Counterfactuals in Logic Programming with Applications to Agent Morality

Luís Moniz Pereira* and Ari Saptawijaya**

NOVA Laboratory for Computer Science and Informatics (NOVA LINCS)
Universidade Nova de Lisboa, Portugal

Abstract. Computational morality is an interdisciplinary field emerging from the need of imbuing autonomous agents with the capacity for moral decision-making. This paper supplies a computational model, via Logic Programming (LP), of counterfactual reasoning of autonomous agents with application to morality. Counterfactuals are conjectures about what would have happened, had an alternative event occurred. The first contribution of the paper is showing how counterfactual reasoning is modeled using LP, benefiting from LP abduction and updating. The approach is inspired by Pearl's structural causal model of counterfactuals, where causal direction and conditional reasoning are captured by inferential arrows of rules in LP. Herein, LP abduction hypothesizes background conditions from given evidences or observations, whereas LP updates frame these background conditions as a counterfactual's context, and then imposes causal interventions on the program through defeasible LP rules. In the second contribution, counterfactuals are applied to agent morality, resorting to this LP-based approach. We demonstrate its potential for specifying and querying moral issues, by examining viewpoints on moral permissibility via classic moral principles and examples taken from the literature. Application results were validated on a prototype implementing the approach on top of an integrated LP abduction and updating system supporting tabling.

Keywords: counterfactuals, abduction, logic programming, non-monotonic reasoning, morality.

1 Introduction

Computational morality [3, 45] is becoming a pressing concern, as agents become ever more sophisticated, autonomous, and act in groups, amidst populations of other agents. Accordingly, computational morality is a burgeoning field that emerges from the need of imbuing autonomous agents with the capacity of moral decision making to enable them to function in an ethically responsible manner via their own ethical decisions. It has attracted the artificial intelligence community, and brought together perspectives from various fields: philosophy, anthropology, cognitive science, neuroscience, and evolutionary biology. The overall result of this interdisciplinary research is not just important

* Email: lmp@fct.unl.pt
** Email: ar.saptawijaya@campus.fct.unl.pt. Affiliated with Faculty of Computer Science at Universitas Indonesia, Depok, Indonesia.
for equipping agents with some capacity for making moral judgments, but also to help better understand morality, via the creation and testing of computational models of ethical theories.

The potential of LP for computational morality has been reported in [16, 22, 33, 38], where the main characteristics of morality aspects can appropriately be expressed by LP-based reasoning, such as abduction, integrity constraints, preferences, updating, and argumentation. In this paper we focus on a computational model of counterfactual reasoning, via LP abduction and updating, for autonomous agents, with application to morality. The use of LP to model counterfactual reasoning therefore aims at taking counterfactuals to a wider context of the aforementioned well-developed LP-based non-monotonic reasoning methods that are all important, appropriate, and promising for moral reasoning.

Counterfactuals capture the process of reasoning about a past event that did not occur, namely what would have happened had this event occurred; or, vice-versa, to reason about an event that did occur but what if it had not. An example from [7]: Lightning hits a forest and a devastating forest fire breaks out. The forest was dry after a long hot summer and many acres were destroyed. One may think of a counterfactual about it, e.g., “if only there had not been lightning, then the forest fire would not have occurred”.

Counterfactuals have been widely studied, in philosophy [9, 15, 23], psychology [7, 11, 24, 25, 27, 35], and from the computational viewpoint [6, 13, 29, 30, 44]. In Pearl’s approach [29], counterfactuals are evaluated based on a probabilistic causal model and a calculus of intervention. Its main idea is to infer background circumstances that are conditional on current evidences, and subsequently to make a minimal required intervention in the current causal model, so as to comply with the antecedent condition of the counterfactual. The modified model serves as the basis for computing the counterfactual consequence’s probability.

Instead of defining a new formalism for counterfactual reasoning, we adopt here Pearl’s approach, abstaining from probabilities – given the lack of pure non-probabilistic counterfactual reasoning in LP – but resorting to LP abduction and updating. LP lends itself to Pearl’s causal model of counterfactuals: (1) The inferential arrow in a LP rule is adept at expressing causal direction; and (2) LP is enriched with functionalities, such as abduction and defeasible reasoning with updates. They can be exploited to establish the counterfactuals evaluation procedure of Pearl’s: LP abduction is employed for providing background conditions from observations made or evidences given, whereas defeasible logic rules allow achieving at points of adjustments to the current model via hypothetical updates of intervention. Our first contribution is to innovatively make use of LP abduction and updating in an implemented procedure for evaluating counterfactuals – taking Pearl’s established approach as reference – for cases when probabilities are not known or needed.

In a second contribution counterfactuals are applied to computational morality. Counterfactual theories are very suggestive of a conceptual relationship to a form of debugging, namely in view of correcting moral blame, since people ascribe abnormal antecedents an increased causal power, and are also more likely to generate counterfactuals concerning abnormal antecedents. Two distinct processes can be identified when people engage in counterfactual thinking. For one, its frequent spontaneous triggers en-
compass bad outcomes and “close calls” (some harm that was close to happening). Sec-
ond, such thinking comprises a process of finding antecedents which, if mutated, would
prevent the bad outcome from arising. When people employ counterfactual thinking,
they are especially prone to change abnormal antecedents, as opposed to normal ones.
Following a bad outcome, people are likely to conceive of the counterfactual “if only
[some abnormal thing] had not occurred, then the outcome would not have happened”.
See [35] for a review.

In this paper, counterfactuals are specifically engaged to distinguish whether an
effect of an action is a cause for achieving a morally dilemmatic goal or merely a side-
effect of that action. The distinction is essential for establishing moral permissibility
from the viewpoints of the Doctrines of Double Effect and of Triple Effect, as scruti-
nized herein through several off-the-shelf classic moral examples from the literature.
Note that, the application of counterfactuals in these examples neither aims at defend-
ring the two doctrines nor resolving the dilemmas appearing in the examples, as even
philosophers split over opinions on them. Instead, its purpose is to show that coun-
terfactuals, supported by our LP approach, are capable and appropriate for expressing
different viewpoints on permissibility according to both doctrines based on views ar-
gued in the literature. By materializing these doctrines in concrete moral dilemmas,
the results of counterfactual evaluation are readily comparable to those from the liter-
ature. Abstaining from probability does not hamper the application of the present LP
approach to computational morality, as moral reasoning in the literature does not in-
volve probability, but rather focuses on issues of definite right or wrong. Furthermore,
by abstaining from formal probability, we want to concentrate on the naturalized logic
of human moral reasoning as formulated by counterfactuals.

The paper is organized as follows. Section 2 reviews basic notation in LP, abduc-
tion in LP, and Pearl’s structure-based counterfactuals. We discuss how causation and
intervention in Pearl’s approach can be expressed in LP, and subsequently detail a LP
approach to evaluate counterfactuals, in Section 3. The application of counterfactuals
to computational morality is elaborated in Section 4. Section 5 frames our contributions
in the context of related work. We conclude in Section 6, by touching upon prospects of
counterfactuals in expressing moral issues, thereby opening up further opportunities of
application in computational morality, within a combination of abduction and updating.

