteq J^\bot$.

   (a) Assume there is an $[A] \in \bigcup X^\top$ such that $[A] \notin J^\top$. By $[A] \in \bigcup X^\top$, there is an $\mathcal{E}$-interpretation $K \in X$ where $[A] \in K^\top$. Because $[A] \notin J^\top$, J is not an upper bound for X.

   (b) Assume there is an $[A] \in \bigcup X^\bot$ such that $[A] \notin J^\bot$. By $[A] \in \bigcup X^\bot$, there is an $\mathcal{E}$-interpretation $K \in X$ where $[A] \in K^\bot$. Because $[A] \notin J^\bot$, J is not an upper bound for X.

   In both cases, the assumption that the least upper bound J of X differs from $\langle \bigcup X^\top, \bigcup X^\bot\rangle$ leads to a contradiction. Hence, $\langle \bigcup X^\top, \bigcup X^\bot\rangle$ is the least upper bound of X. □

Corollary 1. *The set of all $\mathcal{E}$-interpretations $\mathcal{I}$ is a complete partial order with respect to $\subseteq$.*

Proof. Reflexivity, antisymmetry and transitivity hold for $\subseteq$. The least element of $\mathcal{I}$ is $\langle\emptyset, \emptyset\rangle$. By Proposition 3, every directed subset of $\mathcal{I}$ has a least upper bound in $\mathcal{I}$. □

Proposition 4. *For each program $\mathcal{P}$ the mapping $\Phi^{\mathcal{E}}_{\mathcal{P}}$ is monotonic.*

Proof. Assume $I = \langle I^\top, I^\bot\rangle$ and $J = \langle J^\top, J^\bot\rangle$ are $\mathcal{E}$-interpretations for $\mathcal{P}$ with $I^\top \subseteq J^\top$ and $I^\bot \subseteq J^\bot$. We show that $\Phi^{\mathcal{E}}_{\mathcal{P}}(I) = I' = \langle I'^\top, I'^\bot\rangle \subseteq \langle J'^\top, J'^\bot\rangle = J' = \Phi^{\mathcal{E}}_{\mathcal{P}}(J)$.

1. $I'^\top \subseteq J'^\top$: By the definition of $I' = \Phi^{\mathcal{E}}_{\mathcal{P}}(I)$, we have $[A] \in I'^\top$ iff there is a rule $A \leftarrow Body$ in $\mathsf{g}\mathcal{P}$ such that

   (a) $Body = \top$ (i.e., $A \leftarrow Body$ is a fact) and, therefore, $J(Body) = \top$ and $[A] \in J'^\top$ or

   (b) $I(Body) = \min_t\{I(L) \mid L \text{ is a literal occurring in } Body\} = \top$. Then, for all conjuncts L in $Body$ we have one of the following cases:

      i. $L = B$ for a ground atom B and $I(B) = \top$ and, hence, $[B] \in I^\top \subseteq J^\top$.

      ii. $L = \neg B$ for a ground atom B and $I(B) = \bot$ and, hence, $[B] \in I^\bot \subseteq J^\bot$.

   In both cases, $J(Body) = \top$ and by definition of $J' = \Phi^{\mathcal{E}}_{\mathcal{P}}(J)$ we have $[A] \in J'^\top$.

2. $I'^\bot \subseteq J'^\bot$: By the definition of $\Phi^{\mathcal{E}}_{\mathcal{P}}$ we have $[A] \in I'^\bot$ iff

   (a) there exists a rule $A \leftarrow Body \in \mathsf{g}\mathcal{P}$ and

   (b) for all rules $A' \leftarrow Body \in \mathsf{g}\mathcal{P}$ with $[A'] = [A]$ we have $I(Body) = \bot$.

   Hence, for all rules $A_i \leftarrow Body_i \in \mathsf{g}\mathcal{P}$ where $[A_i] = [A]$ we have $Body_i = \bot$ (negative assumption) or $I(Body_i) = \min_t\{I(L) \mid L \text{ is a literal occurring in } Body_i\} = \bot$, i.e. there is a literal L occurring in $Body_i$ such that $I(L) = \bot$. We have one of the following cases:

   (a) If $Body_i = \bot$, then $J(Body_i) = \bot$

   (b) $L = B$ for a ground atom B and $I(B) = \bot$ and, therefore, $[B] \in I^\bot \subseteq J^\bot$.

   (c) $L = \neg B$ for a ground atom B and $I(B) = \top$ and, therefore, $[B] \in I^\top \subseteq J^\top$.

   In any of these cases, for all rules $A_i \leftarrow Body_i \in \mathsf{g}\mathcal{P}$ where $[A_i] = [A]$, we have $J(Body_i) = \min_t\{J(L) \mid L \text{ is a literal occurring in } Body\} = \bot$ and, therefore, $[A] \in J'^\bot$ by definition of $J' = \Phi^{\mathcal{E}}_{\mathcal{P}}(J)$.

$\Phi^{\mathcal{E}}_{\mathcal{P}}$ is monotonic because $I' \subseteq J'$, i.e. $I'^\top \subseteq J'^\top$ and $I'^\bot \subseteq J'^\bot$. □

However, $\Phi^{\mathcal{E}}_{\mathcal{P}}$ is generally not continuous. Consider the program

$$\mathcal{P} = \{q(a) \leftarrow \top,\ q(s(X)) \leftarrow q(X),\ p \leftarrow \neg q(X)\}$$

and the empty equational theory. The least fixed point of $\Phi^{\emptyset}_{\mathcal{P}}$ is

$$\langle \{[q(s^k(a))] \mid k \in \mathbb{N}\}, \{[p]\}\rangle$$

and is reached after iterating $\Phi^{\emptyset}_{\mathcal{P}}$ $\omega + 1$ times, where ω is the first limit ordinal. Hence, by Kleene's fixed point theorem (see e.g. [6]), $\Phi^{\emptyset}_{\mathcal{P}}$ is not continuous. One should observe that the Herbrand $\emptyset$-base contains infinitely many equivalence classes $[p]$, $[q(a)]$, $[q(s(a))]$, …, each of which has one element.

