

Preface

Scope

In recent years, systems or agents have become ever more sophisticated, autonomous, and act in groups, amidst populations of other agents, including humans. Autonomous robots or agents have been actively developed to be involved in a wide range of fields, where more complex issues concerning responsibility are in increased demand of proper consideration, in particular when the agents face situations involving choices on moral or ethical dimensions. In medical or elder care, a robot may be confronted with conflicts between attaining its duties to treat the patient and respecting the patient's decision, e.g., in the case where the patient rejects a critical treatment the robot recommends. In the military, where robots from different makers, with diverse purposes, shapes, and sizes have been built and even deployed in wars, will naturally face moral dilemmas, e.g., whether it is permissible for a drone to fire on a house where a target is known to be hiding, but at the same time it is one that also sheltering civilians. In fact, there has been much attention given recently concerning the ethical issue of such autonomous military robots. More recently, a multidisciplinary team of researchers received funding from the U.S. Navy to explore the challenges of providing autonomous agents with ethics.

As these demands become ever more pervasive and ubiquitous, the requirement that agents should function in an ethically responsible manner is becoming a pressing concern. Accordingly, *machine ethics*, which is also known under a variety of names, such as machine morality, computational morality, artificial morality, and computational ethics, emerges as a burgeoning field of inquiry to attend to that need, by imbuing autonomous agents with the capacity for moral decision-making. Clearly, machine ethics brings together perspectives from various fields, among them: philosophy, psychology, anthropology, evolutionary biology, and artificial intelligence. The overall result of this interdisciplinary research is therefore not just important for equipping agents with some capacity for making moral decisions, but also to help better understand morality, via the creation and testing of computational models of ethical theories.

The field is increasingly more so recognized, and its importance has been emphasized in dedicated scientific meetings, book publications, as well as a heightened public awareness to its importance. The Future of Life Institute explicitly identifies machine ethics as one important research priorities in promoting artificial intelligence (AI) research, which is not only capable, but also robust and beneficial. More recently, this initiative, which is supported by top AI researchers from industry and academia, received a significant amount of funding to run a global research program aiming at those priorities. The topic continues to receive wide attention in a variety of conferences, both soliciting papers and promoting panels.

This book reports our investigations on programming machine ethics in two different realms: the individual and collective realms. In studies of human morality, these distinct interconnected realms are evinced too: one stressing above all individual cognition, deliberation, and behavior; the other stressing collective morals, and how they emerged. Of course, the two realms are necessarily intertwined, for cognizant individuals form the populations, and the twain evolved jointly to cohere into collective norms, and into individual interaction.

Content

The main content of the book starts with the groundbreaking work of Alan Turing, in Chap. 1, to justify our *functionalism* stance regarding the modeling of morality, and thus programming machine ethics. In the realm of the individual, computation is vehicle for the study of morality, namely in its modeling of the dynamics of knowledge and cognition of agents. We begin, in Chap. 2, with a survey of research in the individual realm of machine ethics. It provides the context and the motivation for our investigations in addressing moral facets such as permissibility and the dual process of moral judgments by framing together various *logic programming* (LP) knowledge representation and reasoning features that are essential to moral agency, viz., abduction with integrity constraints, preferences over abductive scenarios, probabilistic reasoning, counterfactuals, and updating. Computation over combinations of these features has become our vehicle for modeling the dynamics of moral cognition within a single agent. The investigation of the individual realm of machine ethics reported in this book is a part of the recent Ph.D. thesis of the second author, supervised by the first author, which explores and exploits, for the first time, the aforementioned LP-based knowledge representation and reasoning features for programming machine ethics.

This investigation, elaborated on in Chaps. 3–8 of this book, does not merely address the appropriateness of LP-based reasoning to machine ethics abstractly, but also provides an implementation technique as a basis and testing ground for experimentation of said moral facets. That is, it serves as proof of concept that our understanding of the considered moral facets can in part be computationally modeled and implemented, testing them through a variety of classic moral examples

taken off-the-shelf from the morality literature, together with their reported empirical results about their judgments serving as validity reference.

- Chapter 3 reports on our study of the literature in moral philosophy and psychology for choosing those moral facets and their conceptual viewpoints that are close to LP-based representation and reasoning. We discuss in particular: (1) *moral permissibility*, taking into account the doctrines of double effect and triple effect, and Scanlonian contractualism; (2) *the dual process model* that stresses the interaction between deliberative and reactive processes in delivering moral decisions; and (3) the role of *counterfactual* thinking in moral reasoning.
- Chapter 4 starts with necessary background in logic programming. The reader who is familiar with semantics of logic programs (particularly the stable model and the well-founded semantics), may skip this technical background (in Sect. 4.1), as we simply need them to justify how the results in the examples are obtained. Other sections in this chapter briefly overview the considered LP reasoning features:
 - *Abduction*, whose roles are of scenario generation and of hypothetical reasoning, including the consideration of counterfactual scenarios about the past;
 - *Preferences*, which are enacted for preferring scenarios obtained by abduction;
 - *Probabilistic LP*, which allows abduction to take scenario uncertainty measures into account;
 - *LP updating*, which enables updating the knowledge of an agent, either actual or hypothetical. The latter is used for back-in-time temporary causal interventions specified by counterfactuals in Chap. 6;
 - *LP counterfactuals* permit hypothesizing into the past, and even taking into account present knowledge when so doing;
 - *Tabling* affords solutions reuse (rather than recomputing them), and is employed in joint combination with abduction and updating, as discussed in Chap. 5. Therefore tabling also enables reactivity by picking up readymade solutions, including priorly abducted ones, and those that result from subsequent incremental updating.

