Counterfactuals, Logic Programming and Agent Morality

Luís Moniz Pereira[1] and Ari Saptawijaya[2]

[1] NOVA Lab. for Computer Science and Informatics, Universidade Nova de Lisboa, Portugal
lmp@fct.unl.pt
[2] Faculty of Computer Science, Universitas Indonesia, Indonesia.
saptawijaya@cs.ui.ac.id

Abstract. 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 updating 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: abduction, counterfactual, logic programming, morality, non-monotonic reasoning.

1 Introduction

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 [11]: *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 [13, 24, 34], psychology [11, 16, 35, 36, 39, 51]. They also have been studied from the computational viewpoint [8, 22, 41, 42, 62], where approaches in Logic Programming (LP), e.g., [8, 42, 62], are mainly based on probabilistic reasoning.

In the first contribution, we innovatively make use of LP abduction and updating in an implemented procedure for evaluating counterfactuals, taking the established approach of Pearl [41] as reference. Our approach concentrates on pure non-probabilistic counterfactual reasoning in LP – thus distinct from but complementing existing probabilistic approaches – by instead resorting to abduction and updating, in order to determine the logical validity of counterfactuals under the Well-Founded Semantics [61]. Nevertheless, the approach is adaptable to other semantics, e.g., Weak Completion Semantics [29] is employed in [45]. Our approach is therefore suited and applicable to instances when probabilities are not known or needed.

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 select points adjustments to the current model via hypothetical updates of intervention.

Counterfactual thinking in moral reasoning has been investigated particularly via psychology experiments (see, e.g., [16, 36, 37, 39]), but it has only been limitedly explored in machine ethics. In the second contribution, counterfactual reasoning is applied to machine ethics, an interdisciplinary field that emerges from the need of imbuing autonomous agents with the capacity for moral decision making to enable them to function in an ethically responsible manner via their own ethical decisions. The potential of LP for machine ethics has been reported in [25,33,46,54], where the main characteristics of morality aspects can appropriately be expressed by LP-based reasoning, such as abduction, integrity constraints, preferences, updating, and argumentation. The application of counterfactual reasoning to machine ethics – herein by resorting to our LP approach – therefore aims at more generally taking counterfactuals to the wider context of the aforementioned well-developed LP-based non-monotonic reasoning methods, these being also all important, appropriate, and promising for moral reasoning.

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 encompass bad outcomes and "close calls" (some harm that was close to happening). Second, 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 [51] 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 defending 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 counterfactuals, supported by our LP approach, are capable and appropriate for expressing different viewpoints on permissibility according to both doctrines based on views argued in the literature. By materializing these doctrines in concrete moral dilemmas, the results of counterfactual evaluation are readily comparable to those from the literature. Abstaining from probability permits focusing on the naturalized logic of human counterfactual moral reasoning. Moreover, people naturally do not compute formal probabilities, nor probabilities are always available, when making moral decisions via counterfactuals, though one can benefit from counterfactuals for inferring intentions through a probabilistic model to explain moral permissibility [32]. Note that, even though the LP technique introduced in this paper is relevant for modeling counterfactual moral reasoning, its use is general, not specific to morality.

The paper is organized as follows. Section 2 reviews basic notation in LP, abduction 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 machine ethics 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 machine ethics, within a combination of abduction and updating.

2 Preliminaries

By an alphabet $\mathcal{A}$ of a language $\mathcal{L}$ we mean a countable disjoint set of constants, function symbols, and predicate symbols. Moreover, an alphabet is assumed to contain a countable set of variable symbols. A term over $\mathcal{A}$ is defined recursively as either a variable, a constant or an expression of the form $f(t_1, \ldots, t_n)$, where f is a function symbol of $\mathcal{A}$, and t_is are terms. An *atom* over $\mathcal{A}$ is an expression of the form $p(t_1, \ldots, t_n)$, where p is a predicate symbol of $\mathcal{A}$, and t_is are terms. A *literal* is either an atom a or its negation $not\ a$. Literals of the latter form is called default literals. The *negation complement of a literal* L is denoted by $compl(L)$, where the negation complement of a positive literal a and its negation $not\ a$ are defined as $compl(a) = not\ a$ and $compl(not\ a) = a$, respectively.

A term (respectively, atom and literal) is *ground* if it does not contain variables. The set of all ground terms (respectively, ground atoms) of $\mathcal{A}$ is called the Herbrand universe (respectively, Herbrand base) of $\mathcal{A}$.

Definition 1 (Logic Program). *A* (normal) logic program *is a countable set of rules of the form:*

$$H \leftarrow L_1, \ldots, L_m$$

where H is an atom, $m \geq 0$, and L_is ($1 \leq i \leq m$) are literals.[3]

The comma operator in rules is read as conjunction. A normal logic program is called *definite* if none of its rules contains default literals. Following the standard convention, rules of the form $H \leftarrow$ are alternatively written as H. A rule of this form is called a *fact*.

The alphabet $\mathcal{A}$ used to write program P is assumed to precisely comprise all the constants, the function and predicate symbols that explicitly appear in P. By Herbrand universe (respectively, base) of P we mean the Herbrand universe (respectively, base) of $\mathcal{A}$. We denote the Herbrand base of P by $\mathcal{H}_P$. By a ground logic program we mean the set of ground rules obtained from P by substituting in all possible ways each of the variables in P by elements of its Herbrand universe.

We define next three-value Herbrand interpretations and models of logic programs.[4] It permits representing incomplete knowledge, where some atoms are neither true nor false, but rather undefined. Let F be a set of atoms, $F = \{a_1, \ldots, a_n\}$, and $not\ F = \{not\ a_1, \ldots, not\ a_n\}$.

Definition 2 (Three-valued Interpretation). *A* three-valued interpretation *I of a logic program P is a set of literals*

$$I = T \cup not\ F$$

such that $T \subseteq \mathcal{H}_P$, $F \subseteq \mathcal{H}_P$ and $T \cap F = \emptyset$.

In a three-valued interpretation, the set T (respectively, F) is the set of atoms that are true (respectively, false) in I, and the truth value of the remaining atoms is undefined.

We may view an interpretation I of a program P as a a function $I : \mathcal{H}_P \rightarrow \mathcal{V}$, where $\mathcal{V} = \{0, 0.5, 1\}$, defined by:

$$I(A) = \begin{cases} 0 & \text{if } not\ A \in I \\ 1 & \text{if } A \in I \\ 0.5 & \text{otherwise} \end{cases}$$

Models are defined as usual, and based on a truth valuation function.