Likewise, consider the program

$$\mathcal{P} = \{q(1) \leftarrow \top,\ q(X \circ a) \leftarrow q(X),\ p \leftarrow \neg q(X)\}$$

and the AC1-theory presented in Section 1. The least fixed point of $\Phi_{\mathcal{P}}^{AC1}$ is

$$\langle \{[q(1 \circ \overbrace{a \circ \ldots \circ a}^{k})] \mid k \in \mathbb{N}\}, \{[p]\}\rangle \tag{1}$$

and is reached after iterating $\Phi_{\mathcal{P}}^{AC1}$ $\omega + 1$ times. Again, $\Phi_{\mathcal{P}}^{AC1}$ is not continuous. One should observe that the Herbrand $AC1$-base contains infinitely many equivalence classes, viz. $[p]$, $[q(1)]$, $[q(a)]$, $[q(a \circ a)]$, $\ldots$. With the exception of $[p]$ each of these equialence classes is infinite because $a \equiv_{AC1} a \circ 1 \equiv_{AC1} a \circ 1 \circ 1 \equiv_{AC1} \ldots$.

Proposition 5. *For each finite propositional program $\mathcal{P}$ the mapping $\Phi_{\mathcal{P}}^{\mathcal{E}}$ is continuous.*

Proof. Because the set of all propositional variables in a finite propositional program $\mathcal{P}$ is finite and we only have finitely many truth values, the set $\mathcal{I}$ of all $\mathcal{E}$-interpretations is finite. By Corollary 1, $\mathcal{I}$ is a complete partial order with respect to $\subseteq$. By Proposition 4, $\Phi_{\mathcal{P}}^{\mathcal{E}}$ is monotonic on $\mathcal{I}$. Hence, $\Phi_{\mathcal{P}}^{\mathcal{E}}$ is continuous on $\mathcal{I}$ because monotonic mappings over finite and complete partial orders are continuous (see e.g. [11]). □

Proposition 6. *If the Herbrand $\mathcal{E}$-base for a program $\mathcal{P}$ and a set of equations $\mathcal{E}$ is finite, then the mapping $\Phi_{\mathcal{P}}^{\mathcal{E}}$ is continuous.*

Proof. Let $\mathcal{P}$ be a program and $\mathcal{E}$ be a set of equations such that the Herbrand $\mathcal{E}$-base is finite. The result follows immediately from Proposition 5 and the fact that there is a bijection between the Herbrand $\mathcal{E}$-base and an equally large set of propositional atoms. □

As an example consider $\mathcal{E} = \{a \approx c\}$ and

$$\mathcal{P} = \{q(a) \leftarrow \top,\ q(b) \leftarrow \top,\ p(X) \leftarrow q(X)\}.$$

In this case, the Herbrand $\mathcal{E}$-base consists of $[q(a)]$, $[q(b)]$, $[p(a)]$, and $[p(b)]$ with $[q(a)] = [q(c)]$ and $[p(a)] = [p(c)]$. Let $r_1 - r_4$ be four propositional variables and define the bijection

$$[q(a)] \Leftrightarrow r_1,\ [q(b)] \Leftrightarrow r_2,\ [p(a)] \Leftrightarrow r_3,\ [p(c)] \Leftrightarrow r_4.$$

If this bijection is applied to each element of an equivalence class, then the following propositional program is equivalent to $\mathtt{g}\mathcal{P}$:

$$\{r_1 \leftarrow \top,\ r_2 \leftarrow \top,\ r_3 \leftarrow r_1,\ r_4 \leftarrow r_2\}.$$

One should observe that the ground instances $p(a) \leftarrow q(a)$ and $p(c) \leftarrow q(c)$ are both mapped onto $r_3 \leftarrow r_1$.

Unfortunately, this result is insufficient for using the fluent calculus in general as the fluent calculus utilizes a binary function symbol $\circ$ in order to represent multisets of fluents. As shown in (1), $\circ$ may be used to define infinitely many equivalence classes in the Herbrand $\mathcal{E}$-base of a program. However, in the fluent calculus the function symbol $\circ$ is only used to represent multisets. If we consider only finite multisets in the same way as Selman, Levesque, and Mitchell consider only finite plans in [27], then the Herbrand $\mathcal{E}$-base for a given program is finite and, consequently, Proposition 6 applies. Likewise, if the initial state is finite and there is no action whose application leads to an increase of the number of fluents occurring in a state, then the Herbrand $\mathcal{E}$-base of such a program can also be restricted to a finite set. In particular, in the context of human reasoning episodes such restrictions appear to be quite reasonable. In the trolley problems discussed in Section 5 the largest multiset has size six, the initial states are always finite, and there is no action which increases the number of fluents occurring in a state.

On the other hand, if we consider finite datalog programs and finite sets of equations between constants, then the Herbrand $\mathcal{E}$-base is also finite. However, such a class is of no particular interest here as we want to model multisets by means of an AC1-operator.

