Computational Modelling of Morality

Luís Moniz Pereira
Centro de Inteligência Artificial
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
E-mail: lmp@di.fct.unl.pt

Ari Saptawijaya
Fakultas Ilmu Komputer,
Universitas Indonesia, 16424 Depok, Jawa Barat, Indonesia
E-mail: saptawijaya@cs.ui.ac.id

Morality no longer belongs only to the realm of philosophers. The study of morality also attracts the artificial intelligence community from the computational perspective, and has been known by several names, including machine ethics, machine morality, artificial morality, and computational morality. Research on modelling moral reasoning computationally has been conducted and reported, e.g. by Anderson et al in (2005: Machine Ethics: Papers from the AAAI Fall Symposium, AAAI Press).

There are at least two reasons to mention the importance of studying morality from the computational point of view. First, with the current growing interest to understand morality in cognitive science, modelling moral reasoning computationally will assist in better understanding morality. For instance, it can greatly benefit in understanding complex interaction of cognitive aspects that build human morality or even to extract moral principles people normally apply when facing moral dilemmas. Modelling moral reasoning computationally can also be useful for intelligent tutoring systems, for instance to aid in teaching morality to children. Second, as artificial agents are more and more expected to be fully autonomous and work on our behalf, equipping agents with the capability to compute moral decisions is an indispensable requirement. This is particularly true when the agents are operating in domains where moral dilemmas occur, e.g. in health care or medical fields.

Our ultimate goal within this topic is to provide a general framework to model moral dilemmas and to draw moral judgments computationally. This framework should serve as a toolkit to codify arbitrarily chosen moral rules as declaratively as possible. We envisage that logic programming is an appropriate paradigm to achieve our purpose. Continuous and active research in logic programming has provided us with necessary ingredients that look promising enough to model morality. For instance, default negation is suitable for expressing exception in moral rules, abductive logic programming in Kakas et al (1998: Handbook of Logic in AI and Logic Programming, Oxford U.P., 235-324) and Kowalski (2006: Lecture Notes in Artificial Intelligence 3900, Springer, 1-22) and stable model semantics in Gelfond and Lifschitz (1988: Logic Programming: The 5th International Conference and Symposium, MIT Press, 1070-1080) can be used to generate possible decisions along with their moral consequences, and preferences are appropriate for preferring among moral decisions or moral rules

in Dell'Acqua and Pereira (2007: Preferential Theory Revision, Journal of Applied Logic, 5(4): 586-601).

In our work (2009: Modelling Morality with Prospective Logic, IJRIS, to appear - http://centria.di.fct.unl.pt/~lmp/publications/online-papers/ijris09-moral.pdf), we present our preliminary attempt to exploit the aforementioned enticing features of logic programming, e.g. default negation, abductive logic programming and preferences, to model moral reasoning. In particular, we employ prospective logic programming by Pereira and Lopes (2007: Prospective Logic Agents, LNAI 4784, Springer, 73-86), an on-going research project that incorporates these features. For the set of moral dilemmas, we take those from the classic trolley problem of Foot (1967: The Problem of Abortion and The Doctrine of Double Effect, Oxford Review, 5: 5-15). This problem is challenging to model since it contains a family of complex moral dilemmas. To make moral judgments on these dilemmas, we model the principle of double effect as the basis of moral reasoning. The principle can be expressed as follows: harming another individual is permissible if it is the foreseen consequence of an act that will lead to a greater good; in contrast, it is impermissible to harm someone else as an intended means to a greater good. This principle is chosen by considering empirical research results in cognitive science by Hauser (2007: Moral Minds, Little Brown) and law by Mikhail (2007: Universal Moral Grammar: Theory, Evidence, and The Future, Trends in Cognitive Science, 11(4):143-152), that show the consistency of this principle to justify similarities of judgments by diverse demographically populations when given this set of dilemmas. Additionally, we also employ prospective logic programming to model another moral principle, the principle of triple effect by Kamm (2006: Intricate Ethics: Rights, Responsibilities, and Permissible Harm, Oxford U.P). This principle refines the double effect principle, in particular on harming someone as an intended means. In this case, the triple effect principle distinguishes an action that is performed *in order to* bring about an evil from an action performed *which directly causes* an evil to occur without production of evil being its goal. The latter is a new category of action, which neither treats the occurrence of evil as a foreseen unintended consequence nor as an action performed in order to intentionally bring about an evil. The model allows us to explain computationally the difference of moral judgments drawn using these two similar but distinct moral principles.

