From Logic Programming to Machine Ethics

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

[1] Faculty of Computer Science, Universitas Indonesia, Indonesia
[2] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS),
Departamento de Informática, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, Portugal

Abstract. This chapter investigates the appropriateness of Logic Programming-based reasoning to machine ethics, an interdisciplinary field of inquiry that emerges from the need of imbuing autonomous agents with the capacity for moral decision making. The first part of the chapter aims at identifying morality viewpoints, as studied in moral philosophy and psychology, which are amenable to computational modeling, and then mapping them to appropriate Logic Programming-based reasoning features. The identified viewpoints are covered by two morality themes: moral permissibility and the dual-process model. In the second part, various Logic Programming-based reasoning features are applied to model these identified morality viewpoints, via classic moral examples taken off-the-shelf from the literature. For this purpose, our QUALM system mainly employs a combination of the Logic Programming features of abduction, updating, and counterfactuals. These features are all supported jointly by Logic Programming tabling mechanisms. The applications are also supported by other existing Logic Programming-based systems, featuring preference handling and probabilistic reasoning, which complement QUALM in addressing the morality viewpoints in question.
Throughout the chapter, many references to our published work are given, providing further examples and details about each topic. Thus, this chapter can be envisaged as an entry point survey on the employment of Logic Programming for knowledge modelling and technically implementing machine ethics.

Keywords: Logic Programming; Machine Ethics; Abduction; Updating; Counterfactuals.

1 Introduction

The need for systems or agents that can function in an ethically responsible manner is becoming a pressing concern, as they become ever more sophisticated, autonomous, and act in groups, amidst populations of other agents, including humans. Its importance has been emphasized as a research priority in artificial intelligence [39] with funding supports [18, 47], in scientific meetings (e.g., [5, 6]), book publications (e.g., [4, 36, 50, 51]), as well as a heightened public awareness to its economic import [10]. The field —bringing together perspectives from various fields, including philosophy, psychology, anthropology, evolutionary biology, and artificial intelligence— is not just important for equipping agents with some capacity for moral decision-making, but also to help better understand morality, via the creation and testing of computational models of ethical theories.

Several logic-based formalisms have been employed to model moral theories or particular morality aspects, e.g., deontic logic in [7, 38, 52], non-monotonic reasoning in [13, 38], and the use of Inductive Logic Programming in [3]. Some of these works only abstractly address the utilization of logic-based formalisms (e.g., [38]), whereas others also provide implementations (e.g., using Inductive Logic Programming-based systems [3], an interactive theorem prover [7], an agent programming language [52], and answer set programming [13]). Despite the aforementioned logic-based formalisms, the usage of Logic Programming —which provides a logic-based programming paradigm supported by a number of reasoning features and practical systems— is rather limited. The potential and suitability of Logic Programming, and of computational logic in general, for machine ethics, is identified in [21, Section 12],

on the heels of our work. This chapter is a condensed report of our recent book [36]. The book discusses at length further original and integrative investigation on the appropriateness of various Logic Programming-based reasoning to machine ethics, not just abstractly, but also furnishing a proof of concept implementation for the morality issues in hand.

The first part of this chapter aims at identifying conceptual morality viewpoints and mapping them into appropriate Logic Programming-based reasoning features. The identified viewpoints are covered in two morality themes: (1) *moral permissibility*, taking into account viewpoints such as the Doctrines of Double Effect [28], Triple Effect [20], and Scanlon's contractualist moral theory [44]; and (2) *the dual-process model* [8, 25], which stresses the interaction between deliberative and reactive behaviors in moral judgment, from both the viewpoints of competing and collaborative interactions. The mapping of all these considered viewpoints into Logic Programming-based reasoning benefits from its features and their integration, such as abduction with integrity constraints [2, 19, 41, 43], preferences over abductive scenarios [9], probabilistic reasoning [15], updating [1, 40], counterfactuals [37], and from Logic Programming tabling technique [46] as well as Logic Programming semantics: either Stable Model [14] or Well-Founded Model [49] semantics.

In the second part, the various Logic Programming-based reasoning features are employed to model the aforementioned morality viewpoints, reifying the discussion of the first part of this chapter through a variety of classic moral examples taken off-the-shelf from the literature, as proof of ability. For this purpose, our QUALM system mainly employs a combination of the Logic Programming features of abduction, updating, and counterfactuals [37, 40, 42, 43]. These features are all supported jointly by Logic Programming tabling mechanisms. The applications are also supported by other existing Logic Programming-based systems featuring preference handling and probabilistic reasoning, which complement QUALM in experimenting with the morality viewpoints. The deployments of these systems include: (1) The use of a priori integrity constraints and a posteriori preferences over abductive scenarios to capture deontological and utilitarian judgments, resp., within a competing interaction viewpoint of the dual-process model [8]; (2) Probabilistic moral reasoning, to reason about actions, under uncertainty, that might have occurred, and thence provide judgment adhering to moral principles within some prescribed uncertainty level. This permits to capture a form of argumentation (with respect to Scanlon's contractualism [44]) in courts, through presenting different evidences as a consideration whether an exception can justify a verdict of guilty (beyond reasonable doubt) or non-guilty; (3) The use of QUALM to examine moral permissibility and its justification, with respect to the Doctrines of Double Effect and Triple Effect, via counterfactual queries. Finally, QUALM is also employed to experiment with the issue of moral updating, allowing for other (possibly overriding) moral rules (themselves possibly subsequently overridden) to be adopted by an agent, on top of those it currently follows.

Note that the choice of morality viewpoints in this chapter neither aims at defending them nor resolving their related moral dilemmas, as even philosophers commonly split opinions over them. Instead, its purpose is to show that diverse Logic Programming-based reasoning features are capable and appropriate for expressing, in combination, different viewpoints on the morality themes tackled —as demonstrated by the prototypes under experimentation— with results conforming to those argued in the literature.