We proceed to show that for a given program $\mathcal{P}$ and a given set of equations $\mathcal{E}$, the least $\mathcal{E}$-model of $wc\mathcal{P}$ and the least fixed point of $\Phi_{\mathcal{P}}^{\mathcal{E}}$ coincide.

Lemma 1. *Let $\mathcal{P}$ be a program, J be the least fixed point of $\Phi_{\mathcal{P}}^{\mathcal{E}}$, and I be an $\mathcal{E}$-model of $wc\mathcal{P}$. Then, for every ground atom A the following holds:*

1. *If $J(A) = \top$, then $I(A) = \top$.*
2. *If $J(A) = \bot$, then $I(A) = \bot$.*

Proof. Let J be the least fixed point of $\Phi_{\mathcal{P}}^{\mathcal{E}}$. It can be computed by iterating $\Phi_{\mathcal{P}}^{\mathcal{E}}$ starting from the empty interpretation as follows:

$$J_0 = \langle \emptyset, \emptyset \rangle, \tag{2}$$
$$J_\alpha = \Phi_{\mathcal{P}}^{\mathcal{E}}(J_{\alpha-1}) \text{ for every non-limit ordinal } \alpha > 0, \tag{3}$$
$$J_\alpha = \bigcup_{\beta<\alpha} J_\beta \text{ for every limit ordinal } \alpha. \tag{4}$$

Then, there must be some ordinal $\alpha_{\mathcal{P}}$ such that $J = J_{\alpha_{\mathcal{P}}}$. We will prove by transfinite induction that for every ordinal α and every ground atom A the following holds:

1. If $J_\alpha(A) = \top$, then $I(A) = \top$.
2. If $J_\alpha(A) = \bot$, then $I(A) = \bot$.

With this result, the claim will follow from Corollary 1 and Proposition 4.

Turning to the induction proof, we consider three cases: the base case when the ordinal $\alpha = 0$ and two inductive cases, one for non-limit ordinals and the other for limit ordinals:

1. Let $\alpha = 0$. Then, by (2) we find $J_\alpha = \langle \emptyset, \emptyset \rangle$. Because there is no atom such that $J_\alpha(A) = \top$ or $J_\alpha(A) = \bot$, the claim follows trivially.
2. Let $\alpha > 0$ be a non-limit ordinal. By the inductive hypothesis we find for every ground atom B that:

$$\text{If } J_{\alpha-1}(B) = \top, \text{ then } I(B) = \top. \tag{5}$$
$$\text{If } J_{\alpha-1}(B) = \bot, \text{ then } I(B) = \bot. \tag{6}$$

Moreover, by (3) we find $J_\alpha = \Phi_{\mathcal{P}}^{\mathcal{E}}(J_{\alpha-1})$ and we distinguish two two cases:

(a) If $J_\alpha(A) = \top$, then according to the definition of $\Phi_{\mathcal{P}}^{\mathcal{E}}$ there must be some rule $A' \leftarrow Body_i$ in $\mathsf{g}\mathcal{P}$ with $[A'] = [A]$ such that $J_{\alpha-1}(Body_i) = \top$. Let

$$Body_i = B_1 \wedge B_2 \wedge \cdots \wedge B_k \wedge \neg B_{k+1} \wedge \neg B_{k+2} \wedge \cdots \wedge \neg B_m,$$

where each B_j, $1 \leq j \leq m$, is a ground atom. Then, for each s with $1 \leq s \leq k$ we have $J_{\alpha-1}(B_s) = \top$ and for each t with $k < t \leq m$ we have $J_{\alpha-1}(B_t) = \bot$. Using the inductive hypothesis and, in particular, (5) and (6) we learn that for each s with $1 \leq s \leq k$ we have $I(B_s) = \top$ and for each t with $k < t \leq m$ we have $I(B_t) = \bot$. Hence, $I(Body_i) = \top$. Furthermore, in $wc\mathcal{P}$ there will be a formula of the form $A \leftrightarrow F$, where F is a disjunction with $Body_i$ as one of the disjuncts. Thus, we have $I(F) = \top$ and also $I(A \leftrightarrow F) = \top$ because I is a model of $wc\mathcal{P}$. This implies that $I(A) = \top$.

(b) If $J_\alpha(A) = \bot$, then according to the definition of $\Phi^{\mathcal{E}}_{\mathcal{P}}$ there must be a rule of the form $A \leftarrow \mathit{Body}$ in $\mathsf{g}\mathcal{P}$ and all rules of the form $A' \leftarrow \mathit{Body}_i$ in $\mathsf{g}\mathcal{P}$ with $[A'] = [A]$ must have $J_{\alpha-1}(\mathit{Body}_i) = \bot$. Pick an arbitrary but fixed j and let

$$\mathit{Body}_j = B_1 \wedge B_2 \wedge \cdots \wedge B_k \wedge \neg B_{k+1} \wedge \neg B_{k+2} \wedge \cdots \wedge \neg B_m,$$

where B_l, $1 \leq l \leq m$ are ground atoms. We have to consider two cases:

i. There is some s with $1 \leq s \leq k$ such that $J_{\alpha-1}(B_s) = \bot$. Then, by (6) we find $I(B_s) = \bot$ and, hence, $I(\mathit{Body}_j) = \bot$.

ii. There is some t with $k < t \leq m$ such that $J_{\alpha-1}(B_t) = \top$. Then, by (5) we obtain $I(B_t) = \top$ and, hence, $I(\mathit{Body}_j) = \bot$.