Definition 3 (Truth Valuation). *If I is an interpretation, the* truth valuation *$\hat{I}$ corresponding to I is a function $\hat{I} : \mathcal{F} \rightarrow \mathcal{V}$, where $\mathcal{F}$ is the set of ground literals, conjunctions of literals, and rules formed over the language. It is defined as follows:*

- *If L is a ground atom, then $\hat{I}(L) = I(L)$.*
- *If L is a default literal, i.e., $L = not\ A$, then $\hat{I}(L) = 1 - \hat{I}(A)$.*
- *If S and T are conjunctions of literals, then $\hat{I}((S, T)) = min(\hat{I}(S), \hat{I}(T))$.*
- *If $H \leftarrow B$ is a rule, where B is a conjunction of literals, then:*

$$\hat{I}(H \leftarrow B) = \begin{cases} 1 & \textit{if } \hat{I}(B) \leq \hat{I}(H) \\ 0 & \textit{otherwise} \end{cases}$$

[3] In the sequel, unless otherwise specified, we generally write *logic programs*, or simply *programs*, to refer to *normal logic programs*.

[4] In the sequel, we simply write *interpretations* and *models* to refer to *Herbrand interpretations* and *Herbrand models*, respectively.

For any $F \in \mathcal{F}$, the values 0, 0.5 and 1 of $\hat{I}(F)$ correspond to the truth values false*,* undefined *and* true*, respectively. We write $I \models F$, for $F \in \mathcal{F}$, iff $\hat{I}(F) = 1$.*

Definition 4 (Model). *A interpretation I is called a* model *of a program P iff for every ground instance $H \leftarrow B$ of a rule in program P we have $\hat{I}(H \leftarrow B) = 1$.*

We define some orderings among interpretations and models as follows.

Definition 5 (Classical Ordering [49]). *If I and J are two interpretations then we say that $I \preceq J$ if $I(A) \leq J(A)$ for any ground atom A. If $\mathcal{I}$ is a collection of interpretations, then an interpretation $I \in \mathcal{I}$ is called* minimal *in $\mathcal{I}$ if there is no interpretation $J \in \mathcal{I}$ such that $J \preceq I$ and $J \neq I$. An interpretation I is called* least *in $\mathcal{I}$ if $I \preceq J$, for any other interpretation $J \in \mathcal{I}$. A model M is called* minimal *(respectively,* least*) if it is minimal (respectively, least) among all models of P.*

Definition 6 (Fitting Ordering [17]). *If I and J are two interpretations then we say that $I \preceq_F J$ iff $I \subseteq J$. If $\mathcal{I}$ is a collection of interpretations, then an interpretation $I \in \mathcal{I}$ is called* F-minimal *in $\mathcal{I}$ if there is no interpretation $J \in \mathcal{I}$ such that $J \preceq_F I$ and $J \neq I$. An interpretation I is called* F-least *in $\mathcal{I}$ if $I \preceq_F J$, for any other interpretation $J \in \mathcal{I}$. A model M is called* F-minimal *(respectively,* F-least*) if it is F-minimal (respectively, F-least) among all models of P.*

Note that the classical ordering is related with the *degree of truth* of their atoms, whereas the Fitting ordering is related with the *degree of information*. Under the latter ordering, the undefined value is less than both values true and false, providing that true and false being incompatible.

In [60], it is shown that every definite program has a unique least model, which determines the so-called *least model semantics* of a definite program. Other semantics for more general programs, allowing default literals in the body of a rule, have been proposed. In [21], *Stable Model Semantics* is introduced. Informally, when one assumes true some set of (hypothetical) default literals, and false all the others, some consequences follow according to the semantics of definite programs. If the consequences completely corroborate the hypotheses made, then they form a stable model.

Despite its advantages, that it provides semantics for more general programs than its predecessors and is closely related to autoepistemic logic and default theory (see [20] and [9]), Stable Model Semantics has some drawbacks. Some programs may have no stable models, e.g., the program $p \leftarrow not\ p$. Even for programs with stable models, their semantics do not always lead to the expected intended semantics (see [3] for a discussion).

The *Well-Founded Semantics* [61], which we refer to in this paper, addresses the difficulties encountered with the Stable Model Semantics. It has been shown in [50] that the Well-Founded Semantics is also equivalent to major formalizations of non-monotonic reasoning.

The Well-Founded Semantics can be viewed as *three-valued Stable Model Semantics* [48]. In order to formalize the notion of three-valued stable models, the language of programs is expanded with the additional propositional constant **u** with the property

of being undefined in every interpretation. It is therefore assumed that every interpretation I satisfies:

$$\hat{I}(\mathbf{u}) = \hat{I}(not\ \mathbf{u}) = 0.5$$

A *non-negative* program is a program whose rules' bodies are either atoms or $\mathbf{u}$. It is proven in [48] that every non-negative logic program has a unique least three-valued model.

The next definition extends the Gelfond-Lifschitz operator [21] to a three-valued operator Γ^*.

Definition 7 (Γ^*-operator). *Let P be a logic program and I be its three-valued interpretation. The* extended GL-transformation of P modulo I *is the program $\frac{P}{I}$ obtained from P by performing the following operations:*

- *Remove from P all rules which contain a default literal $L = not\ A$ such that $\hat{I}(L) = 0$;*
- *Replace in the remaining rules of P those default literals $L = not\ A$ which satisfy $\hat{I}(L) = 0.5$ by $\mathbf{u}$;*
- *Remove from all remaining rules those default literals $L = not\ A$ which satisfy $\hat{I}(L) = 1$.*

Since the resulting program $\frac{P}{I}$ is non-negative, it has a unique three-valued least model J. We define $\Gamma^*(I) = J$.

Definition 8 (Well-Founded Semantics). *A three-valued interpretation I of a logic program P is a three-valued stable model of P if $\Gamma^*(I) = I$. The* Well-Founded Semantics *of P is determined by the unique F-least three-valued stable model of P, and can be obtained by the bottom-up iteration of Γ^* starting from the empty interpretation.*

Example 1. Consider program P:

$$\begin{array}{l} a \leftarrow not\ b. \\ b \leftarrow not\ a. \\ c \leftarrow not\ d. \\ d \leftarrow not\ e. \\ p \leftarrow a. \\ p \leftarrow b. \end{array}$$

Let $I_0 = \emptyset$ be the empty interpretation.

- The least three-valued model of $\frac{P}{I_0}$:

$$\begin{array}{l} a \leftarrow \mathbf{u}. \\ b \leftarrow \mathbf{u}. \\ c \leftarrow \mathbf{u}. \\ d \leftarrow \mathbf{u}. \\ p \leftarrow a. \\ p \leftarrow b. \end{array}$$

is $\Gamma^*(I_0) = \{not\ e\}$.

– Let $I_1 = \Gamma^*(I_0)$. The least three-valued model of $\frac{P}{I_1}$:

$$
\begin{array}{l}
a \leftarrow \mathbf{u}. \\
b \leftarrow \mathbf{u}. \\
c \leftarrow \mathbf{u}. \\
d \leftarrow . \\
p \leftarrow a. \\
p \leftarrow b.
\end{array}
$$

is $\Gamma^*(I_1) = \{d, not\ e\}$.

