Logic Programming Applied to Machine Ethics

Ari Saptawijaya* and Luís Moniz Pereira

NOVA Laboratory for Computer Science and Informatics (NOVA LINCS)
Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
ar.saptawijaya@campus.fct.unl.pt, lmp@fct.unl.pt

Abstract. This paper summarizes our investigation on the application of LP-based reasoning to machine ethics, a field that emerges from the need of imbuing autonomous agents with the capacity for moral decision-making. We identify morality viewpoints (concerning moral permissibility and the dual-process model) as studied in moral philosophy and psychology, which are amenable to computational modeling. Subsequently, various LP-based reasoning features are applied to model these identified morality viewpoints, via classic moral examples taken off-the-shelf from the literature.

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 autonomous and act in groups, amidst populations of other agents, including humans. Its importance has been emphasized as a research priority in AI with funding support [26]. Its field of enquiry, named *machine ethics*, is interdiscplinary, and 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 [2], non-monotonic reasoning in [6], and the use of Inductive Logic Programming (ILP) in [1]; some of them only abstractly, whereas others also provide implementations (e.g., using ILP-based systems [1], an interactive theorem prover [2], and answer set programming (ASP) [6]). Despite the aforementioned logic-based formalisms, Logic Programming (LP) itself is rather limitedly explored. The potential and suitability of LP, and of computational logic in general, for machine ethics, is identified and discussed at length in [11], on the heels of our work. LP permits declarative knowledge representation of moral cases with sufficiently level of detail to distinguish one moral case from other similar cases. It provides a logic-based programming paradigm with a number of practical Prolog systems, thus allowing not only addressing morality issues in an abstract logical formalism, but also via a Prolog implementation as proof of concept and a testing ground for experimentation. Furthermore, LP are also equipped with various reasoning features, as identified in the paragraph below, whose applications to machine ethics are promising, but still unexplored. This paper summarizes our integrative investigation on the appropriateness of various LP-based reasoning to machine ethics, not just abstractly, but also furnishing a proof of concept implementation for the morality issues in hand.

We identify conceptual morality viewpoints, which are covered in two morality themes: (1) *moral permissibility*, taking into account viewpoints such as the Doctrines of Double Effect (DDE) [15], Triple Effect (DTE) [10], and Scanlon's contractualist moral theory [23]; and (2) *the dual-process model* [3, 14], which stresses the interaction between deliberative and reactive behaviors in moral judgment. The mapping of all these considered viewpoints into LP-based reasoning benefits from its features and their integration, such as abduction with integrity constraints (ICs) [22], preferences over abductive scenarios [4], probabilistic reasoning [7], updating [21], counterfactuals [20], and from LP tabling technique [25].

---

* Affiliated also with the Faculty of Computer Science at Universitas Indonesia, Depok, Indonesia.

We show, in Section 2, how these various LP-based reasoning features are employed to model the aforementioned morality viewpoints, including: (1) The use of a priori ICs and a posteriori preferences over abductive scenarios to capture deontological and utilitarian judgments; (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 (wrt. Scanlon's contractualism [23]) 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, which combines LP abduction, updating, and counterfactuals, supported by LP tabling mechanisms (based on [20–22]) to examine moral permissibility wrt. DDE and DTE, 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.

2 Modeling Morality with Logic Programming

2.1 Moral Permissibility with Abduction, a Priori ICs and a Posteriori Preferences

In [17], moral permissibility is modeled through several cases of the classic trolley problem [5], by emphasizing the use of ICs in abduction and preferences over abductive scenarios. The cases, which include moral principles, are modeled in order to deliver appropriate moral decisions that conform with those the majority of people make, on the basis of empirical results in [9]. DDE [15] is utilized in [9] to explain the consistency of judgments, shared by subjects from demographically diverse populations, on a series of trolley dilemmas. In addition to DDE, we also consider DTE [10].

