Agent Morality via Counterfactuals in Logic Programming

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

[1] NOVA-LINCS, 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 presents a computational model, via Logic Programming (LP), of counterfactual reasoning with applications to agent morality. Counterfactuals are conjectures about what would have happened, had an alternative event occurred. In the first part, we show how counterfactual reasoning, inspired by Pearl's structural causal model of counterfactuals, is modeled using LP, by benefiting from LP abduction and updating. In the second part, 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. Finally, we discuss some potential extensions of our LP approach to cover other aspects of counterfactual reasoning and show how these aspects are relevant in modeling agent morality.

Keywords: abduction, counterfactuals, 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 [5]: *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 [6, 19], psychology [5, 21, 31]. They also have been studied from the computational viewpoint [4, 11, 26, 27, 39], where approaches in Logic Programming (LP), e.g., [4, 27, 39], are mainly based on probabilistic reasoning.

In the first part of this paper, we report on our approach of using LP abduction and updating in a procedure for evaluating counterfactuals, taking the established approach of Pearl [26] as reference. 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. Our approach therefore 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 [38].

Counterfactual thinking in moral reasoning has been investigated particularly via psychology experiments (see, e.g., [9, 21]), but it has only been limitedly explored in machine ethics. In the second part of the paper, 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 [13, 18, 29, 32], 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.

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 scrutinized herein through several off-the-shelf classic moral examples from the literature. By materializing these doctrines in concrete moral dilemmas, the results of counterfactual evaluation –supported by our LP approach– are readily comparable to those from the literature. 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.

In the final part of the paper, we discuss some potential extensions of our LP approach to cover other aspects of counterfactual reasoning. These aspects include *assertive counterfactuals*, extending the antecedent of a counterfactual with a LP rule, and abducing the antecedent of a counterfactual in the form of intervention. These aspects are relevant in modeling agent morality, which opens the way for further research towards employing LP-based counterfactual reasoning to machine ethics.

2 Abduction in Logic Programming

We start by recapping basic notation in LP and review how abduction is expressed and computed in LP.

A *literal* is either an atom L or its default negation $not\ L$, named positive and negative literals, respectively. They are negation complements to each other. The atoms *true* and *false* are true and false, respectively, in every interpretation. A *logic program* is a set of rules of the form $H \leftarrow B$, naturally read as "H if B", where its *head* H is an atom and its (finite) *body* B is a sequence of literals. When B is empty (equal to $true$), the rule is called a *fact* and simply written H. A rule in the form of a denial, i.e., with *false* as head, is an *integrity constraint*.

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 its negation complement Ab^* (denotes $not\ Ab$), whose truth value is not initially assumed, and it does not appear in the head of a rule. 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 such that there is no rule in P whose head is in $\mathcal{A}$, and $\mathcal{I}$ is a set of integrity constraints.

An observation O is a set of literals, analogous to a query or goal in LP. Abducing an explanation for O amounts to finding *consistent abductive solutions* $S \subseteq \mathcal{A}$ to a goal O, whilst satisfying the integrity constraints, where abductive solutions S entail O is true in the semantics obtained after replacing the abducibles S in P by their abduced truth value. Abduction in LP can be accomplished by a *top-down query-oriented procedure* for finding a query's abductive 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 3-valued *Well-Founded Semantics* [38], employed by us, enjoys this relevancy property [8], 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, and remain undefined.

3 LP-based Counterfactuals

Our LP approach in evaluating counterfactuals is based Pearl's approach [26]. Therein, 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 consequent's probability.

Since each step of 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 approach to logic programs, but leaving out probabilities, employing instead LP abduction and updating to determine the logic validity of counterfactuals under Well-Founded Semantics.

3.1 Causation and Intervention in LP

Two important ingredients in Pearl's approach of counterfactuals are causal model and intervention. With respect to an abductive framework $\langle P, \mathcal{A}, \mathcal{I} \rangle$, observation O corresponds to Pearl's definition for evidence. 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 consists of set of background variables, whose values are conditional on case-considered observed evidences. These background variables are not causally explained in the model, as they have no parent nodes in the causal diagram of the model. In terms of LP abduction, they correspond to a set of abducibles $E \subseteq \mathcal{A}$ 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.