In either case we have $I(\mathit{Body}_j) = \bot$. Because j was arbitrarily chosen, we can conclude that for every i we have $I(\mathit{Body}_i) = \bot$. Furthermore, in $\mathit{wc}\mathcal{P}$ there is a formula of the form $A \leftrightarrow F$ with $F = \mathit{Body}_1 \vee \mathit{Body}_2 \vee \ldots$. So we have $I(F) = \bot$. Because I is a model of $\mathit{wc}\mathcal{P}$ we find $I(A \leftrightarrow F) = \top$. This implies that $I(A) = \bot$.

3) Let α be a limit ordinal. By the inductive hypothesis we find for every ground atom B and every ordinal $\beta < \alpha$ that:

$$\text{If } J_\beta(B) = \top, \text{ then } I(B) = \top. \tag{7}$$

$$\text{If } J_\beta(B) = \bot, \text{ then } I(B) = \bot. \tag{8}$$

Moreover, by (4) we have $J_\alpha = \bigcup_{\beta<\alpha} J_\beta$. There are again two cases to consider:

(a) If $J_\alpha(A) = \top$, then there is some ordinal $\beta < \alpha$ such that $J_\beta(A) = \top$. By the inductive hypothesis (7) we have $I(A) = \top$.

(b) If $J_\alpha(A) = \bot$, then there is some ordinal $\beta < \alpha$ such that $J_\beta(A) = \bot$. By the inductive hypothesis (8) we have $I(A) = \bot$. □

Lemma 2. *If $\mathcal{P}$ is a program and J a fixed point of $\Phi^{\mathcal{E}}_{\mathcal{P}}$, then J is an $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$.*

Proof. By the definition of $\Phi^{\mathcal{E}}_{\mathcal{P}}$ an $\mathcal{E}$-interpretation $I = \langle I^\top, I^\bot \rangle$ is a fixed point of $\Phi^{\mathcal{E}}_{\mathcal{P}}$ iff

$$\begin{array}{rcl} I^\top & = & \{[A] \mid \text{there exists } A \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P} \text{ and } I(\mathit{Body}) = \top\} \\ I^\bot & = & \{[A] \mid \text{there exists } A \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P} \\ & & \quad \text{and for all } A' \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P} \text{ with } [A] = [A'] \text{ we find } I(\mathit{Body}) = \bot\} \end{array}$$

We show that for every equivalence $A \leftrightarrow F \in \mathit{wc}\mathcal{P}$ we have $I(A \leftrightarrow F) = \top$, i.e. $I(A) = I(F)$.

1. For every $[A] \in I^\top$ there is a rule $A \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P}$ such that $I(\mathit{Body}) = \top$. Hence, for each equivalence $A \leftrightarrow F \in \mathit{wc}\mathcal{P}$, where $F = \mathit{Body} \vee F'$ for a (possibly empty) disjunction F', we have $I(F) = I(\mathit{Body} \vee F') = \max_t(I(\mathit{Body}), I(F')) = \top$. Hence, $I(A) = I(\mathit{Body} \vee F') = I(F)$ and, therefore, $I(A \leftrightarrow F) = \top$.

2. For every $[A] \in I^\bot$ there is a rule $A \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P}$ and for all rules $A' \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P}$ with $[A'] = [A]$ we have $I(\mathit{Body}) = \bot = I(A)$. Hence, for each equivalence $A' \leftrightarrow F \in \mathit{wc}\mathcal{P}$ with $[A'] = [A]$ we have $I(F) = \bot$ and, therefore, $I(A' \leftrightarrow F) = \top$.

3. For every $[A] \not\in I^\top \cup I^\bot$ we have two possibilities:

   (a) There is no rule $A' \leftarrow \mathit{Body} \in \mathsf{g}\mathcal{P}$ with $[A'] = [A]$ and, therefore, there is no equivalence $A' \leftrightarrow F \in \mathit{wc}\mathcal{P}$.

(b) There are rules $A'_i \leftarrow \mathit{Body}_i \in \mathsf{g}\mathcal{P}$ for $i \in \{1, \ldots, n\}$ with $[A'_i] = [A]$ and $I(A'_i) = \mathrm{U}$, but neither $I(\mathit{Body}_i) = \bot$ for all $i \in \{1, \ldots, n\}$ nor there is an $i \in \{1, \ldots, n\}$ such that $I(\mathit{Body}_i) = \top$. Hence, $I(\bigvee_{i \in \{1,\ldots,n\}} \mathit{Body}_i) = \max_t(\{I(\mathit{Body}_i) \mid i \in \{1, \ldots, n\}\}) = \mathrm{U}$ and, therefore, $I(A' \leftrightarrow \bigvee_{i \in \{1,\ldots,n\}} \mathit{Body}_i) = \top$.