Each case of the trolley problem is modeled individually; their details being referred to [17]. The key points of their modeling are as follows. The DDE and DTE are modeled via a priori ICs 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 ICs, 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 ICs In this application, ICs are used for two purposes. First, they are utilized to force the goal in each case (like in [9]), by observing the desired end goal resulting from each possible decision. Such an IC thus enforces all available decisions to be abduced, together with their consequences, from all possible observable hypothetical end goals. The second purpose of ICs is for ruling out impermissible actions, viz., actions that involve intentional killing in the process of reaching the goal, enforced by the IC: $false \leftarrow intentional\_killing$. The definition of $intentional\_killing$ depends on rules in each case considered and whether DDE or DTE is to be upheld. Since this IC 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 ICs 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 [3]).

In this application, a preference predicate (e.g., based on a utility function concerning the number of people died) is defined to select those abductive stable models [4] containing decisions with greater good of overall consequences. The reader is referred to [17] for the results of various trolley problem cases.

2.2 Probabilistic Moral Reasoning

In [8], probabilistic moral reasoning is explored, 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 DDE, 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: $verdict(guilty\_brd)$ and $verdict(not\_guilty)$, where $guilty\_brd$ stands for 'guilty beyond reasonable doubt'. Depending on how probable each verdict (the value of which is determined by the probability $pr_{int\_shove}(P)$ of intentional shoving), a preferred $verdict(guilty\_brd)$ or $verdict(not\_guilty)$ 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 $pr_{int\_shove}(P)$ of intentional shoving is therefore determined by the existence of evidence, expressed as dynamic predicates $evd\_run/1$ and $evd\_slip/1$, whose sole argument is $true$ or $false$, standing for the evidences that the agent was running in a hurry and that the bridge was slippery, resp.

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 [16]– as a common ground for the probability of guilty verdicts to be qualified as 'beyond reasonable doubt', a form of argumentation (à la Scanlon contractualism [23]) may take place through presenting different evidence (via updating of observed evidence atoms, e.g., $evd\_run(true)$, $evd\_slip(false)$, 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 $pr_{int\_shove}(P)$ 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 [8], which details a scenario capturing this moral jurisprudence viewpoint.

2.3 Modeling Morality with QUALM

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

Counterfactuals in Morality We revisit moral permissibility wrt. DDE and DTE, but now applying counterfactuals. Counterfactuals may provide a general way to examine DDE 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 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 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.

We demonstrate in [18] the application of this counterfactual form in machine ethics. First, we use counterfactual queries to distinguish moral permissibility between two off-the-shelf military cases from [24], viz., terror bombing vs. tactical bombing, according to DDE. In the second application, we show that counterfactuals may as well be suitable to justify permissibility, via a process of argumentation (wrt. Scanlon contractualism [23]), using a scenario built from cases of the trolley problem that involve both DDE and DTE. Alternatively, we show that moral justification can also be addressed via 'compound counterfactuals' – *Had I known what I know today, then if I were to have done otherwise, something preferred would have followed* – for justifying with hindsight a moral judgment that was passed under lack of current knowledge.

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 [12], 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 [13]).

The storytelling is reconstructed in this paper using QUALM, to particularly demonstrate: (1) The direct use of LP 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 $fight/1$ is an abducible predicate:

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

The first rule of $intend\_savePrincess$ corresponds to a utilitarian moral rule (wrt. 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 (based on rule name fluent in [21]), 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 [13], 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 wrt. 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.

3 Conclusion and Future Work

The paper summarizes our investigation on the application of LP-based reasoning to the *terra incognita* of machine ethics, a field that is now becoming a pressing concern and receiving wide attention. Our research shows a number of original inroads, exhibiting a proof of possibility to model morality viewpoints systematically using a combination of various LP-based reasoning features (such as LP abduction, updating, preferences, probabilistic LP and counterfactuals) afforded by the-state-of-the-art tabling mechanisms, through moral examples taken off-the-shelf from the literature. Given the broad dimension of the topic, our contributions touch solely on a dearth of morality issues. Nevertheless, it prepares and opens the way for additional research towards employing various features in LP-based reasoning to machine ethics. Several topics can be further explored in the future, as summarized below.