2 Preliminaries

A literal is either an atom $B$ or its default negation $\text{not } B$, named positive and negative
literals, resp. They are negation complements to each other, $L'$ denoting the complement
of literal $L$. The atoms $\text{true}$ and $\text{false}$ are true and false, resp., in every interpretation. A
logic program is a set of rules of the form $H \leftarrow B$, where its head $H$ is an atom and its
(finite) body $B$ is a sequence of literals. When $B$ is empty, the rule is called a fact and
simply written $H$. A rule in the form of a denial, i.e., with $\text{false}$ as head, is an integrity
constraint (IC).
2.1 Abductive Logic Programs

Abduction is a reasoning method where one chooses from available hypotheses those that best explain the observed evidence, in a preferred sense. In LP, an abductive hypothesis (abducible) is a 2-valued positive literal $Ab$ or negation complement $Ab^*$ (denotes not $Ab$), with no rules, whose truth value is not initially assumed. An abductive logic program (ALP) is one allowing abducibles in the body of rules. An ALP framework $\langle P, A, I \rangle$ consists of program $P$, set $A$ of abducibles, and set $I$ of integrity constraints [20]. An observation $O$ is a set of literals, analogous to a query in ALP.

Abduction in LP can be accomplished by a top-down query-oriented procedure for finding a query solution by need. The solution’s abducibles are leaves in its procedural query-rooted call-graph, i.e., the graph recursively engendered by the procedure calls from literals in bodies of rules to heads of rules, and thence to the literals in a rule’s body. The correctness of this top-down computation requires the underlying semantics to be relevant, as it avoids computing a whole model (to warrant its existence) in finding an answer to a query. Instead, it suffices to use only the rules relevant to the query – those in its procedural call-graph – to find its truth value. The 3-valued Well-Founded Semantics or WFS [43] enjoys the relevancy property, i.e., it permits finding only relevant abducibles and their truth value via the aforementioned top-down query-oriented procedure. Those abducibles not mentioned in the solution are indifferent to the query. WFS guarantees the existence of a unique model for every program, the Well-Founded Model (WFM), which can be computed as the least fixpoint of an appropriate semantic operator [34]. For the above reason, our approach is based on WFS. Moreover, the XSB Prolog system [41], wherein the prototype is implemented, computes the WFM. Our approach is adaptable to other semantics as well, e.g., Weak Completion Semantics (WCS) [19] is employed in [4].

Herein the WFM of program $P$ is denoted by $WFM(P)$.

**Definition 1.** An IC is satisfied in $WFM(P)$ iff it is true or undefined in $WFM(P)$, i.e., its body is either false or undefined [31].

**Definition 2.** The logical consequence relation $\models$ is defined such that for formula $F$, $P \models F$ iff $F$ is true in $WFM(P)$.

**Definition 3.** Given an ALP framework $\langle P, A, I \rangle$ and an observation $O$, a consistent abductive solution $E \subseteq A$ is an explanation to observation $O$ iff $P \cup E \models O$ and $I$ is satisfied in $WFM(P \cup E)$.

2.2 Pearl’s Structure-based Counterfactuals

Pearl [29] proposes a structural theory of counterfactuals based on a probabilistic causal model and a calculus of intervention (viz., do-calculus). A causal model $M$ consists of two sets of variables $U$ and $V$, and a set $F$ of functions that decides how values are assigned to each variable $V_i \in V$. The variables in $U$ are background knowledge that have no explanatory mechanism encoded in model $M$. The values of all variables in $V$ are uniquely determined by every instantiation $U = u$ of the background knowledge.
Procedure 1. Given evidence $e$, the probability of the counterfactual sentence “$Y$ would be $y$ had $X$ been $x$” can be evaluated in a three-step process:

1. **Abduction**: Update the probability $P(u)$ by the evidence $e$ to obtain $P(u | e)$. This step explains the past circumstance $U = u$ in the presence of evidence $e$.
2. **Action**: Modify $M$ by the action $do(X = x)$. This step minimally adjusts model $M$ by a hypothetical intervention via the external action $do(X = x)$ to comply with the antecedent condition of the counterfactual.
3. **Prediction**: Compute the probability $Y = y$ in the modified model. In this step the consequence of the counterfactual is predicted based on the evidential understanding of the past (Step 1), and the hypothetical modification performed in Step 2.

In summary, the approach determines the probability of the counterfactual’s consequence $Y = y$ by performing an intervention to impose the counterfactual’s antecedent $X = x$ (other things being equal), given evidence $e$ about $U = u$.

3 LP-based Counterfactuals

Our LP approach is based on an existing procedure of counterfactuals evaluation, rather than being developed from scratch. Inspired by Pearl’s approach with its well-accepted epistemic foundations, we apply the idea of Pearl’s three-step procedure to logic programs, but leaving out probabilities, employing instead LP abduction and updating to determine the logic validity of counterfactuals.

Two important ingredients in Pearl’s approach of counterfactuals are causal model and intervention. Causation denotes a specific relation of cause and effect. Causation can be captured by LP rules, where the inferential arrow in a logic rule represents causal direction. LP abduction is thus appropriate for inferring causation, providing an explanation to a given observation. That said, LP abduction is not immediately sufficient for counterfactuals. Consider a simple logic program $P = \{ b \leftarrow a \}$. Whereas abduction permits obtaining explanation $a$ to observation $b$, the evaluation of counterfactual “if $a$ had not been true, then $b$ would not have been true” cannot immediately be evaluated from the conditional rule $b \leftarrow a$, for if its antecedent is false the counterfactual would be trivially true. That justifies the need for an intervention. That is, it requires explicitly imposing the desired truth value of $a$, and subsequently checking whether the predicted truth value of $b$ consistently follows from this intervention. As described in Pearl’s approach, such an intervention establishes a required adjustment, so as to ensure that the counterfactual’s antecedent be met. It permits the value of the antecedent to differ from its actual one, whilst maintaining the consistency of the modified model. We resort to LP abduction and updating to express causal source and intervention, resp.

3.1 Causal Model and LP Abduction

With respect to an ALP framework $\langle P, A, I \rangle$, observation $O$ corresponds to Pearl’s definition for evidence $e$. That is, $O$ has rules concluding it in program $P$, and hence does
not belong to the set of abducibles $A$. Recall that in Pearl’s approach, a model $M$ consists of set $U$ of background variables, whose values are conditional on case-considered observed evidences. These background variables are not causally explained in $M$, as they have no parent nodes in the causal diagram of $M$. In terms of LP abduction, they correspond to a set of abducibles $E \subseteq A$ that provide abductive explanations to observation $O$. Indeed, these abducibles have no preceding causal explanatory mechanism, as they have no rules concluding them in the program. In a nutshell, an ALP framework $\langle P, A, I \rangle$ that provides an abduced explanation $E \subseteq A$ to the available observation $O$ mirrors Pearl’s model $M$ with its specific $U$ supporting an explanation to the current observed evidence $e$.