– Let $I_2 = \Gamma^*(I_1)$. The least three-valued model of $\frac{P}{I_2}$:

$$
\begin{array}{l}
a \leftarrow \mathbf{u}. \\
b \leftarrow \mathbf{u}. \\
d \leftarrow . \\
p \leftarrow a. \\
p \leftarrow b.
\end{array}
$$

is $\Gamma^*(I_2) = \{d, not\ c, not\ e\}$.

– Let $I_3 = \Gamma^*(I_2)$. The least three-valued model of $\frac{P}{I_3}$:

$$
\begin{array}{l}
a \leftarrow \mathbf{u}. \\
b \leftarrow \mathbf{u}. \\
d \leftarrow . \\
p \leftarrow a. \\
p \leftarrow b.
\end{array}
$$

is $\Gamma^*(I_3) = \{d, not\ c, not\ e\}$.

Therefore, the well-founded model of P is $I_3 = \{d, not\ c, not\ e\}$, where d is true, c and e are both false, and a, b and p are undefined.

In the sequel, we write the well-founded model of program P as *WFM*(P).

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. That is, the best explanation can generally be determined through some integrity constraints or preference rules [14, 44], rather than simply taking a minimal one. In LP, an abductive hypothesis (*abducible*) is a positive literal Ab or its negation complement Ab^* (syntactically an atom, but denoting literal $not\ Ab$), whose truth value is not initially assumed. An *abductive logic program* (ALP) is one allowing abducibles in the body of rules.

We next define an abductive framework in LP [30], which includes integrity constraints for restricting abduction, under the Well-Founded Semantics [61]. The definitions in this section are adapted from those of [4].

Definition 9 (Integrity Constraint). *An* integrity constraint *is a rule in the form of a denial:*

$$\bot \leftarrow L_1, \ldots, L_m.$$

where $\bot/0$ *is a reserved predicate symbol in* $\mathcal{L}$*,* $m \geq 1$*, and* L_i*s (*$1 \leq i \leq m$*) are literals.*

Definition 10 (Abductive Framework). *An* abductive framework *is a triple* $\langle P, \mathcal{A}, \mathcal{I} \rangle$*, where* $\mathcal{A}$ *is the set of abducibles,* P *is a logic program over* $\mathcal{L} \setminus \{\bot\}$ *such that there is no rule in* P *whose head is in* $\mathcal{A}$*, and* $\mathcal{I}$ *is a set of integrity constraints.*

Definition 11 (Abductive Scenario). *Let* F *be an abductive framework* $\langle P, \mathcal{A}, \mathcal{I} \rangle$*. An* abductive scenario *of* F *is a tuple* $\langle P, \mathcal{A}, \mathcal{S}, \mathcal{I} \rangle$*, where* $\mathcal{S} \subseteq \mathcal{A}$ *and there is no* $A \in \mathcal{S}$ *such that* $compl(A) \in \mathcal{S}$*, i.e.,* $\mathcal{S}$ *is consistent.*

The consistency of an abductive scenario can be imposed by an integrity constraint $\bot \leftarrow Ab, Ab^*$.

Let observation O be a set of literals, analogous to a query in LP. Abducing an explanation for O amounts to finding consistent abductive solutions to a goal, whilst satisfying the integrity constraints, where abductive solutions consist in the semantics obtained by replacing in P the abducibles in $\mathcal{S}$ by their truth value. We define formally abductive solutions under the Well-Founded Semantics below.

Given an abductive scenario $\langle P, \mathcal{A}, \mathcal{S}, \mathcal{I} \rangle$ of an abductive framework $\langle P, \mathcal{A}, \mathcal{I} \rangle$, we first define $P_{\mathcal{S}}$ as the smallest set of rules that contains for each $A \in \mathcal{A}$, the fact A if $A \in \mathcal{S}$; and $A \leftarrow \mathbf{u}$ otherwise. Alternatively, and obviously equivalent, instead of adding to $P_{\mathcal{S}}$ the rule $A \leftarrow \mathbf{u}$, one may simply replace the corresponding A with $\mathbf{u}$ both in P and $\mathcal{I}$.

Definition 12 (Abductive Solution). *Let* F=$\langle P, \mathcal{A}, \mathcal{I} \rangle$ *and* $\langle P, \mathcal{A}, \mathcal{S}, \mathcal{I} \rangle$ *be an abductive scenario of* F*. The consistent set of abducibles* $\mathcal{S}$ *is an* abductive solution *to* F *if* $\bot$ *is* false *in* $M_s = WFM(P \cup P_{\mathcal{S}} \cup \mathcal{I})$*.*

We further say that $\mathcal{S}$ *is an* abductive solution for query Q *if* Q *is* true *in* M_s*, written* $M_s \models Q$*.*

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 is recursively generated 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 Well-Founded Semantics (WFS) enjoys this 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.

Example 2. A library in a city is closed with two possible explanations: it is closed in the weekend, or when it is not weekend, there are no librarians working. These days,

librarians are often absent from their work because they participate in a strike. On the other hand, a museum in that city is only closed when there is a special visit by important guests. This example can be expressed by an abductive framework $\langle P, \mathcal{A}, \mathcal{I} \rangle$, where $\mathcal{A} = \{weekend, weekend^*, strike, strike^*, specialVisit, specialVisit^*\}$, P is the program below, and $\mathcal{I} = \emptyset$:

$$closed\_library \leftarrow weekend \qquad closed\_library \leftarrow weekend^*, absent$$
$$absent \leftarrow strike \qquad closed\_museum \leftarrow specialVisit$$

Consider the query $Q = \{closed\_library\}$. This query induces the call-graph with $closed\_library$ as its root and, through procedure calls, it ends with two leaves: one leaf containing abducible $\{weekend\}$, and the other containing $\{weekend^*, strike\}$. That is, there are two abductive solutions for query Q, viz., $\mathcal{S}_1 = \{weekend\}$ and $\mathcal{S}_2 = \{weekend^*, strike\}$. Note that the abducible $specialVisit$ is not mentioned in either solution: its truth value is irrelevant for query Q because the WFS enjoys the relevancy property, namely it finds only relevant abducibles (and their truth value) through the aforementioned procedure calls, driven by the considered query Q.

Example 3. Recall Example 2 plus new information: librarians may also be free from working in case the library is being renovated. The rule below is added to program P, where $renov$ and $renov^*$ are new abducibles in $\mathcal{A}$:

$$absent \leftarrow renov$$

The same query $Q = \{closed\_library\}$ now returns an additional solution, viz. $\mathcal{S}_3 = \{weekend^*, renov\}$. Let us further suppose that the municipal authority permits any renovations to take place only on weekends, expressed as an integrity constraint in $\mathcal{I}$:

$$\mathcal{I} = \{\bot \leftarrow weekend^*, renov\}$$

That is, $\mathcal{S}_3 = \{weekend^*, renov\}$ is now ruled out from the abductive solutions.