The rest of the chapter is organized as follows. Section 2 reports on our background relevant literature study in philosophy and psychology for choosing conceptual viewpoints amenable to Logic Programming-based reasoning. These viewpoints are mapped into Logic Programming-based reasoning features in Section 3. Section 4 discusses how these features are empowered to model the chosen morality viewpoints. Section 5 concludes the chapter with discussion and potential future work.

2 Moral Permissibility and the Dual-Process Model

Moral Permissibility The Doctrine of Double Effect is often invoked to explain the permissibility of an action that causes a harm by distinguishing whether this harm is a mere *side-effect* of bringing about a

good result, or rather a *means* to bringing about the same good end [28]. The Doctrine of Double Effect has been utilized, in [17], to explain the consistency of judgments, shared by subjects from demographically diverse populations, on a series of moral dilemmas developed from the classic trolley problem [12]. These dilemmas inquire whether it is permissible to harm one or more individuals for the purpose of saving others. The original trolley problem, often identified as the *Bystander Case*, concerns the permissibility of diverting the trolley from the main track to a parallel side track, on which a man is standing, for saving five people walking on the main track. The dilemma thus concerns the permissibility to kill the man while saving the five. Other dilemmas of the problem are obtained by adapting this case. In the *Footbridge Case*, there exists only the main track but with a footbridge over it, on which a heavy man stands. The dilemma concerns the permissibility to shove him onto the track for stopping the trolley, consequently saving the five. Though both cases aim at saving the five, albeit killing one, they differ in the Doctrine of Double Effect permissibility of their corresponding actions.

Another principle, related to the Doctrine of Double Effect, is the Doctrine of Triple Effect [20] that refines the Doctrine of Double Effect on the notion about harming someone as an intended means. The Doctrine of Triple Effect distinguishes further between doing an action *in order* that an effect occurs and doing it *because* that effect will occur. Though the Doctrine of Triple Effect also classifies the former as impermissible, it is more tolerant to the latter (the third effect): it treats as permissible those actions performed just and only *because* instrumental harm will occur. The Doctrine of Triple Effect is proposed to accommodate the *Loop Case* [48] of the trolley problem. In this dilemma, the trolley can be redirected onto a side track, which *loops* back towards the five. However, a heavy man sits on this looping side track, so that his body will by itself stop the trolley. The question is whether it is permissible to divert the trolley to the looping side track, thereby killing him, but saving the five. This case strikes most philosophers that diverting the trolley is permissible [31]. To this end, the Doctrine of Triple Effect may provide the justification: it is permissible because it will hit the man, but not in order to intentionally kill him [20]. Nonetheless, the Doctrine of Double Effect views diverting the trolley in this case as impermissible.

The philosopher T. M. Scanlon has developed a distinctive view of moral reasoning called *contractualism* [44], whereby moral permissibility is differently addressed through so-called *deliberative* employing of moral judgments, i.e., the question of the permissibility of actions is answered by identifying the justified though defeasible argumentative considerations, and their exceptions. It is based on a view that moral dilemmas typically share the same structure: they concern general principles that in some cases admit exceptions, and when those exceptions generally apply. This deliberative employment carries three important features: (1) it regards the importance of principles to provide reason for justifying an action to others and flexibility on choosing the principles; (2) reasoning is an important aspect in contractualism; and (3) as reasoning becomes a primary concern in contractualism for providing justification to others, it is achieved by looking for some common ground that others could not reasonably object to.

The Dual-Process Model Psychology research reveals that moral judgment is driven by two systems; a view known as the *dual-process* model [8], or Type I and II processing [45]. The first one, the cognitive system, operates by *controlled* psychological processes, where explicit moral principles are consciously applied via deliberative reasoning. The other, the affective system, operates by *automatic* processes, not entirely accessible to conscious reflection, where moral judgment is intuition-based and more low-level.

The dual-process model is evidenced by psychological experiments in moral dilemmas like those from the trolley problem. The experiments characterize each system with respect to applicable moral principles in these dilemmas: a general principle favoring welfare-maximizing behaviors (cf. utilitarian judgment in the Bystander case) appears to be supported by controlled cognitive processes, whereas that prohibiting the use of harm as a means to a greater good (cf. deontological judgment in the Footbridge case) appears to be part of the process that generates intuitive emotional responses [8, 45]. The dual-process model therefore exhibits a form of *competing interaction* between the two processes.

A *collaborative interaction* between the two systems in producing moral judgment is discussed in [25]. Whereas the dual-process model of [8] supports the view that cognitive processes are not associated with deontological (non-utilitarian) judgment, the collaborative view defends that both reasoning on moral rules and emotion (the cognitive and the affective systems, resp.) work together to produce non-utilitarian judgment, like in the Footbridge case. This study is based on the view that moral judgment is supported by internally represented rules and reasoning about whether particular cases fall under those rules, even in deontological (non-utilitarian) judgment, where the affective system dominates, as claimed in [8]. It asserts that, though several studies demonstrate that people experience difficulty in justifying moral judgment generated by rapid and automatic processes (of the affective system), moral rules may still play an important role without the reasoning process being consciously accessible. It thus provides an explanation that despite this difficulty, moral judgment driven by the affective system mirrors and is consistent with a particular moral rule, e.g., the deontological judgment in the Footbridge case is consistent with The Doctrine of Double Effect, and so is the utilitarian judgment in the Bystander case.

3 Mapping Morality Viewpoints to Logic Programming-based Reasoning Features

We start by defining logic programs and necessary notation used in the sequel.