### 3.2 Intervention and LP Updating

Besides abduction, our approach also benefits from LP updating, which is supported by well-established theory and properties, cf. [1, 2]. It allows a program to be updated by asserting or retracting rules, thus changing the state of the program. LP updating is therefore appropriate for representing changes and dealing with incomplete information.

The specific role of LP updating in our approach is twofold: (1) updating the program with the preferred explanation to the current observation, thus fixing in the program the initial abduced background context of the counterfactual being evaluated; (2) facilitating an apposite adjustment to the causal model by hypothetical updates of causal intervention on the program, affecting defeasible rules. Both roles are sufficiently accomplished by fluent (i.e., state-dependent literal) updates, rather than rule updates. In the first role, explanations are treated as fluents. In the second, reserved predicates are introduced as fluents for the purpose of intervention upon defeasible rules. For the latter role, fluent updates are particularly more appropriate than rule updates (e.g., intervention by retracting rules), because intervention is hypothetical only. Removing away rules from the program would be an overkill, as the rules might be needed to elaborate justifications and introspective debugging.

### 3.3 Evaluating Counterfactuals in LP

The procedure to evaluate counterfactuals in LP essentially takes the three-step process of Pearl’s approach as its reference. That is, each step in the LP approach captures the same idea of its corresponding step in Pearl’s.

In what follows, counterfactuals are distinguished from semifactuals [7], as the LP procedure for the former is slightly different from the latter. The procedure for semifactuals will be discussed later in this section.

The key idea of evaluating counterfactuals with respect to an ALP framework, at some current state (discrete time) $T$, is as follows. In step 1, abduction is performed to explain the factual observation.\footnote{We assume that people are using counterfactuals to convey truly relevant information rather than to fabricate arbitrary subjunctive conditionals (e.g., “If I had been watching, then I would have seen the cheese on the moon melt during the eclipse”). Otherwise, implicit observations} The observation corresponds to the evidence that both
the antecedent and the consequence literals of the present counterfactual were factually false. There can be multiple explanations available to an observation; choosing a suitable one among them is a pragmatic issue, which can be dealt with preferences or ICs. The explanation fixes the abduced context in which the counterfactual is evaluated by updating the program with the explanation.

In step 2, defeasible rules are introduced for atoms forming the antecedent of the counterfactual. Given the past event \( E \), that renders its corresponding antecedent literal false, held at factual state \( T_E < T \), its causal intervention is realized by a hypothetical update \( H \), at state \( T_H = T_E + \Delta_H \), such that \( T_E < T_H < T_E + 1 \leq T \). That is, a hypothetical update strictly takes place between two factual states, thus \( 0 < \Delta_H < 1 \).

In the presence of defeasible rules this update permits hypothetical modification of the program to consistently comply with the antecedent of the counterfactual.

In step 3, the WFM of the hypothetical modified program is examined to verify whether the consequence of the counterfactual holds true at state \( T \). One can easily reinstate to the current factual situation by canceling the hypothetical update, e.g., via a new update of \( H \)’s complement at state \( T_F = T_H + \Delta_F \), such that \( T_H < T_F < T_E + 1 \).

Based on these ideas and analogously to the three-step process of Pearl’s, our approach is defined below, abstracting from the above state transition detail (cf. Section 3.5 for a concrete discussion of this state transition).

Procedure 2. Let \((P, A, I)\) be an ALP framework, where program \( P \) encodes the modeled situation on which counterfactuals are evaluated. Consider a counterfactual “if \( \text{Pre} \) had been true, then \( \text{Conc} \) would have been true”, where \( \text{Pre} \) and \( \text{Conc} \) are finite conjunctions of literals.

1. Abduction: Compute an explanation \( E \subseteq A \) to the observation \( O = O_{\text{Pre}} \cup O_{\text{Conc}} \cup O_{\text{Oth}} \), where \( O_{\text{Pre}} = \{ L_i^c \mid L_i \text{ is in } \text{Pre} \} \), \( O_{\text{Conc}} = \{ L_i^c \mid L_i \text{ is in } \text{Conc} \} \), and \( O_{\text{Oth}} \) is other (possibly empty) observations: \( O_{\text{Oth}} \cap (O_{\text{Pre}} \cup O_{\text{Conc}}) = \emptyset \).

   Update program \( P \) with \( E \), obtaining program \( P \cup E \).

2. Action: For each literal \( L \) in conjunction \( \text{Pre} \), introduce a pair of reserved meta-predicates \( \text{make}(B) \) and \( \text{make_not}(B) \), where \( B \) is the atom in \( L \). These two meta-predicates are introduced for the purpose of establishing causal intervention: they are used to express hypothetical alternative events to be imposed. This step comprises two stages:

   (a) Transformation:
   - Add rule \( B \leftarrow \text{make}(B) \) to program \( P \cup E \).
   - Add \( \text{not make_not}(B) \) to the body of each rule in \( P \) whose head is \( B \). If there is no such rule, add rule \( B \leftarrow \text{not make_not}(B) \) to program \( P \cup E \).

   Let \( (P \cup E)_r \) be the resulting transform.

   (b) Intervention: Update program \( (P \cup E)_r \), with literal \( \text{make}(B) \) or \( \text{make_not}(B) \), for \( L = B \) or \( L = \text{not } B \), resp. Assuming that \( \text{Pre} \) is consistent, \( \text{make}(B) \) and \( \text{make_not}(B) \) cannot be imposed at the same time.

must simply be made explicit observations, to avoid natural language conundrums or ambiguities [14].

2 This interpretation is in line with the corresponding English construct, cf. [18], commonly known as third conditionals.
Let \((P \cup E)_{\tau,1}\) be the program obtained after these hypothetical updates of intervention.

3. **Prediction:** Verify whether \((P \cup E)_{\tau,1} \models Conc\) and \(I\) is satisfied in \(WFM((P \cup E)_{\tau,1})\).

This three-step procedure defines valid counterfactuals.

**Definition 4.** The counterfactual "if Pre had been true, then Conc would have been true" is valid given observation \(O = O_{Pre} \cup O_{Conc} \cup O_{Oth}\) iff \(O\) is explained by \(E \subseteq A\), \((P \cup E)_{\tau,1} \models Conc\), and \(I\) is satisfied in \(WFM((P \cup E)_{\tau,1})\).

Since WFS supports top-down query-oriented procedures for finding solutions (cf. Section 2), checking validity of counterfactuals, i.e., whether their conclusion \(Conc\) follows (step 3), given the intervened program transform (step 2) with respect to the abduced background context (step 1), in fact amounts to checking in a derivation tree whether query \(Conc\) holds true while also satisfying \(I\).