2.2 Pearl's Structure-based Counterfactuals

Pearl [41] 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 (*background variables*) and V (*endogenous variables*), 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.

Every causal model M can be associated with a directed graph, called the *causal diagram* of M. This diagram identifies the endogenous and background variables (U and V, resp.) that have direct influence on each V_i.

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 $P(u \mid e)$. 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, viz., the aforementioned Pearl's three-step procedure. Since the idea of each step in our LP approach mirrors the one corresponding in Pearl's, our approach therefore immediately compares to Pearl's, benefits from its epistemic adequacy, and its properties rely on those of Pearl's. 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 abductive framework $\langle P, \mathcal{A}, \mathcal{I} \rangle$, 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 abductive

framework $\langle P, \mathcal{A}, \mathcal{I} \rangle$ 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.

The key idea of evaluating counterfactuals with respect to an abductive framework, at some current state (discrete time) T, is as follows. In step 1, abduction is performed to explain the factual observation.[5] The observation corresponds to the evidence that both the antecedent and the consequence literals of the present counterfactual were factually false.[6] There can be multiple explanations available to an observation; choosing a suitable one among them is a pragmatic issue, which can be dealt with integrity constraints or preferences [14, 44]. 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

[5] 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 must simply be made explicit observations, to avoid natural language conundrums or ambiguities [23].

[6] This interpretation is in line with the corresponding English construct, cf. [27], commonly known as *third conditionals*.

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 the aforementioned 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). The following definitions are needed by the procedure.

Definition 13. *A set of integrity constraint is* satisfied *in $WFM(P)$ iff none is false in $WFM(P)$. That is, the body of an integrity constraint is either false or undefined [43].*

We next rephrase Definition 12 about abductive solutions and relate them to explanations of observations. As our counterfactual procedure is based on the Well-Founded Semantics, the standard logical consequence relation $P \models F$ used in the definition below presupposes the Well-Founded Model of P in verifying the truth of formula F, i.e., whether F is true in $WFM(P)$.

Definition 14. *Given an abductive framework $\langle P, \mathcal{A}, \mathcal{I} \rangle$ and an observation O, a consistent abductive solution $E \subseteq \mathcal{A}$ is an* explanation *to observation O iff $P \cup E \models O$ and $\mathcal{I}$ is satisfied in $WFM(P \cup E)$, where all abducibles not appearing in E have been replaced by $\mathbf{u}$, both in P and $\mathcal{I}$.*[7]

Procedure 2. Let $\langle P, \mathcal{A}, \mathcal{I} \rangle$ be an abductive framework, where program P encodes the modeled situation on which counterfactuals are evaluated. Consider a counterfactual "if Pre had been true, then $Conc$ would have been true", where Pre and $Conc$ are finite conjunctions of literals.

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

   Update program P with E, obtaining program $P \cup E$.
2. ***Action***: For each literal L in conjunction Pre, introduce a pair of reserved meta-predicates $make(B)$ and $make\_not(B)$, 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*:

[7] This replacement of abducible $A \notin E$ with $\mathbf{u}$ in P and $\mathcal{I}$ is an alternative but equivalent to adding $A \leftarrow \mathbf{u}$ into $P \cup E$, as foreseen by Definition 12.

- Add rule $B \leftarrow make(B)$ to program $P \cup E$.
- Add $not\ make\_not(B)$ to the body of each rule in P whose head is B. If there is no such rule, add rule $B \leftarrow not\ make\_not(B)$ to program $P \cup E$.

Let $(P \cup E)_\tau$ be the resulting transform.

(b) *Intervention*: Update program $(P \cup E)_\tau$ with literal $make(B)$ or $make\_not(B)$, for $L = B$ or $L = not\ B$, resp. Assuming that Pre is consistent, $make(B)$ and $make\_not(B)$ cannot be imposed at the same time.

Let $(P \cup E)_{\tau,\iota}$ be the program obtained after these hypothetical updates of intervention.

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

This three-step procedure defines *valid* counterfactuals.

Definition 15. *Let $\langle P, \mathcal{A}, \mathcal{I} \rangle$ be an abductive framework, where program P encodes the modeled situation on which counterfactuals are evaluated. 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 $E \subseteq \mathcal{A}$, $(P \cup E)_{\tau,\iota} \models Conc$, and $\mathcal{I}$ is satisfied in $WFM((P \cup E)_{\tau,\iota})$.*

Since the Well-Founded Semantics supports top-down query-oriented procedures for finding solutions, 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 $\mathcal{I}$.

Example 4. 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 abductive framework $\langle P, \mathcal{A}, \mathcal{I} \rangle$, using abbreviations b, d, f, g, l, s for *barbecue, dry leaves, forest fire, leaves on the ground, lightning*, and *storm*, resp., where $\mathcal{A} = \{s, b, s^*, b^*\}$, $\mathcal{I} = \emptyset$, and P as follows:

$$f \leftarrow b, d. \qquad f \leftarrow b^*, l, d, g. \qquad l \leftarrow s. \qquad g \leftarrow s. \qquad d.$$

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 $(P \cup E_1)_\tau$:
$$f \leftarrow b, d. \qquad f \leftarrow b^*, l, d, g. \qquad g \leftarrow s. \qquad d.$$
$$l \leftarrow make(l) \qquad l \leftarrow s, not\ make\_not(l)$$
Program $(P \cup E_1)_\tau$ is updated with $make\_not(l)$ as the required intervention. It engenders program $(P \cup E_1)_{\tau,\iota}$ corresponding to a new state with $WFM((P \cup E_1)_{\tau,\iota}) = \{d, s, g, make\_not(l), not\ make(l), not\ b, not\ l, not\ f\}$.
3. ***Prediction***: We verify that $(P \cup E_1)_{\tau,\iota} \models not\ f$, and $\mathcal{I} = \emptyset$ is trivially satisfied in $WFM((P \cup E_1)_{\tau,\iota})$.

We thus conclude that, for this E_1 scenario, the given counterfactual is valid.

Example 5. In the other explanatory scenario of Example 4, 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 [11], 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 $O_{Conc} = \{L_i \mid L_i \text{ is in } Conc\}$, to warrant its consequence factually true. The validity condition for semifactuals is the same as for counterfactuals, cf. Definition 15.

Example 6. Recall Example 5, 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 5, we now have $(P \cup E_2)_{\tau,\iota} \models Conc$, i.e., $(P \cup E_2)_{\tau,\iota} \models f$.