A *logic program* is a set of rules (predicate logic formulas of *implication*) of the form $H \leftarrow B$, read 'H is true, *if* B is true'. Here, B is generally a *conjunction* of several conditions, separated by comma. Variables, written in capital letters, appearing in a rule are quantified universally. For example:

$$father(X, Z) \leftarrow married(X, Y), mother\_of(Y, Z).$$

is a rule expressing 'for all X, Y and Z: X is the father of Z, if X is married to Y and Y is the mother of Z. Note that a rule is ended with a dot (.), as shown by the above example.

When B is empty, the rule is called a *fact* and simply written H (viz., H is unconditionally true).

Abduction It is a reasoning method whereby one chooses from available hypotheses those that best explain the observed evidence, in a preferred sense. We refer to an abductive logic programming framework [19] that comprises a logic program, a set of abducibles (available hypotheses), and a set of integrity constraints, where an integrity constraint is a rule in the form of a denial, viz., with *false* as its conclusion. An observation in abduction is analogous to a query in Logic Programming. In abductive logic programming-based agents, abduction amounts to finding consistent abductive solutions to a goal, whilst satisfying the integrity constraints.

Abduction (with its integrity constraints) is appropriate for dealing with morality themes discussed in this chapter. First, in moral dilemmas, like those from the trolley problem, abducibles may represent available actions, e.g., diverting the trolley, shoving the heavy man, etc. They are abduced to satisfy a given goal and integrity constraints, which reflect moral considerations in a modeled dilemma. Second, integrity constraints may serve a couple of roles: (a) with respect to moral permissibility, they can be used for ruling out the doctrines of double effect-/triple effect-impermissible actions a priori; (b) with respect to the dual-process model (in the sense of [8]), they can be viewed as a mechanism to generate satisfactory intuitive emotional responses in deontological judgment, in prohibiting the use of harm for a greater good. That is, a priori integrity constraints provide agent's reactive behaviors in delivering this kind of judgment, in contrast with utilitarian judgment that requires more involved reasoning, achieved via a posteriori preferences, as described below. Third, abduction may be used to hypothesize incomplete information to explain observations. This role is particularly important for providing the "other things being equal" background context in counterfactual reasoning, whose applications in moral permissibility (and its justification) are explored elsewhere in this chapter.

Preferences over Abductive Scenarios In abduction, the hypotheses generation and integrity constraints a priori exclude irrelevant abducibles with respect to the agent's actual situation. Abductive stable models [9], say, can then be computed for these relevant abducibles, and *a posteriori* preferences can subsequently be enacted, by inspecting consequences of the abductive solutions in the obtained abductive stable models. The evaluation can be done quantitatively (e.g. by utility functions) or qualitatively (e.g. by enforcing some rules over consequences to hold). Here, a posteriori preferences are appropriate to capture utilitarian judgment that favors welfare-maximizing behaviors, supported by controlled cognitive processes from the dual-process model [8]. Reasoning with a posteriori preferences can be viewed as a form of controlled cognitive processes (rather than intuitive ones) as it is achieved via more involved reasoning: specific consequences of the considered abductive solutions have first to be computed and only then are they evaluated to prefer the solution affording the greater good.

Probabilistic Logic Programming It is often the case that one has to pass a moral judgment on a situation without actually observing it thoroughly. In order to deal with such uncertainty, probabilistic logic programming is employed as part of abduction. It allows probabilistically abducing moral decisions by reasoning about actions with respect to the availability of observed evidence and its attending truth value. Such reasoning is relevant, e.g., in courts, where juries may be required to proffer rulings beyond a reasonable doubt.

This use of probabilistic logic programming in court rulings is appropriate to capture the deliberative employment, in line with Scanlon's contractualism, where permissibility of actions is addressed through justified but defeasible argumentative considerations. By insisting on a probability standard of proof beyond reasonable doubt (cf. [30]) as common agreed ground for the verdict of guilty to be qualified as 'beyond reasonable doubt', argumentation may take place through presentation of diverse strength evidence (via Logic Programming updating, to be described below) as a consideration to justify exceptions. Whether such evidence is accepted as a justification (defeating the formerly presented evidence) depends on its influence on the probability of action and intent, which in turn determines its permissibility and thereby the verdict. That is, it all depends on whether this probability is still within the agreed standard of proof beyond reasonable doubt, given that moral permissibility of actions is couched in court terms as verdicts about guilt and non-guilt.

Logic Programming Updating Concomitantly to abduction, an agent may learn new information from the external world or update itself of its own accord in order to pursue present goal. It is therefore natural to accommodate Logic Programming abduction with updating. Logic Programming updating facilitates providing justification to an exception with respect to a moral principle. It allows modeling Scanlon's deliberative employing of moral judgment for abducing permissible actions, involving defeasible argumentation and considerations to justify exception. Three applications of Logic Programming updating are relevant in this respect: (1) In the aforementioned court case: for updating jury's knowledge with new evidence that may defeat former ones; (2) In some trolley problem cases: for updating new information that may support the Doctrine of Triple Effect as a principle to justify an exception, rendering an impermissible action (according to Doctrine of Double Effect) permissible (according to the Doctrine of Triple Effect); (3) In the setting of moral updating (the adoption of new, possibly overriding, moral rules on top of those an agent currently follows): the updating moral rules can be viewed as an exception to the current ones. Such updating is necessary when the currently followed moral rules have to be revised, or qualified by overriding exceptions, in the light of situations faced by the agent.