Hence, for each equivalence $A' \leftrightarrow F \in \mathit{wc}\mathcal{P}$ with $[A'] = [A]$ we have $I(A' \leftrightarrow F) = \top$ and, therefore, I is an $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$. □

Proposition 7. *If J is the least fixed point of $\Phi^{\mathcal{E}}_{\mathcal{P}}$, then J is a minimal $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$.*

Proof. By Lemma 2, the least fixed point J of $\Phi^{\mathcal{E}}_{\mathcal{P}}$ is an $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$. By Lemma 1, for every $\mathcal{E}$-model I of $\mathit{wc}\mathcal{P}$ we have $J^\top \subseteq I^\top$ and $J^\bot \subseteq I^\bot$, i.e. $J \subseteq I$. Hence, J is the unique minimal $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$. □

Proposition 8. *If I is a minimal $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$, then I is the least fixed point of $\Phi^{\mathcal{E}}_{\mathcal{P}}$.*

Proof. Let I be a minimal $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$ and J the least fixed point of $\Phi^{\mathcal{E}}_{\mathcal{P}}$. From Lemma 1 we learn that $J^\top \subseteq I^\top$ and $J^\bot \subseteq I^\bot$. From Proposition 7 we learn that J is a minimal $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$. But then $I = J$ because, otherwise, we have a conflict with the minimality of I. □

Proof of Theorem 2. The claim that $\mathit{wc}\mathcal{P}$ has a least $\mathcal{E}$-model follows from Propositions 7 and 8 and the fact that the least fixed point of $\Phi^{\mathcal{E}}_{\mathcal{P}}$ is unique. □

Theorem 3. *I is the least fixed point of $\Phi^{\mathcal{E}}_{\mathcal{P}}$ iff I is the least $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$.*

Proof. The claim follows immediately from Propositions 7 and 8 and Theorem 2. □

Let $\mathcal{M}^{\mathcal{E}}_{\mathcal{P}}$ denote the least $\mathcal{E}$-model of $\mathit{wc}\mathcal{P}$. $\mathcal{P}$ *entails a formula F under the weak completion semantics with equality*, in symbols $\mathcal{P} \models^{\mathcal{E}}_{wcs} F$, iff $\mathcal{M}^{\mathcal{E}}_{\mathcal{P}}(F) = \top$.

5 Moral Decision Making

We will consider three trolley problems: the bystander, the footbridge, and the loop case. All cases are taken from [25] with some minor adaptations.

The Bystander Case *A trolley, whose conductor has fainted, is headed towards two people walking on the main track.*[1] *The banks of the track are so steep that these two people will not be able to get off the track in time. Hank is standing next to a switch, which can turn the trolley onto a side track, thereby preventing it from killing the two people. However, there is a man standing on the side track. Hank can change the switch, killing him. Or he can refrain from doing so, letting the two die. Is it morally permissible for Hank to change the switch?*

The case is illustrated in Figure 1. The tracks are divided into segments 0, 1, and 2, the arrow represents that the trolley t is moving forward and that the track is clear (c), the switch is in position m (main) but can be changed into position s (side), and a bullet above a track segment represents a human (h) on this track. t, c, and h may be indexed to denote the track to which they apply. In addition, we need a fluent d denoting a dead human.

[1] Note that in the original trolley problem, five people are on the main track. For the sake of simplicity, here and in the following we assume that only two people are on the main track.

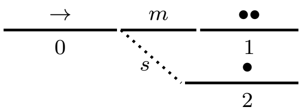

Figure 1: The bystander case

We choose to represent a state by two fluent terms: the causalities and all other fluents. Hence, the initial state of the bystander case is the pair

$$(t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1),^{2}$$

where the causalities are represented in the second element of the pair by 1. As mentioned in Section 1, 1 is the unit in the AC1-theory used in the fluent calculus to represent multisets of fluents. Here it denotes the fact that initially there are no causalities. Casualties will play a special role when preferring one action over another as will be discussed later in this section.

There are two kinds of actions, the ones which can be performed by Hank (*donothing*, *change*), and the actions which are performed by the trolley (*downhill*, *kill*). Herein, we will only represent the actions by the trolley explicitly with the help of a relation symbol *action* specifying the preconditions as well as the immediate effects of actions:

$$\begin{array}{l} \mathit{action}(t_0 \circ c_0 \circ m, 1, \mathit{downhill}, t_1 \circ c_0 \circ m, 1) \leftarrow \top, \\ \mathit{action}(t_0 \circ c_0 \circ s, 1, \mathit{downhill}, t_2 \circ c_0 \circ s, 1) \leftarrow \top, \\ \\ \mathit{action}(t_1 \circ h_1, 1, \mathit{kill}, t_1, d) \leftarrow \top, \\ \mathit{action}(t_2 \circ h_2, 1, \mathit{kill}, t_2, d) \leftarrow \top. \end{array}$$

If the trolley is on track 0, this track is clear, and the switch is in position m, then it will run downhill onto track 1 whereas track 0 remains clear and the switch will remain in position m; if, however, the switch is in position s, it will run downhill onto track 2. If the trolley is on either track 1 or 2 and there is a human on this track, it will kill the human.

The actions of Hank will be the base cases in the definition of causality:[3]

$$\begin{array}{l} \mathit{causes}(\mathit{donothing}, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1) \leftarrow \top, \\ \mathit{causes}(\mathit{change}, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ h_2, 1) \leftarrow \top. \end{array} \tag{9}$$

The recursive case of the definition of causality is given as

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

It checks whether in a given state $(P_1 \circ Z_1, P_2 \circ Z_2)$ an action A' is applicable, which is the case if the preconditions (P_1, P_2) are contained in the given state. If this holds, then the action is

[2] Or, alternatively, the pair of multisets $(\dot{\{}t_0, c_0, m, h_1, h_1, h_2\dot{\}}, \dot{\{}\,\dot{\}})$.