3.4 Some 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 [34], salient properties of counterfactuals in logic are argued (and in [45] for the Weak Completion Semantics), including three counterfactual fallacies that distinguish the counterfactual conditional from the material one. Our approach also satisfies these three properties, as shown below for illustration, for the Well Founded Semantics. Other counterfactual properties, such as reflexive, modus tollens, disjunction in the antecedent, combination of sentences, etc., are postponed for future work; ascertaining their satisfaction is not in the purview of the present paper.

Let $Pre \triangleright Conc$ represent counterfactual statement "if Pre had been true, then $Conc$ would have been true".

Property 1. Fallacy of strengthening the antecedent:
$A \triangleright B$ does not imply $A \wedge C \triangleright B$.

Example 7. Recall Example 4, where E_1 is adopted. We have shown that counterfactual $not\ l \triangleright not\ f$ is valid. Let us strengthen its antecedent with "there had been a barbecue", obtaining counterfactual $not\ l \wedge b \triangleright not\ f$. For this new counterfactual, $O_{Pre} = \{l, not\ b\}$, whereas the other observations (O_{Conc} and O_{Oth}) are the same as in Example 4. The only explanation of $O = O_{Pre} \cup O_{Conc} \cup O_{Oth}$ is $E_{pr_1} = \{s, b^*\}$. The transform $(P \cup E_{pr_1})_\tau$ is as follows:

$$f \leftarrow b, d. \qquad f \leftarrow b^*, l, d, g. \qquad g \leftarrow s. \qquad d.$$
$$l \leftarrow make(l) \qquad l \leftarrow s, not\ make\_not(l)$$
$$b \leftarrow make(b) \qquad b \leftarrow not\ make\_not(b)$$

The required interventions $make\_not(l)$ and $make(b)$ update this program, obtaining $(P \cup E_{pr_1})_{\tau,\iota}$ with $WFM((P \cup E_{pr_1})_{\tau,\iota}) = \{d, s, g, make\_not(l), not\ make(l), not\ l, make(b), not\ make\_not(b), b, f\}$. Observe that, intervention $make(b)$ hypothetically updates the truth value of b from false (in $P \cup E_{pr_1}$) to true (in $(P \cup E_{pr_1})_{\tau,\iota}$). Since $(P \cup E_{pr_1})_{\tau,\iota} \not\models not\ f$, counterfactual $not\ l \wedge b \triangleright not\ f$ is not valid.

Property 2. Fallacy of contraposition:
$A \triangleright B$ does not imply $not\ B \triangleright not\ A$.

Example 8. Recall the abductive framework of Example 4, but with $O_{Oth} = \emptyset$, rendering $O = \{l, f\}$, and thus the two explanations E_1 and E_2 do not change. Therefore, E_1 being preferred, the counterfactual $not\ l \triangleright not\ f$ is valid, as shown in that example. Consider its contraposition: $f \triangleright l$. Its corresponding observation $O_{pr_2} = \{not\ f, not\ l\}$ admits a single explanation $E_{pr_2} = \{s^*, b^*\}$. Though intervention $make(f)$ is imposed on the transform $(P \cup E_{pr_2})_\tau$, we obtain $not\ l \in WFM((P \cup E_{pr_2})_{\tau,\iota})$. Thus, $(P \cup E_{pr_2})_{\tau,\iota} \not\models l$, and $f \triangleright l$ is not a valid counterfactual.

Property 3. Fallacy of transitivity:
$A \triangleright B$ and $B \triangleright C$ do not imply $A \triangleright C$.

Example 9. Let P_t be program:

$$marry \leftarrow pregnant \quad criticized \leftarrow not\ marry, pregnant$$

in the abductive framework $\langle P_t, \mathcal{A}_t, \mathcal{I}_t \rangle$, using abbreviation c, m, p for $criticized$, $marry$, and $pregnant$, resp., where $A_t = \{p, p^*\}$ and $I_t = \emptyset$. Consider counterfactuals C_1, C_2, C_3:
(C_1) $not\ p \triangleright not\ m$; ($C_2$) $not\ m \triangleright c$; and (C_3) $not\ p \triangleright c$.

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 $make\_not(p)$, since $not\ m \in WFM((P \cup E_{C_1})_{\tau,\iota})$, and thus $(P \cup E_{C_1})_{\tau,\iota} \models not\ m$. 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 $O_{C_2} = \{m, not\ c\}$, and the imposed intervention is $make\_not(m)$. But C_3 is not valid, since $(P \cup E_{C_3})_{\tau,\iota} \not\models c$, given that $not\ c \in WFM((P \cup E_{C_3})_{\tau,\iota})$, where $E_{C_3} = \{p\}$ explains $O_{C_3} = \{p, not\ c\}$ with intervention $make\_not(p)$.

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 [53], in XSB Prolog [57]. QUALM allows specifying predicates that are subject to intervention, e.g., predicate l in Example 4. 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., `?-` $query((l, f, g), E_1)$ provides explanation E_1 to the observation $O = \{l, f, g\}$ of Example 4. 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 $updates(L)$ 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 $make\_not(l)$ via query `?-` $updates([make\_not(l)])$. 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 $make\_not(l)$ 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., `?-` $query(f, E)$ 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., `?-` $updates([not\ make\_not(l)])$, 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 $make\_not(l)$ 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 [36] and the influence of counterfactual thoughts in moral judgment [39]. As argued in [16], the function of counterfactual thinking is not just limited to the evaluation process, but occurs also in the reflective 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.

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 available some control of the outcome [63].

The potential of LP to machine ethics has been reported in [25, 33, 46] and with emphasis on LP abduction and updating in [54]. 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) [38] and of Triple Effect (DTE) [31].

DDE is first introduced by Thomas Aquinas in his discussion of the permissibility of self-defense [7]. The current version of this principle emphasizes 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 an *intended means* to bringing about the same good end [38]. According to the Doctrine of Double Effect, the former action is permissible, whereas the latter is impermissible. In [26], DDE has been utilized to explain the consistency of judgments, shared by subjects from demographically diverse populations, on a number of variants of the classic trolley problem [18]: *A trolley is headed toward five people walking on the track, who are unable to get off the track in time. The trolley can nevertheless be diverted onto a side track, thereby preventing it from killing the five people. However, there is a man standing on the side track. The dilemma is therefore whether it is morally permissible to divert the trolley, killing the man but saving the five*. DDE permits diverting the trolley since that action does not intend to harm the man on the side track in order to save the five.

Counterfactuals may provide a general way to examine DDE in moral dilemmas, 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:

If not E *had been true, then* not G *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).