Counterfactuals Counterfactuals capture the process of reasoning about a past event that did not occur, namely what would have happened had this event occurred. They have been investigated in the context of moral reasoning via psychology experiments, e.g., [11, 27, 29], but have been only limitedly explored in machine ethics. In our work, counterfactual reasoning is employed to examine moral permissibility of

actions according to the Doctrines of Double Effect and Triple Effect, achieved by distinguishing between *cause* and *side-effect* as a result of performing an action to achieve a goal. While moral permissibility can be assessed with respect to the Doctrines of Double Effect and Triple Effect using abduction with integrity constraints and a posteriori preferences, the use of counterfactuals provides a different approach to examine moral permissibility. Moreover, counterfactuals permit to justify moral permissibility: (1) In the form of compound counterfactuals for justifying with hindsight what was done in the past, in the absence of current knowledge; (2) In the spirit of Scanlon's contractualism, by providing a conceptual counterfactual query for justifying exception to permissibility of actions.

Tabling Tabling affords solutions reuse, rather than recomputing them, by keeping in tables subgoals and their answers obtained from query evaluation. Given this reuse benefit, tabling seems appropriate for capturing low-level reactive behavior, by relying on the Logic Programming system-level for obtaining solutions from tables rather than deliberatively recomputing them at all times. This feature is thus close to intuition-based psychological processes in the dual-process model that permit rapid and automatic moral judgment. Tabling may benefit Logic Programming abduction and updating: (1) Tabling in contextual abduction [41, 43] permits reusing priorly obtained abductive solutions, from one abductive context to another (insofar keeping the solutions consistent); and (2) Incremental tabling for Logic Programming updating [40] helps realize the causal intervention of counterfactuals [37], via the temporary updates to enact the hypothetical alternative former event.

The roles of tabling a propos morality themes in this chapter are as follows. First, given that abductive solutions represent some actions according to a specific moral principle, tabling in abduction allows an agent to deliver an action in a compatible context without repeating deliberative reasoning, thus establishing a form of low-level reactive behavior (realized by system-level tabling) pertaining to the dual-process model. In this case, though being only retrieved from the table, albeit upwards propagated incrementally, such reactive decisions in the compatible context are consistent with the specific moral principle followed during the deliberative reasoning when these decisions were initially computed. Second, an agent may obtain new information while making a moral decision. In such a dynamic situation, the agent may later be required to achieve new goals in addition to the former ones, due to a moral principle it now follows. While achieving these new goals requires deliberative reasoning, the decisions abduced for former goals can immediately be retrieved from the table and subsequently involved in the deliberative reasoning for the new goals via contextual abduction. It thus provides a computational model of the collaborative interaction between deliberative and reactive reasoning in the dual-process model.

4 Modeling Morality with Logic Programming

4.1 Moral Permissibility with Abduction, a Priori Integrity Constraints and a Posteriori Preferences

In [34], moral permissibility is modeled through several cases of the classic trolley problem [12] (three of them: the Bystander, the Footbridge, and the Loop cases, are described in Section 2), by emphasizing the use of integrity constraints in abduction and preferences over abductive scenarios, using ACORDA [22]. The cases, which include moral principles, are modeled in order to deliver appropriate moral decisions via reasoning. By appropriate moral decisions we mean ones that conform with those the majority of people make, on the basis of empirical results in [17]. Therein, the Doctrine of Double Effect [28] is utilized to explain the consistency of judgments, shared by subjects from demographically diverse populations, on a series of trolley dilemmas. In addition to the Doctrine of Double Effect, the Doctrine of Triple Effect [20] is also considered in [34].

Each case of the trolley problem is modeled individually; their details being referred to [34]. The key points of their modeling are as follows. The doctrines of double effect and triple effect are modeled via

a priori integrity constraints and a posteriori preferences. Possible decisions are modeled as abducibles, encoded in ACORDA by even loops over default negation. Moral decisions are therefore accomplished by satisfying a priori integrity constraints, computing abductive stable models from all possible abductive solutions, and then appropriately preferring amongst them (by means of rules), a posteriori, just some models, on the basis of their abductive solutions and consequences. Such preferred models turn out to conform with the results reported in the literature.

Capturing Deontological Judgment via a Priori Integrity Constraints In this application, integrity constraints are used for two purposes. First, they are utilized to force the goal in each case (like in [17]), by observing the desired end goal resulting from each possible decision. Such an integrity constraint thus enforces all available decisions to be abduced, together with their consequences, from all possible observable hypothetical end goals. The second purpose of integrity constraints is for ruling out impermissible actions, viz., actions that involve intentional killing in the process of reaching the goal, enforced by the integrity constraint: $false \leftarrow intentional\_killing$. The definition of $intentional\_killing$ depends on rules in each case considered and whether the doctrines of double effect or triple effect is to be upheld. Since this integrity constraint serves as the first filter of abductive stable models, by ruling out impermissible actions, it affords us with just those abductive stable models that contain only permissible actions.

Capturing Utilitarian Judgment via a Posteriori Preferences Additionally, one can further prefer amongst permissible actions those resulting in greater good. That is, whereas a priori integrity constraints can be viewed as providing an agent's reactive behaviors, generating intuitively intended responses that comply with deontological judgment (enacted by ruling out the use of intentional harm), a posteriori preferences amongst permissible actions provides instead a more involved reasoning about action-generated models, capturing utilitarian judgment that favors welfare-maximizing behaviors (in line with the dual-process model [8]).

In this application, a preference predicate is defined to select those abductive stable models containing decisions with greater good of overall consequences. For instance, in the Bystander case, this is evaluated by a utility function concerning the number of people that die as a result of possible decisions: among two abductive stable models after satisfying integrity constraints, viz., one containing abducible 'watching trolley go straight' and the other 'throwing switch to divert the trolley', the abductive stable model containing 'throwing switch to divert the trolley' is preferred, as it results in less people being killed. The reader is referred to [34] for the results of other cases.