Possible decisions in a moral dilemma are modelled as abducibles. Abductive stable models are then computed which capture abduced decisions and their consequences. Models violating integrity constraints, i.e. models that contain actions involving intentional killing, are ruled out. Finally, a posteriori preferences are used to prefer models that characterize more preferred moral decisions, including the use of utility functions. These experiments show that preferred moral decisions, i.e. the ones that follow the principle of double effect, are successfully delivered. They conform to the results of empirical experiments conducted in cognitive science and law. Regarding the triple effect principle, the inspection feature of ACORDA can be employed to detect mere consequences of abducibles. Hence, we can distinguish computationally two moral judgments in

line with the triple effect principle, i.e. whether an action is performed in order to bring about an evil or just because an evil will occur.

Our attempt to model moral reasoning on this domain shows encouraging results. Using features of prospective logic programming, we can conveniently model various moral dilemmas of the trolley problem, the principle of double effect, and the principle of triple effect, in a declarative manner. Our experiments on running the model also successfully deliver moral judgments that conform to the human empirical research results by Hauser (2007) and Mikhail (2007: 143-152).

Much research has emphasized using machine learning techniques, e.g. statistical analysis by Rzepka and Araki (2005: Machine Ethics: Papers from the AAAI Fall Symposium, AAAI Press, 85-87), neural networks by Guarini (2005: Machine Ethics: Papers from the AAAI Fall Symposium, AAAI Press, 52-61), case-based reasoning by McLaren (2006: Computational Models of Ethical Reasoning: Challenges, Initial Steps and Future Directions, IEEE Intelligent Systems, 21(4):29-37) and inductive logic programming by Anderson et al {2006: MedEthEx: A Prototype Medical Ethics Advisor, Procs. IAAI'06, AAAI Press) to model moral reasoning from examples of particular moral dilemmas. Our approach differs from them as we do not employ machine learning techniques to deliver moral decisions.

Powers (2006: Prospects for a Kantian Machine, IEEE Intelligent Systems, 21(4):46-51) proposes to use nonmonotonic logic to specifically model Kant's categorical imperatives, but it is unclear whether his approach has ever been realized in a working implementation. On the other hand, Bringsjord et al (2006: Toward a General Logicist Methodology for Engineering Ethically Correct Robots, IEEE Intelligent Systems, 21(4):38-44) propose the use of deontic logic to formalize moral codes. The objective of their research is to arrive at a methodology that allows an agent to behave ethically as much as possible in an environment that demands such behaviour. We share our objective with them to some extent as we also would like to come up with a general framework to model moral judgments computationally. Different from our work, they use an axiomatized deontic logic to decide which moral code is operative to arrive at an expected moral outcome. This is achieved by seeking a proof for the expected moral outcome to follow from candidates of operative moral codes.

To arrive at our ultimate research goal, we envision several possible future directions. We would like to explore how to express metarule and metamoral injunctions. By metarule we mean a rule to resolve two existing conflicting moral rules in deriving moral decisions. Metamorality, on the other hand, is used to provide protocols for moral rules, to regulate how moral rules interact with one another. Another possible direction is to have a framework for generating precompiled moral rules. This will benefit fast and frugal moral decision making which is sometimes needed, cf. heuristics for decision making in law by Gigerenzer and Engel (2006: Heuristics and the Law, MIT Press), rather than to have full deliberative moral reasoning every time.

We envision a final system that can be employed to test moral theories, and also can be used for training moral reasoning, including the automated generation of example tests and their explanation. Finally, we hope our research will help in imparting moral behaviour to autonomous agents.