There has been a number of studies, both in philosophy and psychology, on the relation between causation and counterfactuals. The *counterfactual process view* of causal reasoning [37], for example, advocates counterfactual thinking as an essential part of the process involved in making causal judgments. This relation between causation and counterfactuals can be important for providing explanations in cases involving harm, which underlie people's moral cognition [58] and trigger other related questions, such as "Who is responsible?", "Who is to blame?", "Which punishment would be fair?", etc. Herein, we explore the connection between causation and counterfactuals, focusing on agents' deliberate action, rather than on causation and counterfactuals in general. More specifically, our exploration of this topic links it to the Doctrines of Double Effect and Triple Effect and dilemmas involving harm, such as the trolley problem cases. Such cases have also been considered in psychology experimental studies concerning the role of gender and perspectives (first vs. third person perspectives) in counterfactual thinking in moral reasoning, see [39]. The reader is referred to [13] and [28] 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 [56] – abbreviations in parentheses: terror bombing (teb) vs. tactical bombing (tab). The former refers to bombing a civilian target (civ) during a war, thus killing civilians (kic), in order to terrorize the enemy (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 10. We first model terror bombing with ew as the goal, by considering the abductive framework $\langle P_e, \mathcal{A}_e, \mathcal{I}_e \rangle$, where $\mathcal{A}_e = \{teb, teb^*\}, \mathcal{I}_e = \emptyset$ and P_e:

$$ew \leftarrow ror \quad ror \leftarrow kic \quad kic \leftarrow civ \quad civ \leftarrow teb$$

We consider counterfactual "if civilians had not been killed, then the war would not have ended", where $Pre = not\ kic$ and $Conc = not\ ew$. The observation $O = \{kic, ew\}$, with O_{Oth} being empty, has a single explanation $E_e = \{teb\}$. The rule $kic \leftarrow civ$ transforms into $kic \leftarrow civ, not\ make\_not(kic)$. Given intervention $make\_not(kic)$, the counterfactual is valid, because $not\ ew \in WFM((P_e \cup E_e)_{\tau,\iota})$, and thus $(P_e \cup E_e)_{\tau,\iota} \models not\ ew$. That means the morally wrong kic is instrumental in achieving the goal ew: it is a cause for ew by performing teb and not a mere side-effect of teb. Hence teb is DDE morally impermissible.

Example 11. Tactical bombing with the same goal ew can be modeled by the abductive framework $\langle P_a, \mathcal{A}_a, \mathcal{I}_a \rangle$, where $\mathcal{A}_a = \{tab, tab^*\}, \mathcal{I}_a = \emptyset$ and P_a:

$$ew \leftarrow mil \qquad mil \leftarrow tab \qquad kic \leftarrow tab$$

Given the same counterfactual, we now have $E_a = \{tab\}$ as the only explanation to the same observation $O = \{kic, ew\}$. Note that the transform contains rule $kic \leftarrow tab, not\ make\_not(kic)$, which is obtained from $kic \leftarrow tab$. By imposing the intervention $make\_not(kic)$, one can verify that the counterfactual is not valid, because $ew \in WFM((P_a \cup E_a)_{\tau,\iota})$, and thus $(P_a \cup E_a)_{\tau,\iota} \not\models not\ ew$. Therefore, the morally wrong kic is just a side-effect in achieving the goal ew. Hence tab is DDE morally permissible.

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 semifactuals to machine ethics, to identify indifferent actions.

Example 12. Consider two countries a and its ally, b, that concert a terror bombing, modeled by the abductive framework $\langle P_{ab}, \mathcal{A}_{ab}, \mathcal{I}_{ab} \rangle$, where $\mathcal{A}_{ab} = \{teb, teb^*\}, \mathcal{I}_{ab} = \emptyset$ and P_{ab} below. The abbreviations $kic(X)$ and $civ(X)$ refer to 'killing civilians by country X' and 'bombing a civilian target by country X'. As usual in LP, underscore (_) represents an anonymous variable.

$$ew \leftarrow ror \quad ror \leftarrow kic(\_)$$
$$kic(X) \leftarrow civ(X) \qquad civ(\_) \leftarrow teb$$

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: $Pre = not\ kic(a) \wedge not\ kic(b)$. One can easily verify that $not\ ew \in WFM((P_{ab} \cup E_{ab})_{\tau,\iota})$, where $E_{ab} = \{teb\}$. Thus, $(P_{ab} \cup E_{ab})_{\tau,\iota} \models not\ ew$ and the counterfactual is valid: the concerted 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 [31] 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 [31] proposed DTE to accommodate a variant of the trolley problem, viz., the *Loop Case* [59]: *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 [40]. Referring to a psychology study [26], 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 [31]. 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 13. We model the Loop case with the abductive framework $\langle P_o, \mathcal{A}_o, \mathcal{I}_o \rangle$, 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, $\mathcal{A}_o = \{div, div^*\}, \mathcal{I}_o = \emptyset$, and P_o:

$$sav \leftarrow hit \quad hit \leftarrow tst, mst \quad tst \leftarrow div \quad mst.$$

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 $O = O_{Pre} \cup O_{Conc} = \{hit, sav\}$, its only explanation is $E_o = \{div\}$. Note that rule $hit \leftarrow tst, mst$ transforms into $hit \leftarrow tst, mst, not\ make\_not(hit)$, and the required intervention is $make\_not(hit)$. The counterfactual is therefore valid, because $not\ sav \in WFM((P_o \cup E_o)_{\tau,\iota})$, hence $(P_o \cup E_o)_{\tau,\iota} \models not\ sav$. This means hit, as a consequence of action div, is instrumental as a cause of goal sav. Therefore, 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 mst ready, without abducing 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 – 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 $E \subseteq \mathcal{A}_o$, i.e., $psh \notin \mathcal{A}_o$, 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 14. Consider a variant of the Loop case, viz., the *Loop-Push Case* (see also Extra Push Case in [31]). 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 abductive framework $\langle P_p, \mathcal{A}_p, \mathcal{I}_p \rangle$, where $\mathcal{A}_p = \{div, psh, div^*, psh^*\}, \mathcal{I}_p = \emptyset$, and P_p:

$$sav \leftarrow hit \quad hit \leftarrow tst, mst \quad tst \leftarrow div \quad mst \leftarrow psh$$

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 abduced for having the man on the side track, so the trolley can be stopped by hitting him. The same intervention $make\_not(hit)$ is applied to the same transform, resulting in a valid counterfactual: $(P_p \cup E_{p_1})_{\tau,\iota} \models not\ sav$, because $not\ sav \in WFM((P_p \cup E_{p_1})_{\tau,\iota})$.
- "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 $psh \leftarrow not\, make\_not(psh)$ in the transform and intervention $make\_not(psh)$, we verify that $(P_p \cup E_{p_2})_{\tau,\iota} \models not\ hit$, as $not\ hit \in WFM((P_p \cup E_{p_2})_{\tau,\iota})$.