4.2 Probabilistic Moral Reasoning

In [16], probabilistic moral reasoning is explored, via PROBABILISTIC EPA [15,33], where an example is contrived to reason about actions, under uncertainty, and thence provide judgment adhering to moral rules within some prescribed uncertainty level. The example takes a variant of the Footbridge case within the context of a jury trials in court, in order to proffer verdicts beyond reasonable doubt: *Suppose a board of jurors in a court is faced with the case where the actual action of an agent shoving the man onto the track was not observed. Instead, they are just presented with the fact that the man on the bridge died on the side track and the agent was seen on the bridge at the occasion. Is the agent guilty (beyond reasonable doubt), in the sense of violating the Doctrine of Double Effect, of shoving the man onto the track intentionally?*

To answer it, abduction is enacted to reason about the verdict, given the available evidence. Considering the active goal *judge*, to judge the case, two abducibles are available: the 'verdict of guilty beyond reasonable doubt' and 'verdict of not guilty'. Depending on how probable each verdict (the value of which is determined by the probability of intentional shoving), a preferred 'verdict of guilty beyond reasonable doubt' or 'verdict of not guilty' is abduced as a solution.

The probability with which shoving is performed intentionally is causally influenced by evidences and their attending truth values. Two evidences are considered, viz., (1) Whether the agent was running on

the bridge in a hurry; and (2) Whether the bridge was slippery at the time. The probability of intentional shoving is therefore determined by the existence of these evidences and their truth value.

Based on this representation, different judgments can be delivered, subject to available (observed) evidences and their attending truth value. By considering the standard probability of proof beyond reasonable doubt —here the value of 0.95 is adopted [30]— as a common ground for the probability of guilty verdicts to be qualified as 'beyond reasonable doubt', a form of argumentation may take place through presenting different evidence (via updating of observed evidence atoms, e.g., the agent was indeed running on the bridge in a hurry, the bridge was not slippery at the time, etc.) as a consideration to justify an exception. Whether the newly available evidence is accepted as a justification to an exception —defeating the judgment based on the priorly presented evidence— depends on its influence on the probability of intentional shoving, and thus eventually influences the final verdict. That is, it depends on whether this probability is still within the agreed standard of proof beyond reasonable doubt. The reader is referred to [16], which details a scenario capturing this moral jurisprudence viewpoint.

4.3 Modeling Morality with QUALM

Distinct from the two previous applications, QUALM emphasizes the interplay between Logic Programming abduction, updating and counterfactuals, supported furthermore by their joint tabling techniques.

Moral Permissibility and Its Justification We revisit moral permissibility with respect to the Doctrines of Double Effect and Triple Effect, but now applying counterfactuals. Counterfactuals may provide a general way to examine the Doctrine of Double Effect in dilemmas, like the classic trolley problem, 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 the Doctrine of Double Effect. 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 E is 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: 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, the evaluation of counterfactuals in this application is considered from the perspective of agents who perform the action, rather than from that of observers. Moreover, the emphasis on causation in this application focuses on agents' deliberate actions, rather than on causation and counterfactuals in general.

Related to side-effects in abduction is the concept of inspection points [32]. Therefore, it can alternatively be employed to distinguish a cause from a mere side-effect, and thus provide a supplement to counterfactuals employed for the same above purpose.

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* [26], may provide moral justification for what was done due to a lack in the current knowledge. This is accomplished by evaluating what would have followed if the intent would then have been otherwise, other things (including present knowledge) being equal. It may occasionally justify that what would have followed is no morally better than the actual ensued consequence.

Example 1. Consider a scenario developed from the Loop case of the trolley problem (see Section 2 for case description), which happens on a particularly foggy day. Due to low visibility, the agent saw only part of the looping side track, so the side track appeared to the agent rather as a straight non-looping one. The

agent was faced with a situation whether it was permissible for him to divert the trolley. The knowledge base of the agent with respect to this scenario is the simplified program below. Note, $divert(X)$ is an abducible, meaning 'diverting object X'.

$$run\_sidetrack(X) \leftarrow divert(X). \qquad hit(X,Y) \leftarrow run\_sidetrack(X), on\_sidetrack(Y).$$
$$save\_from(X) \leftarrow sidetrack(straight), run\_sidetrack(X).$$
$$save\_from(X) \leftarrow sidetrack(loop), hit(X,Y), heavy\_enough(Y).$$
$$sidetrack(straight) \leftarrow foggy. \qquad sidetrack(loop) \leftarrow not\ foggy.$$
$$foggy. \qquad on\_sidetrack(man). \qquad heavy\_enough(man).$$

Taking $save\_from(trolley)$ as the goal, the agent can perform counterfactual reasoning "if the man had not been hit by the trolley, the five people would not have been saved". Given the abduced background context $divert(trolley)$, one can verify that the counterfactual is not valid. That is, the man hit by the trolley is just a side-effect of achieving the goal, and thus $divert(trolley)$ is morally permissible according to the Doctrine of Double Effect. Indeed, this case resembles the Bystander case of the trolley problem.

The scenario continues. At some later time, the fog has subsided, and by then it was clear to the agent that the side track was looping to the main track. This results by updating the program with $not\ foggy$, rendering $sidetrack(loop)$ true. There are two standpoints on how the agent can justify its action $divert(trolley)$. For one, it can employ the aforementioned form of compound counterfactual "Had I known that the side track is looping, then if I had not diverted the trolley, the five would have been saved" as a form of self-justification. Given present knowledge that the side track loops, the inner counterfactual is not valid, meaning that to save the five people diverting the trolley (with the consequence of the man being hit) is required. Moreover, the counterfactual employed in the initial scenario "if the man had not been hit by the trolley, the five people would not have been saved", in the abduced context $divert(trolley)$, is now valid, meaning that this action is the Doctrine of Double Effect-impermissible. Therefore, the agent can justify that what would have followed from its action (given its present knowledge, i.e., $sidetrack(loop)$) is no morally better than the one at the time, when there was lack of that knowledge: its decision $divert(trolley)$ at that time was instead permissible according to the Doctrine of Double Effect. Though the agent would not have done so, had it had already the subsequent knowledge.

