AI & Soc (2017) 32:299–300
DOI 10.1007/s00146-017-0690-y

BOOK REVIEW

# Programming Machine Ethics by Luís Moniz Pereira and Ari Saptawijaya

**Robert Kowalski**[1]

Published online: 3 March 2017


Self-driving vehicles, robotic warriors, and autonomous machines of all kinds are on our doorstep. Soon they may be making decisions, with or without our having any say, about who will live and who will die. How those decisions are made, whether we can understand them, and whether we can argue with them, will be important considerations determining whether or not these technological advances will be to our human advantage.

Much of the progress in AI that underpins these developments has been made using statistical and deep learning techniques, trained on huge amounts of data, vastly larger than those encountered in ordinary human experience. These techniques have the potential, therefore, to greatly outperform human decision making procedures. But they come at the price of their not being able to explain their decisions using concepts that ordinary people can understand. Moreover, they bypass not only rule-based and logic-based approaches in AI, but also traditional studies of human decision making in psychology, economics and philosophy.

This is where Programming Machine Ethics (PME) presents an alternative, combining traditional approaches to decision making with logic-based AI techniques. In PME, intelligent behaviour is obtained by using logic, both to generate alternative plans of actions and to derive their likely consequences, and by using preferences to choose among alternative plans of actions, using such criteria as the expected utility of the resulting state of affairs. Ethical behaviour is obtained by choosing plans that benefit society as a whole, in preference to those that serve only the selfish interests of individuals. PME shows how such intelligent and ethical behaviour can be programmed using extensions of existing logic programming techniques and implementations.

PME includes an extensive survey of research on machine ethics, and investigates a number of examples that have been studied in the ethics literature, including the classic trolley problem: A runaway trolley is headed straight for five people walking on a railway track, with no means of escape. A railwayman, observing the situation from a distance, can throw a switch to divert the train onto a side track. However, there is a single man standing on the side track, who also has no means of escape. Is it morally permissible for the railwayman to throw the switch? Most people in psychological experiments and professional ethicists alike agree that it is.

Killing one person to save five people can be justified by utilitarian considerations. But what about the situation in which there is no side track, and there is a bystander standing on a bridge next to a heavy man. Is it morally permissible for the bystander to push the heavy man onto the track, if it is guaranteed to stop the train and save the five men on the track? Most people agree that it is not, and that in general the end does not justify the means.

The distinction between the two situations exemplified by the trolley problem has been described in the literature as being due to the Doctrine of Double Effect, attributed to Thomas Aquinas. According to the Doctrine, an action that causes harm is permissible if the harm is a mere side-effect of bringing about a good result, and it is not permissible if the harm is an intended means of bringing about the same good result. Giving a computational and logical interpretation to such notions as intended means and mere side-effect is what PME is largely about.

✉ Robert Kowalski
rak@doc.ic.ac.uk

[1] Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2BZ, UK

The computational logic of PME builds upon the use of clauses of the logical form *conclusion if conditions,* to relate goals that match the *conclusion* of the clause to subgoals corresponding to the *conditions,* which may include actions as a special case. It builds upon the distinction between using clauses top-down, to deliberately reduce goals to subgoals, and using them bottom-up, to derive logical *conclusions* of the *conditions*.

For example, the railwayman in the trolley problem uses the clause:

> *You save five people if you divert the train onto the side track.*

top-down, to reduce the top-level goal of saving the five men on the track to the subgoal of diverting the train onto the side track. The subgoal itself is morally neutral, but it has the undesirable side effect:

> *You kill one person if you divert the train onto the side trac*k.

In contrast, the bystander uses the two clauses:

> *You save five people if you stop the train.*
> *You stop the train if you push a heavy person in front of the train.*

top-down, to reduce the same top-level goal to the subgoal of pushing the heavy person in front of the train. But now the subgoal is not morally neutral. It involves the use of an action that is morally undesirable in and of itself, with the foreseeable consequence that the heavy person will inevitably be killed.

The computational logic of PME is a variant of abductive logic programming (ALP), which also employs, in addition to "closed" predicates defined by means of clauses, "open" predicates used for making assumptions. In ordinary abduction, the top-level goal represents observations, and the assumptions represent candidate explanations of the observations. In applications to moral reasoning, the assumptions represent candidate actions.

Assumptions in ALP, like updates to a database, need to satisfy integrity constraints. In moral reasoning, integrity constraints can be used to represent such morally absolute principles as *thou shall not intentionally kill*.

In ALP, there can be many alternative candidate solutions that satisfy the integrity constraints. In applications of ALP to explaining observations, it is desirable to choose a candidate that is a best explanation. In deciding on a plan of actions, it is desirable to choose a candidate that accomplishes the most good. In both cases, the choice can be made by reasoning bottom-up to derive the consequences of the alternatives, and by evaluating and comparing their desirability.

In this way, PME argues that ALP is able to accommodate both deontological principles in the form of integrity constraints, and utilitarian judgements in terms of preferences between alternatives. PME also argues that the two kinds of moral judgements, deontological and utilitarian, mirror respectively reactive and deliberative thinking in dual-process models of human thinking. Reactive thinking is fast, automatic and unconscious. Deliberative thinking is slow, controlled and conscious.

In the dual-process model, the two kinds of thinking are complementary, and interact in various ways. For example, in some cases, reactive thinking quickly proposes solutions to problems as they arise, whereas deliberative thinking monitors the quality of those solutions if time allows. In other cases, deliberative thinking generates solutions, and packages them for later reactive use. PME models such latter cases by the use of tabling in Prolog systems such as XSB.

PME covers a vast range of material, from the very philosophical to the nitty gritty engineering detail. It includes such topics as the use of counter-factual reasoning to determine individual responsibility for actions (would the same effect have been obtained if the action had not been performed?). It also includes the use of probability to reason with uncertainty, and the use of evolutionary game theory to study the emergence of norms and co-operation in populations of individuals.

To cater for the wide variety of readers who will benefit from reading the book, the authors have provided a number of alternative reading paths.

The focus in PME is on applications of computational logic to morality, employing both existing techniques, such as ALP, and new ones, such as tabled contextual abduction and counterfactual reasoning. Like ALP, the new techniques have broader applications. Moreover, they have been the subject of numerous journal and conference publications by the authors, testifying to the originality and technical quality of the work.

PME draws upon a huge range of relevant work in philosophy, psychology, artificial intelligence and other disciplines, giving that work both computational and logical interpretations, and realising it by means of well-crafted computer implementations. The book fulfils an important need at a time when computers are interposing themselves into every aspect of our human lives. There is nothing else like it, and it needs to reach a wide audience.