[3] In the original version of the fluent calculus, *causes* is a ternary predicate stating that the execution of a plan transfers an initial into a goal state. Its base case is of the form $\mathit{causes}(X, [\,], X)$, i.e., the empty plans transforms arbitrary states X into X. Generating models bottom up using a semantic operator one has to consider all ground instances of this atom, which is usually too large to consider as a base case for human reasoning episodes. The solution presented in this paper overcomes this problem in that we only have a small number of base cases depending on the number of options an agent like Hank may consider.

executed leading to the successor state $(E_1 \circ Z_1, E_2 \circ Z_2)$, where (E_1, E_2) are the direct effects of the action A'. In other words, if an action is applied, then its preconditions are consumed and its direct effects are produced. Such an action application is considered to be a ramification [29] with respect to the initial action performed by Hank. Hence, the first argument A of *causes* is not changed. The execution of an action is also conditioned by $\neg ab(A')$, where ab is an abnormality predicate. Such abnormalities were introduced in [28] to represent conditionals as licenses for inference. In this example, there is nothing abnormal known with respect to the actions *downhill* and *kill* and, consequently, the assumptions

$$\begin{array}{l} ab(\mathit{downhill}) \leftarrow \bot, \\ ab(\mathit{kill}) \leftarrow \bot \end{array}$$

are added to the program. But we can imagine situations, where the trolley will only cross the switch if the switch is not broken.[4]

One should observe that negative assumptions are overridden once positive information becomes available and is added as fact to the program. This will be demonstrated in the footbridge case below. If we would replace $\neg ab(A')$ in the definition of the recursive case of *causes* above by some normality atom like $\mathit{normal}(A')$ and add a positive fact like $\mathit{normal}(\mathit{downhill}) \leftarrow \top$ to the program, then this fact cannot be overridden by the addition of a negative assumption. Rather the program must be revised in case negative information becomes available. Such a revision operation is more complex than the addition of a fact.

Let $\mathcal{PB}$ be the program consisting of the clauses mentioned in this paragraph. In this program, the largest multiset has size six and there is no action whose execution will increase the size of a multiset. Hence, $\Phi^{AC1}_{\mathcal{PB}}$ is continuous and, hence, its least fixed point can be computed by iterating $\Phi^{AC1}_{\mathcal{PB}}$ starting with the empty interpretation $\langle \emptyset, \emptyset \rangle$.

Hank has the choice to do nothing or to change the switch. Depending on his decision, the trolley will execute its actions which are computed as ramifications in the fluent calculus [29]. If Hank is doing nothing, then the least fixed $AC1$-model of $\mathcal{PB}$ – which is equal to the least fixed point of $\Phi^{AC1}_{\mathcal{PB}}$ – is computed by iterating $\Phi^{AC1}_{\mathcal{PB}}$ starting with the empty interpretation $\langle \emptyset, \emptyset \rangle$. The following equivalence classes will be mapped to true in subsequent iterations:

$$\begin{array}{ll} [\mathit{causes}(\mathit{donothing}, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)] & \text{(initial state)}, \\ [\mathit{causes}(\mathit{donothing}, t_1 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)] & \text{(trolley moving downhill to track } t_1\text{)}, \\ [\mathit{causes}(\mathit{donothing}, t_1 \circ c_0 \circ m \circ h_1 \circ h_2, d)] & \text{(trolley killing the first human)}, \\ [\mathit{causes}(\mathit{donothing}, t_1 \circ c_0 \circ m \circ h_2, d \circ d)] & \text{(trolley killing the second human)}. \end{array}$$

No further action is applicable to the representative of the final equivalence class. The two people on the main track will be killed.

However, if Hank is changing the switch, then the least fixed point of $\Phi^{AC1}_{\mathcal{PB}}$ contains

$$[\mathit{causes}(\mathit{change}, t_2 \circ c_0 \circ s \circ h_1 \circ h_1, d)]$$

and no further action is applicable in this case. The two people on the main track will be saved but the person on the side track will be killed.

The two final cases can be compared by means of a *prefer* clause:

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

[4] If the switch is broken, the trolley may derail. Such a scenario can be modeled in WCSE as well, but it is beyond the scope of this paper to discuss it in detail.

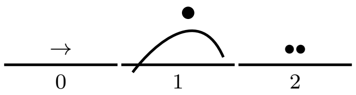

Figure 2: The footbridge case

Comparing D_1 and $D_1 \circ d \circ D_2$, action A_2 leads to at least one more dead person than action A_1. Hence, A_1 is preferred over A_2 if nothing abnormal is known about A_1.

Under an utilitarian point of view [3], the *change* action is preferable to the *donothing* action as it will kill fewer humans. On the other hand, we know that a purely utilitarian view is not allowed in case of human causalities. Hank may ask himself: *Would I still save the humans on the main track if there were no human on the side track and I changed the switch?* This is a counterfactual. But we can easily deal with it in WCSE by starting a new computation with the additional fact

$$causes(change, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1) \leftarrow \top. \qquad (10)$$

Comparing (9) with (10), h_2 has been replaced by c_2. There is no human on track 2 anymore and, hence, this track is clear. This is a minimal change necessary to satisfy the precondition of the counterfactual. In this case, the least $\mathcal{E}$-model of the extended program will contain

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

and no further action is applicable in this case. 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 Hank to conclude that changing the switch is permissible within the doctrine of double effect [1].

The Footbridge Case *The case is similar to the bystander case except that instead of the switch a footbridge is crossing the main track. Ian is standing on the footbridge next to a heavy man, which he can throw on the track in the path of the trolley to stop it. Is it morally permissible for Ian to throw the man down?*

This case is illustrated in Figure 2. The track is again segmented. We use b_1 to denote that there is a heavy human on the footbridge crossing segment 1 of the track. Ian has two possibilities: *donothing* and *throwdown*. They are represented as the base cases in the definition of causality:

$$\begin{aligned} & causes(donothing, t_0 \circ c_0 \circ c_1 \circ b_1 \circ h_2 \circ h_2, 1) \leftarrow \top, \\ & causes(throwdown, t_0 \circ c_0 \circ h_2 \circ h_2, d) \leftarrow \top. \end{aligned}$$

One should observe that in the case of *donothing* track 1 is clear (c_1), whereas this does not hold if Ian has decided to throw down the heavy man. In the latter case, a dead body is blocking track 1.