So far, our application of counterfactuals in machine ethics is based on the evaluation of counterfactuals in order to determine their validity. It is interesting to explore in future other aspects of counterfactual reasoning relevant for moral reasoning. First, we can consider *assertive counterfactuals*: rather than evaluating the truth validity of counterfactuals, they are asserted (known) as being a valid statement. The causality expressed by such a valid counterfactual may be useful for refining moral rules, which can be achieved through incremental rule updating. Second, we may extend the antecedent of a counterfactual with a rule, instead of just literals, allowing to express exception in moral rules, such as "If killing the giant spider had been done by a noble knight, then it would not have been wrong". Third, we can imagine the situation where the counterfactual's antecedent is not given, though its conclusion is, the issue being that the conclusion is some moral wrong. In this case, we want to abduce the antecedent in the form of interventions that would prevent some wrong: "What could I have done to prevent a wrong?".

This paper 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 [19] for references). The bridging of the gap between the two realms [19] 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 paper, by using LP-based reasoning features) is applicable to the evolution of populations of such agents, and vice versa.

References

1. M. Anderson and S. L. Anderson. EthEl: Toward a principled ethical eldercare robot. In *Procs. AAAI Fall 2008 Symposium on AI in Eldercare*, 2008.
2. 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.
3. 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.
4. P. Dell'Acqua and L. M. Pereira. Preferential theory revision. *Journal of Applied Logic*, 5(4):586–601, 2007.
5. P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967.
6. J.-G. Ganascia. Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, 9(1):39–47, 2007.
7. 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.
8. 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.
9. 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.
10. F. M. Kamm. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford U. P., 2006.
11. R. Kowalski. *Computational Logic and Human Thinking: How to be Artificially Intelligent*. Cambridge U. P., 2011.
12. G. Lopes and L. M. Pereira. Prospective storytelling agents. In *PADL 2010*, volume 5937 of *LNCS*. Springer, 2010.
13. G. Lopes and L. M. Pereira. Visual demo of "Princess-saviour Robot". Available from `http://centria.di.fct.unl.pt/~lmp/publications/slides/padl10/quick_moral_robot.avi`, 2010.
14. R. Mallon and S. Nichols. Rules. In J. M. Doris, editor, *The Moral Psychology Handbook*. Oxford University Press, 2010.
15. 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/`.
16. 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.
17. 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.
18. L. M. Pereira and A. Saptawijaya. Abduction and beyond in logic programming with application to morality. Accepted in *"Frontiers of Abduction", Special Issue in IfCoLog Journal of Logics and their Applications*. Available from `http://goo.gl/yhmZzy`, 2015.
19. L. M. Pereira and A. Saptawijaya. Bridging two realms of machine ethics. In J. B. White, editor, *Rethinking Machine Ethics in the Age of Ubiquitous Technology*. IGI Global, 2015.
20. L. M. Pereira and A. Saptawijaya. Counterfactuals in logic programming with applications to agent morality. Available from `http://goo.gl/yOYxkQ`, 2015.
21. 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.
22. A. Saptawijaya and L. M. Pereira. TABDUAL: a tabled abduction system for logic programs. Accepted in *IfCoLog Journal of Logics and their Applications*. Available from `http://goo.gl/lcQGes`, 2015.
23. T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 1998.
24. T. M. Scanlon. *Moral Dimensions: Permissibility, Meaning, Blame*. Harvard University Press, 2008.
25. T. Swift. Tabling for non-monotonic programming. *Annals of Mathematics and Artificial Intelligence*, 25(3-4):201–240, 1999.
26. The Future of Life Institute. Research Priorities for Robust and Beneficial Artificial Intelligence. `http://futureoflife.org/static/data/documents/research_priorities.pdf`, 2015.