A different standpoint where from to justify the agent's action is by resorting to Scanlon's contractualism [44], where an action is determined impermissible through deliberative employment if there is no countervailing consideration that would justify an exception to the applied general principle. In this vein, for the example we are currently discussing, the Doctrine of Triple Effect serves as the exception to justify the permissibility of action $divert(trolley)$ when the side track was known to be looping, as shown through counterfactual reasoning in Example 10 of [37].

We extend now Example 1 to further illustrate moral permissibility of actions, as rather justified through defeasible argumentative considerations according to Scanlon's contractualism.

Example 2. As the trolley approached, the agent realized that the man on the side track was not heavy enough to stop it, acknowledged by the agent through updating $not\ heavy\_enough(man)$ into its knowledge base. But there was a heavy cart on the bridge over the looping side track that the agent could push to place it on the side track, and thereby stop the trolley. This scenario is modeled by rules below ($push(X)$ is an abducible predicate, meaning 'pushing object X'), in addition to the program of Example 1:

$$on\_sidetrack(X) \leftarrow on\_bridge(X), push(X). \qquad on\_sidetrack(Y) \leftarrow push(X), inside(Y,X).$$
$$on\_bridge(cart). \qquad heavy\_enough(cart)$$

The second rule of $on\_sidetrack(Y)$ is an extra knowledge of the agent, that if an object Y is inside the pushed object X, then Y will be on the side track too.

The goal $save\_from(trolley)$ now succeeds with $[divert(trolley), push(cart)]$ as its abductive solution. But the agent subsequently learned that a fat man, who was heavy enough, unbeknownst to the

agent, was inside the cart: the agent updates its knowledge base with $heavy\_enough(fat\_man)$ and $inside(fat\_man)$. As a consequence, this man was also on the side track and hit by the trolley, which can be verified by query `?-` $hit(trolley, heavy\_man)$.

In this scenario, a deliberate action of pushing was involved that consequently placed the fat man on the side track (verified by query `?-` $on\_sidetrack(fat\_man)$) and the man being hit by the trolley is instrumental to save the five people on the track (verified by the counterfactual "if the fat man had not been hit by the trolley, the five people would not have been saved"). The agent may justify the permissibility of its action by arguing that it is admitted by the Doctrine of Triple Effect. In this case, the fat man being hit by the trolley is just a side-effect of the agent's action $push(cart)$ in order to save the five people. This justification can be shown through reasoning on the counterfactual "if the fat man had not been pushed, then he would not have been hit by the trolley". In QUALM, checking the validity of this counterfactual amounts to fixing and updating the background context of the hypothetical action $push(fat\_man)$ into the program, and performing a corresponding intervention to falsify it. This counterfactual is not valid with respect to the intervened modified program: the fat man would still have been hit by the trolley. That means, the fat man being hit by the trolley was not caused by the hypothetical action $push(fat\_man)$, but instead by another cause. The agent may further support its argument by showing that indeed the action $push(cart)$ is the cause for the man being hit by the trolley: the counterfactual "if the cart had not been pushed, then the fat man would not have been hit by the trolley" is valid given the abduced background context $push(cart)$.

Moral Updating Moral updating (and evolution) concerns the adoption of new (possibly overriding) moral rules on top of those an agent currently follows. Such adoption often happens in the light of situations freshly faced by the agent, e.g., when an authority contextually imposes other moral rules, or due to some cultural difference. In [23], moral updating is illustrated in an interactive storytelling (using ACORDA), where the robot must save the princess imprisoned in a castle, by defeating either of two guards (a giant spider or a human ninja), while it should also attempt to follow (possibly conflicting) moral rules that may change dynamically as imposed by the princess (for the visual demo, see [24]).

The storytelling is reconstructed using QUALM, to particularly demonstrate: (1) The direct use of Logic Programming updating so as to place a moral rule into effect; and (2) The relevance of contextual abduction to rule out tabled but incompatible abductive solutions, in case a goal is invoked by a non-empty initial abductive context (the content of this context may be obtained already from another agent, e.g., imposed by the princess). A simplified program modeling the knowledge of the princess-savior robot in QUALM is shown below, where $fight(G)$ is an abducible predicate, meaning 'fighting guard G':

$$guard(spider). \qquad guard(ninja). \qquad human(ninja).$$
$$survive\_from(G) \leftarrow utilVal(G, V), V > 0.6. \qquad utilVal(spider, 0.4). \qquad utilVal(ninja, 0.7).$$
$$intend\_savePrincess \leftarrow guard(G), fight(G), survive\_from(G).$$
$$intend\_savePrincess \leftarrow guard(G), fight(G).$$

The first rule of $intend\_savePrincess$ corresponds to a utilitarian moral rule (with respect to the robot's survival), whereas the second one to a 'knight' moral, viz., to intend the goal of saving the princess at any cost (irrespective of the robot's survival chance). Since each rule in QUALM is assigned a unique name in its transform (see [40,42]), the name of each rule for $intend\_savePrincess$ may serve as a unique moral rule identifier for updating by toggling the rule's name, say via rule name fluents $\#rule(utilitarian)$ and $\#rule(knight)$, resp. In the subsequent plots, query `?-` $intend\_savePrincess$ is referred, representing the robot's intent on saving the princess.