**Example 1.** Recall the example in the introduction. Let us slightly complicate it by having two alternative abductive causes for the forest fire, viz., storm (which implies lightning hitting the ground) or barbecue. Storm is accompanied by strong wind that causes the dry leaves falling onto the ground. Note that dry leaves are important for forest fire in both cases. This example is expressed by ALP framework \(\langle P, A, I \rangle\), using abbreviations \(b, d, f, g, l, s\) for barbecue, dry leaves, forest fire, leaves on the ground, lightning, and storm, resp., where \(A = \{s, b, s^*, b^*\}\), \(I = \emptyset\), and \(P\) as follows:

\[ f \leftarrow b, d. \quad f \leftarrow b^*, l, d, g. \quad l \leftarrow s. \quad g \leftarrow s. \quad d. \]

The use of \(b^*\) in the second rule of \(f\) is intended so as to have mutual exclusive explanations.

Consider counterfactual "if only there had not been lightning, then the forest fire would not have occurred", where \(Pre = not l\) and \(Conc = not f\).

1. **Abduction:** Besides \(O_{Pre} = \{l\}\) and \(O_{Conc} = \{f\}\), say that \(g\) is observed too: \(O_{Oth} = \{g\}\). Given \(O = O_{Pre} \cup O_{Conc} \cup O_{Oth}\), there are two possible explanations: \(E_1 = \{s, b^*\}\) and \(E_2 = \{s, b\}\). Consider a scenario where the minimal explanation \(E_1\) (in the sense of minimal positive literals) is preferred to update \(P\), to obtain \(P \cup E_1\). Note, program \(P \cup E_1\) corresponds to a state with \(WFM(P \cup E_1) = \{d, s, g, l, f, not b\}\). This updated program reflects the evaluation context of the counterfactual, where all literals of \(Pre\) and \(Conc\) were false in the initial factual situation.

2. **Action:** The transformation results in program \((P \cup E_1)_{\tau}\):

\[ f \leftarrow b, d. \quad f \leftarrow b^*, l, d, g. \quad g \leftarrow s. \quad d. \]

\[ l \leftarrow \text{make}(l). \quad l \leftarrow s, not \text{make_not}(l). \]

Program \((P \cup E_1)_{\tau}\) is updated with \(\text{make_not}(l)\) as the required intervention. It engenders program \((P \cup E_1)_{\tau,1}\) corresponding to a new state with \(WFM((P \cup E_1)_{\tau,1}) = \{d, s, g, \text{make_not}(l), not \text{make}(l), not b, not l, not f\}\).

3. **Prediction:** We verify that \((P \cup E_1)_{\tau,1} \models not f\), and \(I = \emptyset\) is trivially satisfied in \(WFM((P \cup E_1)_{\tau,1})\).
We thus conclude that, for this $E_1$ scenario, the given counterfactual is valid.

**Example 2.** In the other explanatory scenario of Example 1, where $E_2$ (instead of $E_1$) is preferred to update $P$, the counterfactual is no longer valid, because $WFM((P \cup E_2)_{\tau,\iota}) = \{d, s, g, b, make, not(l), not make(l), not l, f\}$, and thus $(P \cup E_2)_{\tau,\iota} \not\models not f$. Indeed, the forest fire would still have occurred but due to an alternative cause: barbecue. Skeptical and credulous counterfactual evaluations could ergo be defined, i.e., by evaluating the presented counterfactual for each abduced background context. Given that step 2 can be accomplished by a one-time transformation, such skeptical and credulous counterfactual evaluations require only executing step 3 for each background context fixed in step 1.

**Semifactuals Reasoning** Another form related to counterfactuals is semifactuals, i.e., one that combines a counterfactual antecedent and an enduring factual consequence [7], with a typical form of statement “Even if . . . “. Other comparable linguistic constructs also exist, e.g., “No matter if . . . “, “Though . . . “, etc. The LP procedure for counterfactuals (Procedure 2) can easily be adapted to evaluating semifactuals. Like in counterfactuals, the antecedent of a semifactual is supposed false in the factual situation. But different from counterfactuals, the consequence of a semifactual should instead be factually ensured true (rather than false).

Consider semifactual “even if $Pre$ had been true, $Conc$ would still have been true”. Its LP evaluation follows Procedure 2 with the only modification on the definition of $O_{Conc}$ in Step 1, i.e., for semifactuals, $O_{Conc}$ is defined as $O_{Conc} = \{L_i | L_i \text{ is in } Conc\}$, to warrant its consequence factually true. The validity condition for semifactuals is the same as for counterfactuals, cf. Definition 4.

**Example 3.** Recall Example 2, where $E_2 = \{s, b\}$ is preferred. Consider semifactual “even if there had not been lightning, the forest fire would still have occurred”, where $Pre = not l$ and $Conc = f$. This semifactual is valid, because given the same $WFM((P \cup E_2)_{\tau,\iota})$ as in Example 2, we now have $(P \cup E_2)_{\tau,\iota} \models Conc$, i.e., $(P \cup E_2)_{\tau,\iota} \models f$.

### 3.4 Properties

Since the idea of each step in the LP approach mirrors the one corresponding in Pearl’s, the LP approach therefore immediately compares to Pearl’s, its epistemic adequacy and properties relying on those of Pearl’s.

In [23], salient logic properties of counterfactuals are argued, including three counterfactual fallacies that distinguish counterfactual conditional from the material one. Our approach also satisfies these three properties, as shown below. Other counterfactual properties, such as reflexive, modus tollens, disjunction in the antecedent, combination of sentences, etc., are left for future work; ascertaining their satisfaction is not in the purview of the present paper.

Let $Pre \models Conc$ represent counterfactual statement “if $Pre$ had been true, then $Conc$ would have been true”.


\textit{Property 1.} Fallacy of strengthening the antecedent:

\[ A \implies B \text{ does not imply } A \land C \implies B. \]

\textit{Example 4.} Recall Example 1, where \( E_1 \) is adopted. We have shown that counterfactual \( \text{not } l \implies \text{not } f \) is valid. Let us strengthen its antecedent with “there had been a barbecue”, obtaining counterfactual \( \text{not } l \land \text{b} \implies \text{not } f \). For this new counterfactual, \( O_{prc} = \{l, \text{not } b\} \), whereas the other observations \( O_{conc} \) and \( O_{oth} \) are the same as in Example 1. The only explanation of \( O = O_{prc} \cup O_{conc} \cup O_{oth} \) is \( E_{prc} = \{s, b^*\} \). The transform \( (P \cup E_{prc})_{\tau,\iota} \) is as follows:

\[
\begin{align*}
f &\leftarrow b, d, f \leftarrow b^*, l, d, g, g \leftarrow s, d, \\
l &\leftarrow \text{make}(l), l \leftarrow s, \text{not make}(l), \\
b &\leftarrow \text{make}(b), b \leftarrow \text{not make}(b).
\end{align*}
\]

The required interventions \( \text{make}(l) \) and \( \text{make}(b) \) update this program, obtaining \( (P \cup E_{prc})_{\tau,\iota} \) with \( WF_{\tau}(P \cup E_{prc})_{\tau,\iota} = \{d, s, g, \text{make}(l), \text{not make}(l), \text{not make}(b), b, f\} \). Observe that, intervention \( \text{make}(b) \) hypothetically updates the truth value of \( b \) from false (in \( P \cup E_{prc} \)) to true (in \( (P \cup E_{prc})_{\tau,\iota} \)). Since \( (P \cup E_{prc})_{\tau,\iota} \not\models \text{not } f \), counterfactual \( \text{not } l \implies \text{not } b \implies \text{not } f \) is not valid.