As in the footbridge case, one is tempted to reason that the *throwdown* action is preferable to the *donothing* action as it will kill fewer humans. But throwing down a heavy man involves

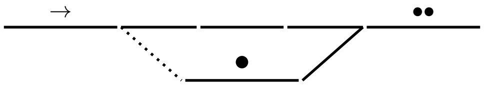

Figure 3: The loop case

an intentional direct kill, and intentional kills are not allowed under the doctrine of double effect. This can be modeled with the help of the abnormality predicate *ab* by

$$\begin{aligned}
&\mathit{prefer}(A_1, A_2) \leftarrow \mathit{causes}(A_1, Z_1, D_1)\\
&\qquad\qquad\qquad\quad \wedge\ \mathit{causes}(A_2, Z_2, D_1 \circ d \circ D_2)\\
&\qquad\qquad\qquad\quad \wedge\ \neg \mathit{ab}_{\mathit{prefer}}(A_1),\\
&\mathit{ab}_{\mathit{prefer}}(\mathit{throwdown}) \leftarrow \bot,\\
&\mathit{ab}_{\mathit{prefer}}(\mathit{throwdown}) \leftarrow \mathit{intensionaldirectkill}(\mathit{throwdown}),\\
&\mathit{intensionaldirectkill}(\mathit{throwdown}) \leftarrow \top.
\end{aligned}$$

Hence, throwing down the heavy man is not preferred and, thus, not permissible. The example demonstrates the way abnormalities are used in the WCSE. If nothing is known, then a negative assumption about the abnormality is made. This assumption can be overridden once additional knowledge becomes available. In this case we learn that intentional direct kills override the negative assumption. Moreover, from the specification of the *throwdown* action we can derive that the killing of the heavy man was intentional as it is a direct effect of this action.

The Loop Case *The case is similar to the bystander case. Ned is standing next to a switch, which he can throw, that will temporarily turn the trolley onto a loop side track. There is a heavy man on the side track. If the trolley hits the heavy man, then this will slow down the trolley, giving the two people on the main track sufficient time to escape. But it will kill the heavy man. Is it morally permissible for Ned to throw the switch?*

This case is illustrated in Figure 3. Ned can reason that if he does nothing, then the humans on the main track will be killed. Likewise, if he changes the switch, then the humans on the main track will be saved whereas the human on the side track will be killed. But the counterfactual *if there were no human on the side and he changes the switch, then he would still save the humans on the main track* will be false. Hence, according to the doctrine of double effect changing the switch is not permissible. However, the doctrine of triple effect [17] allows to distinguish between direct and indirect intentional kills such that the *change* action becomes permissible under the doctrine of triple effects.

This example can also be modeled in WCSE. Because killing a human is not a direct effect of the *change* action we may add

$$\begin{aligned}
&\mathit{ab}_{\mathit{prefer}}(\mathit{change}) \leftarrow \mathit{intensionaldirectkill}(\mathit{change}),\\
&\mathit{intensionaldirectkill}(\mathit{change}) \leftarrow \bot
\end{aligned}$$

to the previous program. Consequently, the *change* action will be preferred over the *donothing* action. A properly revised definition for permissibility will allow Ned to conclude that changing the switch is permissible under the doctrine of triple effect.

Table 1 gives a summary according to which view which action is permissible for each case.

	Bystander Case	Loop Case	Footbridge Case
Doctrine of double effect	*change*	-	-
Doctrine of triple effect	*change*	*change*	-
Utilitarian view	*change*	*change*	*throwdown*

Table 1: The three cases and the permissible actions according to the different views.

6 Conclusion

In this paper we have generalized the weak completion semantics (WCS) to handle equalities (WCSE). In particular, we have rigorously proven that key properties of WCS like the existence of a least model for normal logic programs and the fact that this model can be computed as least fixed point of an appropriate semantic operator can be extended to WCSE. We have shown that the semantic operator is not continuous in general, but continuous for many practical cases.

WCSE allows us to reason about actions and causality based on the fluent calculus if we put a bound on the size of the multisets representing states. But, the fluent calculus approach had to be further modified such that the computation of the least $\mathcal{E}$-model by the corresponding semantic operator does not lead to large or even infinite sets. In particular, the possible actions of an agent are considered as base cases in the definition of causality, whereas the actions of the system influenced by the agent are applied as ramifications in the recursive cases in the definition of causality.

As examples we have considered three trolley problems taken from [25]. We have shown that these problems can be solved in a coherent form under a unified framework using WCSE by applying the principle of utilitarianism and the doctrines of double and triple effect. These lead to different results concerning the permissibility of actions according to certain moral theories. We do not aim at making any moral judgments, but an agent using WCSE can reason about these different views. According to [30], 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.

The examples also demonstrate that within WCSE we can evaluate counterfactuals by making minimal changes of the initial situations considered by an agent.

WCSE is a conservative extension of WCS. All human reasoning examples which can be adequately modeled by WCS like the suppression task, the selection task, and human syllogistic reasoning can be modeled by WCSE as well.