In the first plot, when both rule name fluents are retracted, the robot does not adopt any moral rule to save the princess, i.e., the robot has no intent to save the princess, and thus the princess is not saved. In the second (restart) plot, in order to maximize its survival chance in saving the princess, the robot updates itself with the utilitarian moral: the program is updated with $\#rule(utilitarian)$. The robot thus abduces $fight(ninja)$ so as to successfully defeat the ninja instead of confronting the humongous spider.

The use of tabling in contextual abduction is demonstrated in the third (start again) plot. Assuming that the truth of $survive\_from(G)$ implies the robot's success in defeating (killing) guard G, the princess argues that the robot should not kill the *human* ninja, as it violates the moral rule she follows, say a 'Gandhi' moral, expressed by the following rule in her knowledge (the first three facts in the robot's knowledge are shared with the princess): $follow\_gandhi \leftarrow guard(G), human(G), not fight(G)$. That is, the princess abduces $not\ fight(ninja)$ and imposes this abductive solution as the initial (input) abductive context of the robot's goal (viz., $intend\_savePrincess$). This input context is inconsistent with the tabled abductive solution $fight(ninja)$, and as a result, the query fails: the robot may argue that the imposed 'Gandhi' moral conflicts with its utilitarian rule (in the visual demo [24], the robot reacts by aborting its mission). In the final plot, as the princess is not saved yet, she further argues that she definitely has to be saved, by now additionally imposing on the robot the 'knight' moral. This amounts to updating the rule name fluent $\#rule(knight)$ so as to switch on the corresponding rule. As the goal $intend\_savePrincess$ is still invoked with the input abductive context $not\ fight(ninja)$, the robot now abduces $fight(spider)$ in the presence of the newly adopted 'knight' moral. Unfortunately, it fails to survive, as confirmed by the failing of the query `?-` $survive\_from(spider)$.

The plots in this story reflect a form of deliberative employment of moral judgments within Scanlon's contractualism. For instance, in the second plot, the robot may justify its action to fight (and kill) the ninja due to the utilitarian moral it adopts. This justification is counter-argued by the princess in the subsequent plot, making an exception in saving her, by imposing the 'Gandhi' moral, disallowing the robot to kill a human guard. In this application, rather than employing updating, this exception is expressed via contextual abduction with tabling. The robot may justify its failing to save the princess (as the robot leaving the scene) by arguing that the two moral rules it follows (viz., utilitarian and 'Gandhi') are conflicting with respect to the situation it has to face. The argumentation proceeds, whereby the princess orders the robot to save her whatever risk it takes, i.e., the robot should follow the 'knight' moral.

5 Conclusion and Future Work

The chapter investigates the appropriateness of Logic Programming-based reasoning to the *terra incognita* of machine ethics, a field that is now becoming a pressing concern and receiving wide attention. The chapter discusses a number of original inroads, exhibiting a proof of possibility to model morality viewpoints systematically (within the two morality themes) through moral examples taken off-the-shelf from the literature, from our larger collection. The experiments are realized in QUALM, demonstrating the interplay of Logic Programming abduction, updating, and counterfactuals, afforded by the-state-of-the-art tabling mechanisms of XSB Prolog. The application also takes into account other Logic Programming-based reasoning features from existing systems, e.g., to deal with preference handling and probabilistic reasoning, which complement QUALM's very own features.

Given the broad dimension of the topic, the contributions in our work touch solely on a dearth of morality issues. Nevertheless, it prepares and opens the way for additional research towards employing various features in Logic Programming-based reasoning to machine ethics. Several topics can be further explored in the future, as summarized below.

Deliberative employing, within Scanlon's contractualism, addresses the question of moral permissibility by identifying the justified but defeasible argumentative considerations. This chapter shows only an informal form of argumentation, achieved through an admixture of Logic Programming abduction, updating, counterfactuals, and probabilistic reasoning. The follow-up investigation for an appropriate formal argumentation framework modeling this moral viewpoint with a system for its experimentation is a whole different research topic worth pursuing.

This chapter contemplates the individual realm of machine ethics: it stresses individual moral cognition, deliberation, and behavior. A complementary realm stresses collective morals, and emphasizes instead the emergence, in a population, of evolutionarily stable moral norms, of fair and just cooperation,

to the advantage of the whole evolved population. The latter realm is commonly studied via Evolutionary Game Theory by resorting to simulation techniques, typically with pre-determined conditions, parameters, and game strategies (see [35] for references). The bridging of the gap between the two realms [35] would appear to be promising for future work. Namely, how the study of individual cognition of morally interacting multi-agent (in the context of this chapter, by using Logic Programming-based reasoning features) is applicable to the evolution of populations of such agents, and vice versa.