\textit{Property 2.} Fallacy of contraposition:

\[ A \implies B \text{ does not imply } \text{not } B \implies \text{not } A. \]

\textit{Example 5.} Recall the ALP framework of Example 1, but with \( O_{oth} = \emptyset \), rendering \( O = \{l, f\} \), and thus the two explanations \( E_1 \) and \( E_2 \) do not change. Therefore, \( E_3 \) being preferred, the counterfactual \( \text{not } l \implies \text{not } f \) is valid, as shown in that example. Consider its contraposition: \( f \implies l \). Its corresponding observation \( O_{prc} = \{\text{not } f, \text{not } l\} \) admits a single explanation \( E_{prc} = \{s, b^*\} \). Though intervention \( \text{make}(f) \) is imposed on the transform \( (P \cup E_{prc})_{\tau,\iota} \), we obtain \( \text{not } l \in WF_{\tau}(P \cup E_{prc})_{\tau,\iota} \). Thus, \( (P \cup E_{prc})_{\tau,\iota} \not\models \text{not } f \), and \( f \implies l \) is not a valid counterfactual.

\textit{Property 3.} Fallacy of transitivity:

\[ A \implies B \text{ and } B \implies C \text{ do not imply } A \implies C. \]

\textit{Example 6.} Let \( P_1 \) be program:

\[
\text{marry} \leftarrow \text{pregnant. criticized} \leftarrow \text{not marry, pregnant.}
\]

in the ALP framework \( (P_1, A_t, I_t) \), using abbreviation \( c, m, p \) for \( \text{criticized, marry, and pregnant, resp.} \), where \( A_t = \{p, p^*\} \) and \( I_t = \emptyset \). Consider counterfactuals \( C_1, C_2, C_3; (C_1) \not\models p \not\models m; (C_2) \not\models m \implies c; \text{and } (C_3) \not\models p \implies c. \)

We can verify that \( C_1 \) is valid, given the only explanation \( E_{C_1} = \{p\} \) to the observation \( O_{C_1} = \{p, m\} \) and the intervention \( \text{make}(p) \), since \( \not\models m \in WF_{\tau}(P \cup E_{C_1})_{\tau,\iota} \), and thus \( (P \cup E_{C_1})_{\tau,\iota} \models \not\models m \). We can similarly verify that \( C_2 \) is valid, since \( (P \cup E_{C_2})_{\tau,\iota} \models c \), where \( E_{C_2} = \{p\} \) explains \( O_{C_2} = \{m, \not\models c\} \), and the imposed intervention is \( \text{make}(m) \). But \( C_3 \) is not valid, since \( (P \cup E_{C_3})_{\tau,\iota} \not\models c \), given that \( \not\models c \in WF_{\tau}(P \cup E_{C_3})_{\tau,\iota} \), where \( E_{C_3} = \{p\} \) explains \( O_{C_3} = \{p, \not\models c\} \) with intervention \( \text{make}(p) \).
3.5 Implementation

We have developed a prototype, QUALM, that implements the procedure on top of an existing integrated system of LP abduction and updating supporting tabling, based on [37], in XSB Prolog [41]. QUALM allows specifying predicates that are subject to intervention, e.g., predicate \( l \) in Example 1. This information is useful for the transformation stage, in step 2 of the procedure.

In QUALM, the state transition of the program, as a consequence of program updating (asserting or retracting fluents for our case), is facilitated by timestamps that are internally managed. By convention the program is initially inserted at state \( T = 1 \). The state subsequently progresses to \( T = 2 \).

Observations are explained by posing a top-level query, e.g., \( \text{?- query}((l, f, g), E_1) \) provides explanation \( E_1 \) to the observation \( O = \{l, f, g\} \) of Example 1. Thanks to WFS that underlies XSB, QUALM enjoys the relevancy property (cf. Section 2.1) in computing explanations to observations. In order to fix \( E_1 = \{s, b^*\} \) as the abduced context in evaluating counterfactual at the present state \( T = 2 \), both fluents \( s \) and \( b^* \), that held at the factual state \( T_{E_1} = 1 \), are asserted (via QUALM’s reserved predicate \( \text{updates}(L) \) to assert fluents in list \( L \)). These updates render them as facts in the updated program \( P \cup E_1 \).

A causal intervention “there had not been lightning” is enacted by the hypothetical update of fluent \( \text{make, not}(l) \) via query \( \text{?- updates([make, not(l)])} \). As described in Section 3.3, this update strictly takes place between two consecutive factual states; in this case between \( T_{E_1} = 1 \) and the current state \( T = 2 \). QUALM internally assigns a fraction of timestamp, say 0.01, just after \( T_{E_1} \), viz., the hypothetical update \( \text{make, not}(l) \) is imposed at state \( T_H = 1.01 \). It thus simulates an intervention via an update in the past, while keeping the present state at \( T = 2 \). After this update, the validity of the present counterfactual (at \( T = 2 \)) can be checked by testing its conclusion, e.g., \( \text{?- query}(f, E) \) to query whether forest fire would have occurred after the hypothetical update. QUALM answers ‘no’, verifying the counterfactual’s validity that the forest fire would not have occurred.

To reinstate the current factual situation from a counterfactual mode, a hypothetical update can be canceled by updating the program with its fluent complement, e.g., \( \text{?- updates([not make, not(l)])} \), occurring at a fraction of time after \( T_H \) (also internally assigned by QUALM), e.g., at \( T_F = T_H + 0.01 = 1.02 \). It thus supervenes the hypothetical update \( \text{make, not}(l) \) that was enacted at \( T_H = 1.01 \), which is equivalent to retracting it. Consequently, the intervention is no longer imposed on the program.

4 Counterfactuals in Morality

People typically reason about what they should or should not have done when they examine decisions in moral situations. It is therefore natural for them to engage counterfactual thoughts in such settings. Counterfactual thinking has been investigated in the context of moral reasoning, notably by psychology experimental studies, e.g., to understand the kind of counterfactual alternatives people tend to imagine in contemplating moral behaviors [25] and the influence of counterfactual thoughts in moral judgment [27].
Morality and normality judgments typically correlate. Normality mediates morality with causation and blame judgments. The controllability in counterfactuals mediates between normality, blame and cause judgments. The importance of control, namely the possibility of counterfactual intervention, is highlighted in theories of blame that presume someone responsible only if they had some control of the outcome [46].