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.

3.2 Evaluating Counterfactuals in LP

The procedure to evaluate counterfactuals in LP essentially takes the three-step process of Pearl's approach as its reference. 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.[3] The observation corresponds to the evidence that both the antecedent and the consequence literals of the present counterfactual were, in the considered past moment, factually false.[4] For otherwise the counterfactual would be trivially true when making the antecedent false, or irrelevant for the aim of making the consequent true. 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 [7, 28]. The explanation fixes the abduced context in which the counterfactual is evaluated by updating the program with the explanation.

[3] 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 [12].

[4] This interpretation is in line with the corresponding English construct, cf. [15], commonly known as *third conditionals*.

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 Well-Founded Model (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 restorative new update with 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, our approach is defined below, abstracting from the above state transition detail. In the sequel, the Well-Founded Model of program P is denoted by $WFM(P)$. As our counterfactual procedure is based on the Well-Founded Semantics, the standard logical consequence relation $P \models F$ used below presupposes the Well-Founded Model of P in verifying the truth of formula F, i.e., whether F is true in $WFM(P)$.

Procedure 1. 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*:
   - 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. 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 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 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 $P \cup E_1$. 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, viz., "if there had not been lightning".
3. ***Prediction***: We verify that $(P \cup E_1)_{\tau,\iota} \models not\ f$. That is, $not\ f$ holds with respect to the intervened modified program for explanation $E_1 = \{s, b^*\}$ and the intervention $make\_not(l)$. Note, $\mathcal{I} = \emptyset$ 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 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. In this case, $(P \cup E_1)_\tau = (P \cup E_2)_\tau$, and the required causal intervention is also the same: $make\_not(l)$. But we now have $(P \cup E_2)_{\tau,\iota} \not\models not\ f$. Indeed, the forest fire would still have occurred but due to an alternative cause: barbecue.

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 [21] and the influence of counterfactual thoughts in moral judgment [24]. As argued in [9], 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 [40].

The potential of LP to machine ethics has been reported in [13, 18, 29] and with emphasis on LP abduction and updating in [32]. 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) [23] and of Triple Effect (DTE) [17].

DDE is first introduced by Thomas Aquinas in his discussion of the permissibility of self-defense [3]. 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 [23]. According to the Doctrine of Double Effect, the former action is permissible, whereas the latter is impermissible. In [14], 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 [10]: *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 [22], 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 [36] 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 [24]. The reader is referred to [6] and [16] 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 [35] – 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 3. 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 $(P_e \cup E_e)_{\tau,\iota} \models not\ ew$. 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 4. 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 $(P_a \cup E_a)_{\tau,\iota} \not\models not\ ew$. Therefore, the morally wrong kic is just a side-effect in achieving the goal ew. Hence tab is DDE morally permissible.

Example 5. 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: $Pre = not\ kic(a) \wedge not\ kic(b)$. Given $E_{ab} = \{teb\}$, one can easily verify that $(P_{ab} \cup E_{ab})_{\tau,\iota} \models not\ ew$, 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 [17] 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 [17] proposed DTE to accommodate a variant of the trolley problem, viz., the *Loop Case* [37]: *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 [25]. Referring to a psychology study [14], 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 [17]. 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 6. 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 $(P_o \cup E_o)_{\tau,\iota} \models not\ sav$. 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 7. Consider a variant of the Loop case, viz., the *Loop-Push Case* (see also Extra Push Case in [17]). 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: $(P_p \cup E_{p_1})_{\tau,\iota} \models not\ sav$.
- "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 $(P_p \cup E_{p_2})_{\tau,\iota} \models not\ hit$.