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 [45]. 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 [45] – albeit based on different semantics (WCS vs. WFS) – along with some properties specific to our LP-based approach. The present paper complements [45] 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 [45]. Due to the similarity of common features to WFS and WCS, the Propositions and Proofs in [45] 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.[8]

[8] Both WFS and WCS are 3-valued semantics that differ in dealing with close world assumption (CWA) and rules with positive loops (e.g., $p \leftarrow p$). WFS enforces CWA, i.e., atom a that has no

LP abduction and revision are employed in [15] 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 [42], counterfactuals are evaluated using contradiction removal semantics of LP. The work is based on Lewis's counterfactuals [34], 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 is employed, in [8], 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 stable model semantics, but instead on WFS with its relevancy property, which is more appropriate for LP abduction by need as argued earlier. In [62], 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 $make\_not$ to accomplish intervention via defeasible rules and fluent updates, without eliminating rules, as CP-logic does.

Several logic-based approaches have been employed in the above machine ethics research, e.g., in [5, 6, 10, 19, 47]. While some approaches provide implementations in LP, such as in [5, 6, 19], they have not exploited LP-based reasoning features and recent techniques in LP systems that appear essential and promising for moral reasoning and decision making. The approach in [19] mainly just emphasizes the use of default negation in defeasible rules to capture non-monotonic reasoning, whereas the use of LP in [5, 6] is constrained to its purpose of learning rules from cases. Clearly, the potential of Logic Programming goes beyond that.

LP abduction is used in [46] 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. Impermissible actions are ruled out using an integrity constraint, and a posteriori preferences are eventually enacted to come up with a moral decision from the remaining alternatives of action. The subsequent work [25] refines it with uncertainty of actions and consequences in several scenarios of the trolley problem by resorting to P-log.

---

rule is interpreted as false, whereas in WCS undefined. Nevertheless, they can be transformed one to another: adding rules $a \leftarrow u$ and $u \leftarrow not\ u$ for a reserved atom u renders a unknown in WFS; alternatively, adding $a \leftarrow false$ enforces CWA in WCS. In this paper, positive loops are not needed and do not appear throughout examples we consider.

The use of causation, based on structural approach, to define and model issues related to morality, such as blame and responsibility, is discussed in [12, 24]. 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 machine ethics (cf. [25, 46]), particularly in addressing moral permissibility by referring to the Doctrines of Double Effect and Triple Effect. It is nevertheless interesting to explore in the future the application of LP-based probabilistic reasoning to study degrees of blame and moral responsibility.

One of the difficulties in using an integrity constraint to express impermissibility is that it requires the representation to be crafted sufficiently in detail in order for the integrity constraint to be applicable. The examples in the present paper have not exploited the full potential of integrity constraints yet. While we use counterfactuals to examine permissibility (so we are not bound to have a subtle problem representation), integrity constraints can be used for other purposes, e.g., if LP programs for teb and tab examples are combined, integrity constraint: $\bot \leftarrow teb, tab$ 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, integrity constraints should be treated carefully in counterfactuals, because an intervention may render integrity constraints 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 [44] 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* [35], 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 [52]. 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 [55].

Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions. Both authors acknowledge the support from Fundação para a Ciência e a Tecnologia (FCT/MEC) NOVA LINCS PEst UID/CEC/04516/2013. Ari Saptawijaya acknowledges the support from FCT/MEC with the doctoral grant SFRH/BD/72795/2010. We thank Emmanuelle-Anna Dietz for the fruitful discussions.

References

1. J. J. Alferes, A. Brogi, J. A. Leite, and L. M. Pereira. Evolving logic programs. In *Procs. European Conference on Artificial Intelligence (JELIA 2002)*, volume 2424 of *LNCS*, pages 50–61. Springer, 2002.
2. J. J. Alferes, J. A. Leite, L. M. Pereira, H. Przymusinska, and T. Przymusinski. Dynamic updates of non-monotonic knowledge bases. *Journal of Logic Programming*, 45(1-3):43–70, 2000.
3. J. J. Alferes and L. M. Pereira. *Reasoning with Logic Programming*, volume 1111 of *LNAI*. Springer, Berlin, 1996.
4. J. J. Alferes, L. M. Pereira, and T. Swift. Abduction in well-founded semantics and generalized stable models via tabled dual programs. *Theory and Practice of Logic Programming*, 4(4):383–428, 2004.
5. M. Anderson and S. L. Anderson. EthEl: Toward a principled ethical eldercare robot. In *Procs. AAAI 2008 Fall Symposium on AI in Eldercare*, 2008.
6. M. Anderson, S. L. Anderson, and C. Armen. MedEthEx: a prototype medical ethics advisor. In *Procs. 18th Innovative Applications of Artificial Intelligence Conference (IAAI 2006)*, 2006.
7. 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*. Hackett, 1988.
8. C. Baral and M. Hunsaker. Using the probabilistic logic programming language P-log for causal and counterfactual reasoning and non-naive conditioning. In *Procs. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.
9. N. Bidoit and C. Froidevaux. General logic databases and programs: default logic semantics and stratification. *Journal of Information and Computation*, 91(1):15–54, 1991.
10. S. Bringsjord, K. Arkoudas, and P. Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44, 2006.

11. R. M. J. Byrne. *The Rational Imagination: How People Create Alternatives to Reality*. MIT Press, Cambridge, MA, 2007.
12. H. Chockler and J. Y. Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.
13. J. Collins, N. Hall, and L. A. Paul, editors. *Causation and Counterfactuals*. MIT Press, Cambridge, MA, 2004.
14. P. Dell'Acqua and L. M. Pereira. Preferential theory revision. *Journal of Applied Logic*, 5(4):586–601, 2007.
15. E.-A. Dietz, S. Hölldobler, and L. M. Pereira. On indicative conditionals. In *Procs. 1st International Workshop on Semantic Technologies (IWOST)*, volume 1339 of *CEUR Workshop Proceedings*, 2015.
16. K. Epstude and N. J. Roese. The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2):168–192, 2008.
17. M. Fitting. A Kripke-Kleene semantics for logic programs. *Journal of Logic Programming*, 2(4):295–312, 1985.
18. P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967.
19. J.-G. Ganascia. Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, 9(1):39–47, 2007.
20. M. Gelfond. On stratified autoepistemic theories. In *Procs. 6th National Conference on Artificial Intelligence (AAAI)*, 1987.
21. M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. In *Procs. 5th International Logic Programming Conference*. MIT Press, 1988.
22. M. L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(1):35–79, 1986.
23. Paul Grice. *Studies in the Way of Words*. Harvard University Press, Cambridge, MA, 1991.
24. J. Y. Halpern and C. Hitchcock. Graded causation and defaults. *British Journal for the Philosophy of Science*, 66:413–457, 2015.
25. T. A. Han, A. Saptawijaya, and L. M. Pereira. Moral reasoning under uncertainty. In *Procs. 18th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR)*, volume 7180 of *LNCS*, pages 212–227. Springer, 2012.
26. M. Hauser, F. Cushman, L. Young, R. K. Jin, and J. Mikhail. A dissociation between moral judgments and justifications. *Mind and Language*, 22(1):1–21, 2007.
27. M. Hewings. *Advanced Grammar in Use with Answers: A Self-Study Reference and Practice Book for Advanced Learners of English*. Cambridge University Press, New York, NY, 2013.
28. C. Hoerl, T. McCormack, and S. R. Beck, editors. *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*. Oxford University Press, Oxford, UK, 2011.
29. S. Hölldobler and C. D. P. Kencana Ramli. Logic programs under three-valued Łukasiewicz semantics. In *Procs. 25th International Conference on Logic Programming (ICLP)*, volume 5649 of *LNCS*, pages 464–478. Springer, 2009.
30. A. Kakas, R. Kowalski, and F. Toni. Abductive logic programming. *Journal of Logic and Computation*, 2(6):719–770, 1992.
31. F. M. Kamm. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford University Press, Oxford, UK, 2006.
32. M. Kleiman-Weiner, T. Gerstenberg, S. Levine, and J. B. Tenenbaum. Inference of intention and permissibility in moral decision making. In *Procs. 37th Annual Conference of the Cognitive Science Society*, 2015.
33. R. Kowalski. *Computational Logic and Human Thinking: How to be Artificially Intelligent*. Cambridge University Press, New York, NY, 2011.
34. D. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.