As argued in [11], the function of counterfactual thinking is not just limited to the evaluation process, but occurs also in the reflection one. Through evaluation, counterfactuals help correct wrong behavior in the past, thus guiding future moral decisions. Reflection, on the other hand, permits momentary experiential simulation of possible alternatives, thereby allowing careful consideration before a moral decision is made, and to subsequently justify it.

The potential of LP to computational morality has been reported in [16, 22, 33] and with emphasis on LP abduction and updating in [38]. Here we investigate how moral issues can innovatively be expressed with counterfactual reasoning by resorting to a LP approach. We particularly look into its application for examining viewpoints on moral permissibility, exemplified by classic moral dilemmas from the literature on the Doctrines of Double Effect (DDE) [26] and of Triple Effect (DTE) [21].

DDE is often invoked to explain the permissibility of an action that causes a harm by distinguishing whether this harm is a mere side-effect of bringing about a good result, or rather a means to bringing about the same good end [26]. In [17], DDE has been utilized to explain the consistency of judgments, shared by subjects from demographically diverse populations, on a series of moral dilemmas.

Counterfactuals may provide a general way to examine DDE in dilemmas, e.g., the classic trolley problem [12], by distinguishing between a cause and a side-effect as a result of performing an action to achieve a goal. This distinction between causes and side-effects may explain the permissibility of an action in accordance with DDE. That is, if some morally wrong effect \( E \) happens to be a cause for a goal \( G \) that one wants to achieve by performing an action \( A \), and not a mere side-effect of \( A \), then performing \( A \) is impermissible. This is expressed by the counterfactual form below, in a setting where action \( A \) is performed to achieve goal \( G \):

\[
\text{If not } E \text{ had been true, then not } G \text{ would have been true.}
\]

The evaluation of this counterfactual form identifies permissibility of action \( A \) from its effect \( E \), by identifying whether the latter is a necessary cause for goal \( G \) or a mere side-effect of action \( A \). That is, if the counterfactual proves valid, then \( E \) is instrumental as a cause of \( G \), and not a mere side-effect of action \( A \). Since \( E \) is morally wrong, achieving \( G \) that way, by means of \( A \), is impermissible; otherwise, not. Note that the evaluation of counterfactuals in this application is considered from the perspective of agents who perform the action, rather than from others’ (e.g., observers). Moreover, our emphasis on causation in this application focuses on agents’ deliberate actions, rather than on causation and counterfactuals in general. See [9, 29] for a more general and broad discussion on causation and counterfactuals.

We exemplify an application of this counterfactual form in two off-the-shelf military cases from [40] – abbreviations in parentheses: terror bombing (\( \text{teb} \)) vs. tactical bombing (\( \text{tab} \)). The former refers to bombing a civilian target (\( \text{civ} \)) during a war, thus killing civilians (\( \text{kic} \)), in order to terrorize the enemy (\( \text{ror} \)), and thereby get them to
end the war (ew). The latter case is attributed to bombing a military target (mil), which will effectively end the war (ew), but with the foreseen consequence of killing the same number of civilians (kic) nearby. According to DDE, terror bombing fails permissibility due to a deliberate element of killing civilians to achieve the goal of ending the war, whereas tactical bombing is accepted as permissible.

**Example 7.** We first model terror bombing with ew as the goal, by considering the ALP framework \( \langle P_e, A_e, I_e \rangle \), where \( A_e = \{ \text{teb}, \text{teb}^* \} \), \( I_e = \emptyset \) and \( P_e \):

\[
\text{ew} \leftarrow \text{ror}. \quad \text{ror} \leftarrow \text{kic}. \quad \text{kic} \leftarrow \text{civ}. \quad \text{civ} \leftarrow \text{teb}.
\]

We consider counterfactual “if civilians had not been killed, then the war would not have ended”, where \( \text{Pre} = \text{not kic} \) and \( \text{Conc} = \text{not ew} \). The observation \( O = \{ \text{kic}, \text{ew} \} \), with \( O_{\text{teb}} \) being empty, has a single explanation \( E_e = \{ \text{teb} \} \). The rule \( \text{kic} \leftarrow \text{civ} \) transforms into \( \text{kic} \leftarrow \text{civ}, \text{not make not(kic)} \). Given intervention \( \text{make not(kic)} \), the counterfactual is valid, because \( \text{not ew} \in \text{WFM}(\langle P_e \cup E_e \rangle) \). Hence \( \text{teb} \) is DDE morally impermissible.

**Example 8.** Tactical bombing with the same goal ew can be modeled by the ALP framework \( \langle P_a, A_a, I_a \rangle \), where \( A_a = \{ \text{tab}, \text{tab}^* \} \), \( I_a = \emptyset \) and \( P_a \):

\[
\text{ew} \leftarrow \text{mil}. \quad \text{mil} \leftarrow \text{tab}. \quad \text{kic} \leftarrow \text{tab}.
\]

In this tactical bombing example, we could alternatively employ a semifactual: “even if civilians had not been killed, the war would still have ended”. It will be interesting to explore in the future the applicability of semifacts to computational morality, to identify indifferent actions.

**Example 9.** Consider two countries \( a \) and its ally, \( b \), that concert a terror bombing, modeled by the ALP framework \( \langle P_{ab}, A_{ab}, I_{ab} \rangle \), where \( A_{ab} = \{ \text{teb}, \text{teb}^* \} \), \( I_{ab} = \emptyset \) and \( P_{ab} \) below. The abbreviations \( \text{kic}(X) \) and \( \text{civ}(X) \) refer to ‘killing civilians by country \( X \)’ and ‘bombing a civilian target by country \( X \)’. As usual in LP, underscore (_) represents an anonymous variable.

\[
\text{ew} \leftarrow \text{ror}. \quad \text{ror} \leftarrow \text{kic}(\_). \\
\text{kic}(X) \leftarrow \text{civ}(X). \quad \text{civ}(\_) \leftarrow \text{teb}.
\]

Being represented as a single program (rather than a separate knowledge base for each agent), this scenario should appropriately be viewed as if a joint action performed by a single agent. Therefore, the counterfactual of interest is “if civilians had not been killed by \( a \) and \( b \), then the war would not have ended”. That is, the antecedent of the counterfactual is a conjunction: \( \text{Pre} = \text{not kic}(a) \land \text{not kic}(b) \). One can easily verify that \( \text{not ew} \in \text{WFM}(\langle P_{ab} \cup E_{ab} \rangle) \). Thus, \( \langle P_{ab} \cup E_{ab} \rangle \models \text{not ew} \) and the counterfactual is valid: the concerted \( \text{teb} \) is DDE impermissible.
This application of counterfactuals can be challenged by a more complex scenario, to distinguish moral permissibility according to DDE vs. DTE. DTE [21] refines DDE particularly on the notion about harming someone as an intended means. That is, DTE distinguishes further between doing an action in order that an effect occurs and doing it because that effect will occur. The latter is a new category of action, which is not accounted for in DDE. Though DTE also classifies the former as impermissible, it is more tolerant to the latter (the third effect), i.e., it treats as permissible those actions performed just because instrumental harm will occur. Kamm [21] proposed DTE to accommodate a variant of the trolley problem, viz., the Loop Case [42]: A trolley is headed toward five people walking on the track, and they will not be able to get off the track in time. The trolley can be redirected onto a side track, which loops back towards the five. A fat man sits on this looping side track, whose body will by itself stop the trolley. Is it morally permissible to divert the trolley to the looping side track, thereby hitting the man and killing him, but saving the five? This case strikes most moral philosophers that diverting the trolley is permissible [28]. Referring to a psychology study [17], 56% of its respondents judged that diverting the trolley in this case is also permissible. To this end, DTE may provide the justification, that it is permissible because it will hit the man, and not in order to intentionally hit him [21]. Nonetheless, DDE views diverting the trolley in the Loop case as impermissible.