Acknowledgments

LMP acknowledges support from 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 *JELIA 2002*, volume 2424 of *LNCS*, pages 50–61. Springer, 2002.
2. 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.
3. M. Anderson and S. L. Anderson. EthEl: Toward a principled ethical eldercare robot. In *Procs. AAAI Fall 2008 Symposium on AI in Eldercare*, 2008.
4. M. Anderson and S. L. Anderson, editors. *Machine Ethics*. Cambridge U. P., 2011.
5. M. Anderson, S. L. Anderson, and C. Armen. AAAI Fall Symposium on Machine Ethics. `http://www.aaai.org/Library/Symposia/Fall/fs05-06`, 2005.
6. O. Boissier, G. Bonnet, and C. Tessier. 1st Workshop on Rights and Duties of Autonomous Agents (RDA2). `https://rda2-2012.greyc.fr/`.
7. 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.
8. F. Cushman, L. Young, and J. D. Greene. Multi-system moral psychology. In J. M. Doris, editor, *The Moral Psychology Handbook*. Oxford University Press, 2010.
9. P. Dell'Acqua and L. M. Pereira. Preferential theory revision. *Journal of Applied Logic*, 5(4):586–601, 2007.
10. The Economist. Morals and the machine. Main Front Cover and Leaders (page 13), June 2nd-8th 2012.
11. K. Epstude and N. J. Roese. The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2):168–192, 2008.
12. P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967.
13. J.-G. Ganascia. Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, 9(1):39–47, 2007.
14. M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. In *5th Intl. Logic Programming Conf.* MIT Press, 1988.
15. T. A. Han, C. D. K. Ramli, and C. V. Damásio. An implementation of extended P-log using XASP. In *Procs. 24th Intl. Conf. on Logic Programming (ICLP'08)*, volume 5366 of *LNCS*. Springer, 2008.
16. T. A. Han, A. Saptawijaya, and L. M. Pereira. Moral reasoning under uncertainty. In *LPAR-18*, volume 7180 of *LNCS*, pages 212–227. Springer, 2012.
17. 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.
18. C. Higgins. US Navy funds morality lessons for robots. `http://goo.gl/EHNjzz`, 2014.
19. A. Kakas, R. Kowalski, and F. Toni. Abductive logic programming. *Journal of Logic and Computation*, 2(6):719–770, 1993.
20. F. M. Kamm. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford U. P., 2006.
21. R. Kowalski. *Computational Logic and Human Thinking: How to be Artificially Intelligent*. Cambridge U. P., 2011.
22. G. Lopes and L. M. Pereira. Prospective programming with ACORDA. In *ESCoR 2006 Workshop*, IJCAR'06, 2006.

23. G. Lopes and L. M. Pereira. Prospective storytelling agents. In *PADL 2010*, volume 5937 of *LNCS*. Springer, 2010.
24. G. Lopes and L. M. Pereira. Visual demo of "Princess-saviour Robot". Available from `https://goo.gl/vFnma2`, 2010.
25. R. Mallon and S. Nichols. Rules. In J. M. Doris, editor, *The Moral Psychology Handbook*. Oxford University Press, 2010.
26. 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.
27. R. McCloy and R. M. J. Byrne. Counterfactual thinking about controllable events. *Memory and Cognition*, 28:1071–1078, 2000.
28. 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/`.
29. S. Migliore, G. Curcio, F. Mancini, and S. F. Cappa. Counterfactual thinking in moral judgment: an experimental study. *Frontiers in Psychology*, 5:451, 2014.
30. J. O. Newman. Quantifying the standard of proof beyond a reasonable doubt: a comment on three comments. *Law, Probability and Risk*, 5(3-4):267–269, 2006.
31. M. Otsuka. Double effect, triple effect and the trolley problem: Squaring the circle in looping cases. *Utilitas*, 20(1):92–110, 2008.
32. 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.
33. L. M. Pereira and T. A. Han. Evolution prospection. In *Procs. KES International Conference on Intelligence Decision Technologies*, volume 199, pages 139–150, 2009.
34. 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.
35. L. M. Pereira and A. Saptawijaya. Bridging two realms of machine ethics. In J. B. White and R. Searle, editors, *Rethinking Machine Ethics in the Age of Ubiquitous Technology*. IGI Global, 2015.
36. L. M. Pereira and A. Saptawijaya. *Programming Machine Ethics*. Springer, 2016.
37. L. M. Pereira and A. Saptawijaya. Counterfactuals, logic programming and agent morality. In R. Urbaniak and G. Payette, editors, *Applied Formal/Mathematical Philosophy*, volume (forthcoming) of *Logic, Argumentation & Reasoning*. Springer, 2017. Available from `https://goo.gl/TKFPYo`.
38. T. M. Powers. Prospects for a Kantian machine. *IEEE Intelligent Systems*, 21(4):46–51, 2006.
39. S. Russell, D. Dewey, and M. Tegmark. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, 36(4), 2015.
40. A. Saptawijaya and L. M. Pereira. Incremental tabling for query-driven propagation of logic program updates. In *LPAR-19*, volume 8312 of *LNCS*, pages 694–709. Springer, 2013.
41. A. Saptawijaya and L. M. Pereira. Tabled abduction in logic programs (Technical Communication of ICLP 2013). *Theory and Practice of Logic Programming, Online Supplement*, 13(4-5), 2013. `http://journals.cambridge.org/downloadsup.php?file=/tlp2013008.pdf`.
42. 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`.
43. A. Saptawijaya and L. M. Pereira. TABDUAL: a tabled abduction system for logic programs. *IfCoLog Journal of Logics and their Applications*, 2(1), 2015.
44. T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 1998.
45. K. E. Stanovich. *Rationality and the Reflective Mind*. Oxford U.P., 2011.
46. T. Swift. Tabling for non-monotonic programming. *Annals of Mathematics and Artificial Intelligence*, 25(3-4):201–240, 1999.
47. The Future of Life Institute. International Grant Competition for Robust and Beneficial AI. `http://futureoflife.org/grants/large/initial`, 2015.
48. J. J. Thomson. The trolley problem. *The Yale Law Journal*, 279:1395–1415, 1985.
49. A. van Gelder, K. A. Ross, and J. S. Schlipf. The well-founded semantics for general logic programs. *Journal of ACM*, 38(3):620–650, 1991.

50. W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford U. P., 2009.
51. J. B. White and R. Searle, editors. *Rethinking Machine Ethics in the Age of Ubiquitous Technology*. IGI Global, Hershey, 2015.
52. V. Wiegel. *SophoLab: Experimental Computational Philosophy*. PhD thesis, Delft University of Technology, 2007.