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 Extending LP-based Counterfactuals

Our approach, in Section 3, specifically focuses on evaluating counterfactuals in order to determine their validity. We identify some potential extensions of this LP-based approach to other aspects of counterfactual reasoning:

1. We consider the so-called *assertive counterfactuals*, where a counterfactual is given as being a valid statement, rather than a statement whose truth validity has to be determined. The causality expressed by such a valid counterfactual may be useful for refining an existing knowledge base. For instance, suppose we have a rule stating that the lamp is on if the switch is on, written as $lamp\_on \leftarrow switch\_on$. Clearly, providing the fact $switch\_on$, we have $lamp\_on$ true. Now consider that the following counterfactual is given as being a valid statement:

   "If the bulb had not functioned properly, then the lamp would not be on"

   There are two ways that this counterfactual may refine the rule about $lamp\_on$. First, the causality expressed by this counterfactual can be used to transform the rule into:

$$\begin{aligned} lamp\_on &\leftarrow switch\_on, bulb\_ok. \\ bulb\_ok &\leftarrow not\ make\_not(bulb\_ok). \end{aligned}$$

   So, the lamp will be on if the switch is on – that is still granted – but subject to an update $make\_not(bulb\_ok)$, which captures the condition of the bulb. In the other alternative, an assertive counterfactual is rather directly translated into an updating rule, and need not transform existing rules.
2. We may extend the antecedent of a counterfactual with a rule, instead of just literals. For example, consider the following program (assuming an empty abduction, so as

to focus on the issue):

$$\begin{aligned} warm\_blood(M) &\leftarrow mammal(M). \\ mammal(M) &\leftarrow dog(M). \\ mammal(M) &\leftarrow bat(M). \\ dog(d). \qquad & bat(b). \end{aligned}$$

Querying ?- $bat(B), warm\_blood(B)$ assures us that there is a warm blood bat, viz., $B = b$.
Now consider the counterfactual:

> "If bats were not mammals they would not have warm blood".

Transforming the above program using our procedure obtains:

$$\begin{aligned} warm\_blood(M) &\leftarrow mammal(M). \\ mammal(M) &\leftarrow make(mammal(M)). \\ mammal(M) &\leftarrow dog(M), not\ make\_not(mammal(M)). \\ mammal(M) &\leftarrow bat(M), not\ make\_not(mammal(M)). \\ dog(d). \qquad & bat(b). \end{aligned}$$

The antecedent of the given counterfactual can be expressed as the rule:

$$make\_not(mammal(B)) \leftarrow bat(B).$$

We can check using our procedure that, given this rule intervention, the above counterfactual is valid: $not\ warm\_blood(b)$ is true in the intervened modified program.

3. Finally, we can easily imagine the situation where the antecedent Pre of a counterfactual is not given, though the conclusion $Conc$ is, and we want to abduce Pre in the form of interventions. That is, the task is to abduce $make$ and $make\_not$, rather than imposing them, while respecting the integrity constraints, such that the counterfactual is valid.
Tabling abductive solutions [33] may be relevant in this problem. Suppose that we already abduced an intervention Pre_1 for a given $Conc_1$, and we now want to find Pre_2 such that the counterfactual "If Pre_1 and Pre_2 had been the case, then $Conc_1$ and $Conc_2$ would have been the case" is valid. In particular, when abduction is performed for a more complex conclusion $Conc_1$ and $Conc_2$, the solution Pre_1, which has already been abduced and tabled, can be reused in the abduction of such a more complex conclusion, leading to the idea that problems of this kind of counterfactual reasoning can be solved in parts or in a modular way.

Indeed, the above three aspects may have relevance in modeling agent morality:

1. In assertive counterfactuals, the causality expressed by a given valid counterfactual can be useful for refining moral rules, which can be achieved through incremental rule updating. This may further the application of moral updating and evolution.