We use counterfactuals to capture the distinct views of DDE and DTE in the Loop case.

Example 10. We model the Loop case with the ALP framework \((P_o, A_o, I_o)\), where sav, div, hit, tst, mst stand for save the five, divert the trolley, man hit by the trolley, train on the side track and man on the side track, resp., with sav as the goal, \(A_o = \{\text{div}, \text{div}^*\}\), \(I_o = \emptyset\), and \(P_o:\)

\[
\begin{align*}
\text{sav} & \leftarrow \text{hit}. \\
\text{hit} & \leftarrow \text{tst}, \text{mst}. \\
\text{tst} & \leftarrow \text{div}. \\
\text{mst}. 
\end{align*}
\]

DDE views diverting the trolley impermissible, because this action redirects the trolley onto the side track, thereby hitting the man. Consequently, it prevents the trolley from hitting the five. To come up with the impermissibility of this action, it is required to show the validity of the counterfactual “if the man had not been hit by the trolley, the five people would not have been saved”. Given observation \(O = O_{\text{pre}} \cup O_{\text{conc}} = \{\text{hit}, \text{sav}\}\), its only explanation is \(E_o = \{\text{div}\}\). Note that rule \(\text{hit} \leftarrow \text{tst}, \text{mst}, \text{not make}_\text{not}(\text{hit})\), and the required intervention is \(\text{make}_\text{not}(\text{hit})\). The counterfactual is therefore valid, because \(\text{not sav} \in \text{WFM}(\langle P_o \cup E_o \rangle_{\tau,i})\), hence \((P_o \cup E_o)_{\tau,i} \models \text{not sav}\). This means hit, as a consequence of action div, is instrumental as a cause of goal sav. Therefore, \text{div} is DDE morally impermissible.

DTE considers diverting the trolley as permissible, since the man is already on the side track, without any deliberate action performed in order to place him there. In \(P_o\), we have the fact \(\text{mst} \) ready, without abducting any ancillary action. The validity of the counterfactual “if the man had not been on the side track, then he would not have been hit by the trolley”, which can easily be verified, ensures that the unfortunate event of the man being hit by the trolley is indeed the consequence of the man being on the side track. The lack of deliberate action (exemplified here by pushing the man – \text{psh} for short) in order to place him on the side track, and whether the absence of this action still causes the unfortunate event (the third effect) is captured by the counterfactual “if
the man had not been pushed, then he would not have been hit by the trolley”. This counterfactual is not valid, because the observation \( O = O_{Pre} \cup O_{Conc} = \{psh, hit\} \) has no explanation \( E \subseteq A_o \), i.e., \( psh \notin A_o \), and no fact \( psh \) exists either. This means that even without this hypothetical but unexplained deliberate action of pushing, the man would still have been hit by the trolley (just because he is already on the side track). Though \( hit \) is a consequence of \( div \) and instrumental in achieving \( sav \), no deliberate action is required to cause \( mst \), in order for \( hit \) to occur. Hence \( div \) is DTE morally permissible.

Next, we consider a more involved trolley example.

**Example 11.** Consider a variant of the Loop case, viz., the *Loop-Push Case* (see also Extra Push Case in [21]). Differently from the Loop case, now the looping side track is initially empty, and besides the diverting action, an ancillary action of pushing a fat man in order to place him on the side track is additionally performed. This case is modeled by the ALP framework \( \langle P_p, A_p, I_p \rangle \), where \( A_p = \{div, psh, div^*, psh^*\} \), \( I_p = \emptyset \), and \( P_p:\)

\[
\begin{align*}
sav & \leftarrow \text{hit}. \\
\text{hit} & \leftarrow \text{tst, mst}. \\
\text{tst} & \leftarrow \text{div}. \\
mst & \leftarrow \text{psh}.
\end{align*}
\]

Recall the counterfactuals considered in the discussion of DDE and DTE of the Loop case:

- “If the man had not been hit by the trolley, the five people would not have been saved.” The same observation \( O = \{hit, sav\} \) provides an extended explanation \( E_{p_1} = \{div, psh\} \). That is, the pushing action needs to be abducted for having the man on the side track, so the trolley can be stopped by hitting him. The same intervention \( \text{make/not}(hit) \) is applied to the same transform, resulting in a valid counterfactual: \( (P_p \cup E_{p_1})_{\tau,i} \models \text{not} \ text{sav}, \text{because not} \ text{sav} \in WFM((P_p \cup E_{p_1})_{\tau,i}). \)

- “If the man had not been pushed, then he would not have been hit by the trolley.”

  The relevant observation is \( O = \{psh, hit\} \), explained by \( E_{p_2} = \{div, psh\} \). Whereas this counterfactual is not valid in DTE of the Loop case, it is valid in the Loop-Push case. Given rule \( psh \leftarrow \text{not make/not}(psh) \) in the transform and intervention \( \text{make/not}(psh) \), we verify that \( (P_p \cup E_{p_2})_{\tau,i} \models \text{not} \ text{hit, as not} \text{hit} \in WFM((P_p \cup E_{p_2})_{\tau,i}). \)

From the validity of these two counterfactuals it can be inferred that, given the diverting action, the ancillary action of pushing the man onto the side track causes him to be hit by the trolley, which in turn causes the five to be saved. In the Loop-Push, DTE agrees with DDE that such a deliberate action (pushing) performed in order to bring about harm (the man hit by the trolley), even for the purpose of a good or greater end (to save the five), is likewise impermissible.

## 5 Related Work

In Pearl’s approach, intervention is realized by superficial revision, by imposing the desired value to the intervened node and cutting it from its parent nodes. This is also the case in our approach, by means of hypothetical updates affecting defeasible rules. Other subtle ways of intervention may involve deep revision, which can be realized in
LP. It is beyond the scope of the paper, but amply discussed in [4]. Unlike Pearl’s, our approach is non-probabilistic, which corresponds to assigning probability to abductive explanations (or variables in $U$ of Pearl’s causal model) of 0 or 1.