On the other hand, there are many open questions. The examples discussed in this paper are hand-crafted and we would like to develop an extension, where examples taken from the moral machine project (moralmachine.mit.edu) can be automatically treated under WCSE. We also would like to generalize the reasoning such that if an action does something good and nothing abnormal is known, then it is permissible. This, however, requires a formalization of 'something good' and very likely a formalization of 'something bad'. And, we should have a closer look at counterfactuals and minimal change.

References

[1] T. Aquinas. Summa Theologica II-II, q. 64, art. 7, "Of Killing". In W. P. Baumgarth and R.J. Regan, editors, *On Law, Morality, and Politics*, pages 226–227. Hackett Publishing Co., Indianapolis, 1988.

[2] L. Barrett. *An Architecture for Structured, Concurrent, Real-time Action.* PhD thesis, Computer Science Division, University of California at Berkeley, 2010.

[3] J. Bentham. *An Introduction to the Principles of Morals and Legislation.* Dover Publications Inc., 2009.

[4] R. M. J. Byrne. Suppressing valid inferences with conditionals. *Cognition*, 31:61–83, 1989.

[5] K. L. Clark. Negation as failure. In H. Gallaire and J. Minker, editors, *Logic and Databases*, pages 293–322. Plenum, New York, 1978.

[6] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order.* Cambridge University Press, 2nd edition, 2002.

[7] E.-A. Dietz, S. Hölldobler, and M. Ragni. A computational logic approach to the suppression task. In N. Miyake, D. Peebles, and R. P. Cooper, editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1500–1505. Cognitive Science Society, 2012.

[8] E.-A. Dietz, S. Hölldobler, and M. Ragni. A computational logic approach to the abstract and the social case of the selection task. In *Proceedings Eleventh International Symposium on Logical Formalizations of Commonsense Reasoning*, 2013. commonsensereasoning.org/2013/proceedings.html.

[9] M. Fitting. A Kripke–Kleene semantics for logic programs. *Journal of Logic Programming*, 2(4):295–312, 1985.

[10] G. Große, S. Hölldobler, and J. Schneeberger. Linear deductive planning. *Journal of Logic and Computation*, 6(2):233–262, 1996.

[11] Carl A. Gunter and Dana S. Scott. Semantic domains. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Sematics (B)*, pages 633–674. North-Holland, 1990.

[12] S. Hölldobler. *Foundations of Equational Logic Programming*, volume 353 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, Heidelberg, 1989.

[13] S. Hölldobler and F. Jovan. Advanced petri nets and the fluent calculus. In S. Hölldobler, A. Malikov, and C. Wernhard, editors, *Proceedings of the Young Scientists' International Workshop on Trends in Information Processing*, volume 1145 of *CEUR Workshop Proceedings*, pages 15–24. CEUR-WS.org, 2014. http://ceur-ws.org/Vol-1145/.

[14] S. Hölldobler and C. D. P. Kencana Ramli. Logic programs under three-valued Łukasiewicz's semantics. In P. M. Hill and D. S. Warren, editors, *Logic Programming*, volume 5649 of *LNCS*, pages 464–478, Berlin, Heidelberg, 2009. Springer.

[15] S. Hölldobler and J. Schneeberger. A new deductive approach to planning. *New Generation Computing*, 8:225–244, 1990.

[16] J. Jaffar, J-L. Lassez, and M. J. Maher. A theory of complete logic programs with equality. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pages 175–184. ICOT, 1984.

[17] F. M. Kamm. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm.* Oxford University Press, Oxford, 2006.

[18] C. D. P. Kencana Ramli. Logic programs and three-valued consequence operators. Master's thesis, International Center for Computational Logic, TU Dresden, 2009.

[19] S. Khemlani and P. N. Johnson-Laird. Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, 138(3):427–457, 2012.

[20] R.A. Kowalski and M. Sergot. A logic-based calculus of events. *New Generation Computing*, 4:67–95, 1986.

[21] J. Łukasiewicz. O logice trójwartościowej. *Ruch Filozoficzny*, 5:169–171, 1920. English translation: On Three-Valued Logic. In: *Jan Łukasiewicz Selected Works.* (L. Borkowski, ed.), North Holland, 87-88, 1990.

[22] J. McCarthy. Situations and actions and causal laws. Stanford Artificial Intelligence Project:

Memo 2, 1963.

[23] R. S. Nickerson. *Conditional Reasoning.* Oxford University Press, 2015.

[24] A. Oliviera da Costa, E.-A. Dietz Saldanha, S. Hölldobler, and M. Ragni. A computational logic approach to human syllogistic reasoning. In G. Gunzelmann, A. Howes, T. Tenbrink, and E. J. Davelaar, editors, *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, pages 883–888, Austin, TX, 2017. Cognitive Science Society.

[25] L. M. Pereira and A. Saptawijaya. *Programming Machine Ethics.* Springer, Berlin, Heidelberg, 2016.

[26] R. Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation — Papers in Honor of John McCarthy*, pages 359–380. Academic Press, 1991.

[27] B. Selman, H. Levesque, and D. Mitchell. A new method for solving hard satisfiability problems. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 440–446, Menlo Park, 1992. AAAI Press.

[28] K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science.* MIT Press, 2008.

[29] M. Thielscher. Controlling semi-automatic systems with FLUX (extended abstract). In C. Palamidessi, editor, *Logic Programming*, volume 2916 of *Lecture Notes in Computer Science*, pages 515–516, Berlin, Heidelberg, 2003. Springer.

[30] Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press, Inc., New York, NY, USA, 2008.

[31] P. C. Wason. Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, 20:273–281, 1968.