2. The extension of a counterfactual with a rule antecedent opens up another possibility to express exceptions in moral rules. For instance, one can express an exception about lying, such as "If lying had been done to save an innocent from a murderer, then it would not have been wrong". That is, given a knowledge base about lying for human H:

$$lying\_wrong(H) \leftarrow lying(H), not\ make\_not(lying\_wrong(H)).$$

   The antecedent of the above counterfactual can be represented as a rule:

$$make\_not(lying\_wrong(H)) \leftarrow save\_from\_murderer(H, I), innocent(I).$$

3. Given that the conclusion of a counterfactual is some moral wrong W, abducing its antecedent in the form of intervention can be used for expressing a prevention of W, viz., "What could I have done to prevent a wrong W?".

6 Concluding Remarks

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.

The three potential extensions of our LP approach to cover other aspects of counterfactual reasoning, as well as their applications to machine ethics, are worth exploring in future. Apart from these identified extensions, our present LP-based approach for evaluating 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* [20], may provide moral justification for what was done due to lack, at the time, 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 not morally superior than the actual ensued consequence. We have started, in [30], to explore the application of our present LP-based approach to evaluate compound counterfactuals for moral justification. Further application of compound counterfactuals, to justify an exception for an action to be permissible, that may lead to agents' argumentation following Scanlon's contractualism [34], is another path of future investigation.

Acknowledgements

Luís Moniz Pereira acknowledges the support from Fundação para a Ciência e a Tecnologia (FCT/MEC) NOVA LINCS PEst UID/CEC/04516/2013.

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. 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.
4. 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.
5. R. M. J. Byrne. *The Rational Imagination: How People Create Alternatives to Reality*. MIT Press, Cambridge, MA, 2007.
6. J. Collins, N. Hall, and L. A. Paul, editors. *Causation and Counterfactuals*. MIT Press, Cambridge, MA, 2004.
7. P. Dell'Acqua and L. M. Pereira. Preferential theory revision. *Journal of Applied Logic*, 5(4):586–601, 2007.
8. J. Dix. A classification theory of semantics of normal logic programs: II. weak properties. *Fundamenta Informaticae*, 3(22):257–288, 1995.
9. K. Epstude and N. J. Roese. The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2):168–192, 2008.
10. P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967.
11. M. L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(1):35–79, 1986.
12. Paul Grice. *Studies in the Way of Words*. Harvard University Press, Cambridge, MA, 1991.
13. 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.
14. 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.
15. 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.
16. C. Hoerl, T. McCormack, and S. R. Beck, editors. *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*. Oxford University Press, Oxford, UK, 2011.
17. F. M. Kamm. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford University Press, Oxford, UK, 2006.
18. R. Kowalski. *Computational Logic and Human Thinking: How to be Artificially Intelligent*. Cambridge University Press, New York, NY, 2011.
19. D. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.
20. 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.
21. R. McCloy and R. M. J. Byrne. Counterfactual thinking about controllable events. *Memory and Cognition*, 28:1071–1078, 2000.
22. 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.

23. 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/`.
24. S. Migliore, G. Curcio, F. Mancini, and S. F. Cappa. Counterfactual thinking in moral judgment: an experimental study. *Frontiers in Psychology*, 5:451, 2014.
25. M. Otsuka. Double effect, triple effect and the trolley problem: Squaring the circle in looping cases. *Utilitas*, 20(1):92–110, 2008.
26. J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, MA, 2009.
27. 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.
28. 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.
29. 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.
30. L. M. Pereira and A. Saptawijaya. *Programming Machine Ethics*, volume 26 of *Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE)*. Springer, 2016.
31. N. J. Roese. Counterfactual thinking. *Psychological Bulletin*, 121(1):133–148, 1997.
32. 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.
33. A. Saptawijaya and L. M. Pereira. TABDUAL: a tabled abduction system for logic programs. *IfCoLog Journal of Logics and their Applications*, 2(1):69–123, 2015.
34. T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, Cambridge, MA, 1998.
35. T. M. Scanlon. *Moral Dimensions: Permissibility, Meaning, Blame*. Harvard University Press, Cambridge, MA, 2008.
36. 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.
37. J. J. Thomson. The trolley problem. *The Yale Law Journal*, 279:1395–1415, 1985.
38. 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.
39. 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.
40. B. Weiner. *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. The Guilford Press, New York, NY, 1995.