A formalization of our procedure is reported in [4] – albeit based on different semantics (WCS vs. WFS) – along with some properties specific to our LP-based approach. The present paper complements [4] in the sense that we provide an implemented procedure employing our WFS-based LP abduction and updating, realized in our prototype QUALM. That is, it lays emphasis more on the LP engineering aspect for relating the role of LP abduction and updating to Pearl’s causal model and hypothetical intervention, and in realizing the procedure in QUALM. Moreover, this paper also shows how counterfactuals apply to examine morality issues, which is not touched at all in [4]. Due to the similarity of common features to WFS and WCS, the Propositions and Proofs in [4] can be transposed to the WFS setting, which we do not repeat here, given the distinct emphasizes just made salient about each of these two otherwise conceptually similar complementary approaches.3

LP abduction and revision are employed in [5] to evaluate indicative conditionals, but not counterfactual conditionals. LP abduction is employed through a rewrite system to find solutions for an abductive framework; the rewrite system intuitively captures the natural semantics of indicative conditionals. Rule revisions are additionally used to satisfy conditions whose truth-value is unknown and which cannot be explained by abduction.

In [30], counterfactuals are evaluated using contradiction removal semantics of LP. The work is based on Lewis’s counterfactuals [23], where a model of a logic program represents a world in Lewis’s concept. The semantics defines the most similar worlds by removing contradictions from the associated program, obtaining the so-called maximal non-contradictory submodels of the program. It does not concern itself with LP abduction and updating; both being relevant for our work, which is based on Pearl’s concept rather than Lewis’s, without the need of a world distance measure.

Probabilistic LP (PLP) language P-log with the Stable Model Semantics (SMS) is employed, in [6], to encode Pearl’s Probabilistic Causal Model (PCM), without involving abduction. It does not directly encode Pearl’s three-step process, but focuses on P-log probabilistic approach to compute the probability of a counterfactual query. Our work does not deal with probability, but logic, though it epistemically mirrors Pearl’s three-step process, via LP abduction and updating. Our approach is also not based on SMS, but instead on WFS with its relevancy property, which is more appropriate for LP abduction by need as argued earlier. In [44], Pearl’s PCM is encoded using PLP CP-logic, without involving abduction either. Whereas P-log has its own do-operator to achieve intervention in its probabilistic reasoning, CP-logic achieves it by eliminating rules. Similar to P-log, our approach introduces meta-predicates make and make_not.

3 Both WFS and WCS are 3-valued semantics that differ in dealing with close world assumption (CWA) and rules with positive loops (e.g., $p ← p$). WFS enforces CWA, i.e., atom $a$ that has no rule is interpreted as false, whereas in WCS undefined. Nevertheless, they can be transformed one to another: adding rules $a ← u$ and $u ← not u$ for a reserved atom $u$ renders $a$ unknown in WFS; alternatively, adding $a ← false$ enforces CWA in WCS. In this paper, positive loops are not needed and do not appear throughout examples we consider.
to accomplish intervention via defeasible rules and fluent updates, without eliminating rules, as CP-logic does.

The use of causation, based on structural approach, to define and model issues related to morality, such as blame and responsibility, is discussed in [8, 15]. The interest of the present work is to bring counterfactuals (rather than causation), inspired by the structural approach, into a wider context of LP-based non-monotonic reasoning, given the lack of pure non-probabilistic counterfactual reasoning in LP, and to foster the interplay of various LP-based reasoning for the application of computational morality. It is nevertheless interesting to explore in the future the application of LP-based probabilistic reasoning to study degrees of blame and moral responsibility.

LP abduction is used in [33] to model moral reasoning in various scenarios of the trolley problem, both from DDE and DTE viewpoints, sans counterfactuals. Abducibles are used to represent decisions, e.g., diverting the trolley, pushing the man, etc. Impemissible actions are ruled out using an IC, and a posteriori preferences are eventually enacted to come up with a moral decision from the remaining alternatives of action. The subsequent work [16] refines it with uncertainty of actions and consequences in several scenarios of the trolley problem by resorting to P-log.

One of the difficulties in using an IC to express impermissibility is that it requires the representation to be crafted sufficiently in detail in order for the IC to be applicable. The examples in the present paper have not exploited the full potential of ICs yet. While we use counterfactuals to examine permissibility (so we are not bound to have a subtle problem representation), ICs can be used for other purposes, e.g., if LP programs for teb and tab examples are combined, IC: false ← teb, tab can be introduced to choose among mutually exclusive abducibles, teb or tab. The decision to have separate models for them in this paper is solely for clearer presentation. Nevertheless, ICs should be treated carefully in counterfactuals, because an intervention may render ICs unsatisfiable, and hence their body’s support may need to be abductively revised in order to re-impose satisfaction.

Side-effects in abduction have been investigated in [32] through the concept of inspection points; the latter are construed in a procedure by ‘meta-abducing’ a specific abducible abduced(A) whose function is only checking that its corresponding abducible A is indeed already adopted elsewhere. Therefore, the consequence of the action that triggers this ‘meta-abducing’ is merely a side-effect. Indeed, inspection points may be employed to distinguish a cause from a mere side-effect, and thus may provide an alternative or supplement to counterfactuals employed for the same purpose.

6 Conclusion and Future Work

This paper presents a formulation of counterfactuals evaluation by means of LP abduction and updating. The approach corresponds to the three-step process in Pearl’s structural theory, but omits probability to concentrate on a naturalized logic. We addressed too how to examine (non-probabilistic) moral reasoning about permissibility, employing this LP approach to distinguish between causes and side-effects as a result of agents’ actions to achieve a goal.
Counterfactuals may as well be suitable to address moral justification, via ‘compound counterfactuals’: Had I known what I know today, then if I were to have done otherwise, something preferred would have followed. Such counterfactuals, typically imagining alternatives with worse effect – the so-called downward counterfactuals [24], may provide moral justification for what was done due to lack of the current knowledge. This is accomplished by evaluating what would have followed if the intent had been otherwise, other things (including present knowledge) being equal. It may justify that what would have followed is no morally better than the actual ensued consequence. QUALM can evaluate such compound counterfactuals, thanks to its implemented incremental tabling of fluents [36]. Because fluents are tabled and time-stamped, events in the past subjected to hypothetical updates of intervention can readily be accessed. Indeed, these hypothetical updates take place without requiring any undoing of other fluent updates, from the state those past events occurred in up to the current one; more recent updates are kept in tables and readily provide the current knowledge. We are investigating the application of compound counterfactuals, e.g., to justify an exception for an action to be permissible, that may lead to agents’ argumentation following Scanlon’s contractualism [39].

Acknowledgements

Ari Saptawijaya acknowledges the support of Fundação para a Ciência e a Tecnologia (FCT/MEC) Portugal with grant SFRH/BD/72795/2010. We thank Emmanuelle-Anna Dietz for the fruitful discussions.

References