35. K. D. Markman, I. Gavanski, S. J. Sherman, and M. N. McMullen. The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology*, 29:87–109, 1993.
36. R. McCloy and R. M. J. Byrne. Counterfactual thinking about controllable events. *Memory and Cognition*, 28:1071–1078, 2000.
37. T. McCormack, C. Frosch, and P. Burns. The relationship between children's causal and counterfactual judgements. In C. Hoerl, T. McCormack, and S. R. Beck, editors, *Understanding Counterfactuals, Understanding Causation*. Oxford University Press, Oxford, UK, 2011.
38. A. McIntyre. Doctrine of double effect. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information, Stanford University, Fall 2011 edition, 2004. `http://plato.stanford.edu/archives/fall2011/entries/double-effect/`.
39. S. Migliore, G. Curcio, F. Mancini, and S. F. Cappa. Counterfactual thinking in moral judgment: an experimental study. *Frontiers in Psychology*, 5:451, 2014.
40. M. Otsuka. Double effect, triple effect and the trolley problem: Squaring the circle in looping cases. *Utilitas*, 20(1):92–110, 2008.
41. J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, MA, 2009.
42. L. M. Pereira, J. N. Aparício, and J. J. Alferes. Counterfactual reasoning based on revising assumptions. In *Procs. International Symposium on Logic Programming (ILPS 1991)*, pages 566–577. MIT Press, 1991.
43. L. M. Pereira, J. N. Aparício, and J. J. Alferes. Hypothetical reasoning with well founded semantics. In *Procs. 3rd Scandinavian Conference on Artificial Intelligence*. IOS Press, 1991.
44. L. M. Pereira, P. Dell'Acqua, A. M. Pinto, and G. Lopes. Inspecting and preferring abductive models. In K. Nakamatsu and L. C. Jain, editors, *The Handbook on Reasoning-Based Intelligent Systems*, pages 243–274. World Scientific Publishers, 2013.
45. L. M. Pereira, E.-A. Dietz, and S. Hölldobler. An abductive counterfactual reasoning approach in logic programming. Available from `http://goo.gl/bx0mIZ`, 2015.
46. L. M. Pereira and A. Saptawijaya. Modelling Morality with Prospective Logic. In M. Anderson and S. L. Anderson, editors, *Machine Ethics*, pages 398–421. Cambridge U. P., 2011.
47. T. M. Powers. Prospects for a Kantian machine. *IEEE Intelligent Systems*, 21(4):46–51, 2006.
48. H. Przymusinska and T. C. Przymusinski. Semantic issues in deductive databases and logic programs. In *Formal Techniques in Artificial Intelligence: A Sourcebook*, pages 321–367. North-Holand, 1990.
49. T. C. Przymusinski. Every logic program has a natural stratification and an iterated least fixed point model. In *Procs. 8th ACM Symposium on Principles Of Database Systems (PODS)*, pages 11–21, 1989.
50. T. C. Przymusinski. Three-valued non-monotonics formalisms and logic programming. In *Procs. 1st International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 1989.
51. N. J. Roese. Counterfactual thinking. *Psychological Bulletin*, 121(1):133–148, 1997.
52. A. Saptawijaya and L. M. Pereira. Incremental tabling for query-driven propagation of logic program updates. In *Procs. 19th International Conference on Logic Programming, Artificial Intelligence and Reasoning (LPAR)*, volume 8312 of *LNCS*, pages 694–709. Springer, 2013.
53. A. Saptawijaya and L. M. Pereira. Joint tabling of logic program abductions and updates (Technical Communication of ICLP 2014). *Theory and Practice of Logic Programming, Online Supplement*, 14(4-5), 2014. Available from `http://arxiv.org/abs/1405.2058`.

54. A. Saptawijaya and L. M. Pereira. Towards modeling morality computationally with logic programming. In *PADL 2014*, volume 8324 of *LNCS*, pages 104–119. Springer, 2014.
55. T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, Cambridge, MA, 1998.
56. T. M. Scanlon. *Moral Dimensions: Permissibility, Meaning, Blame*. Harvard University Press, Cambridge, MA, 2008.
57. T. Swift and D. S. Warren. XSB: Extending Prolog with tabled logic programming. *Theory and Practice of Logic Programming*, 12(1-2):157–187, 2012.
58. P. E. Tetlock, P. S. Visser, R. Singh, M. Polifroni, A. Scott, S. B. Elson, P. Mazzocco, and P. Rescober. People as intuitive prosecutors: the impact of social-control goals on attributions of responsibility. *Journal of Experimental Social Psychology*, 43:195–209, 2007.
59. J. J. Thomson. The trolley problem. *The Yale Law Journal*, 279:1395–1415, 1985.
60. M. H. van Emden and R. Kowalski. The semantics of predicate logic as a programming language. *Journal of the ACM*, 4(23):733–742, 1976.
61. A. van Gelder, K. A. Ross, and J. S. Schlipf. The well-founded semantics for general logic programs. *Journal of the ACM*, 38(3):620–650, 1991.
62. J. Vennekens, M. Bruynooghe, and M. Denecker. Embracing events in causal modeling: Interventions and counterfactuals in CP-logic. In *JELIA 2010*, volume 6341 of *LNCS*, pages 313–325. Springer, 2010.
63. B. Weiner. *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. The Guilford Press, New York, NY, 1995.