In Chap. 4, we also discuss the appropriateness of the above features for representing and reasoning about diverse issues of moral facets tackled in this book.

- Chapter 5 details novel approaches for employing tabling in abduction and updating, viz., tabling abductive solutions in *contextual abduction* (TABDUAL) and the *incremental tabling* of fluents (EVOLP/R), respectively. Tabling in contextual abduction allows the reuse of priorly obtained abduction results in a different context, whereas the use of incremental tabling in updating permits bottom-up propagation of updates, and hence avoids top-down frame axiom computations at a high level. The combined use of tabling in contextual abduction and updating, discussed in Chap. 7, therefore permits the interaction

between the deliberative and reactive processes foreseen by the dual process model.

- Chapter 6 elaborates our LP-based counterfactuals evaluation procedure, concentrating on pure non-probabilistic counterfactual reasoning by resorting to abduction and updating, in order to determine the logical validity of counterfactuals.
- Chapter 7 discusses the three LP systems (ACORDA, PROBABILISTIC EPA, and QUALM), emphasizing how each of them distinctively incorporates a combination of the LP-based representation and reasoning features discussed in Chap. 4.
- Chapter 8 details the applications of ACORDA, PROBABILISTIC EPA, and QUALM for modeling various issues relevant to the chosen moral facets.

The other realm of machine ethics, viz., the collective one, concerns itself with computational moral emergence. The mechanisms of emergence and evolution of cooperation in populations of abstract individuals, with diverse behavioral strategies in co-presence, have been undergoing mathematical study via *evolutionary game theory* (EGT), inspired in part on evolutionary psychology. Their systematic study resorts to simulation techniques, thus enabling the study of aforesaid mechanisms under a variety of conditions, parameters, and alternative virtual evolutionary games. The theoretical and experimental results have continually been surprising, rewarding, and promising.

In the collective realm, the computational study of norms and moral emergence via EGT is typically conducted in populations of rather simple-minded agents. That is, these agents are not equipped with any cognitive capability, and thus simply act from a predetermined set of actions. Our research (of the first author and other co-authors) has shown that the introduction of cognitive capabilities, such as intention recognition, commitment, revenge, apology, and forgiveness, reinforce the emergence of cooperation in the population, comparatively to the absence of such cognitive abilities.

We discuss, in Chap. 9 of this book, how modeling the aforesaid cognitive capabilities in individuals within a networked population shall allow them to fine tune game strategies, and in turn may lead to the evolution of high levels of cooperation. Such driving strategies are associated with moral “emotions” that motivate moral discernment and substantiate ethical norms, leading to improved general conviviality on occasion, or not. To wit, we can model moral agency without explicitly representing embodied emotions, as we know them. Rather, such software-instantiated “emotions” are modeled as (un)conscious strategic heuristics empowered in complex evolutionary games. Moreover, modeling such capabilities in individuals within a population may help us understand the emergent behavior of ethical agents in groups, in order to implement them not just in a simulation, but also in the real world of future robots and their swarms.

The EGT techniques mentioned in Chap. 9 are rather standard and hence use rather standard EGT software, whose originality of use relies, as usual, on infusing it with the specific equations of our analytical mathematical models, plus the parameters for the attending simulations. They can be followed up in detail in our

own references therein, many of them in open-access publications. Because they are substantially mathematically sophisticated and extensive in nature, we thought best to leave them out of the book. Suffice it to say that these EGT techniques enable us to obtain our results, regarding mixed populations of individuals and their strategies, either analytically, by simulation, or both.

Having contemplated the two distinct realms of machine ethics, a fundamental question then arises, concerning the study of individual cognition in small groups of frequently morally interacting multi-agents that can choose to defect or cooperate with others. That is, whether from such a study we can obtain results equally applicable to the evolution of populations of such agents; and vice versa, whether the results obtained in the study of populations carry over to groups of frequently interacting multi-agents, and under what conditions. This issue is discussed in Chap. 10, by bringing to the fore congenial present views and research on the evolution of human morality, hoping to reinforce the bridging ideas and computational paradigms we set forth. Moreover, we take for granted that computational and robotic models can actually provide abstract and concrete insight on emerged human moral reality, irrespective of the distinct embodiments of man and machine.

Reading Paths

The book is best read sequentially, chapter by chapter. Nevertheless, several alternative reading paths are possible, as shown below. These reading paths can also be first read cursively, by simply reading the introductions and concluding remarks of each of the corresponding chapters. Those in parentheses provide some necessary background, technical details, and further references for their respective topic. They may safely be skipped and read only if needed. Moreover, the explanations set forth in the previous section, about the motivation and resulting use of the main techniques, may well suffice for the general reader who is not concerned with the details of how they are brought about and integrated, and is willing to believe that our implemented techniques actually work.

- A general survey of approaches to machine ethics:

$$1 \rightarrow 2 \rightarrow 3 \rightarrow (4.1) \rightarrow 4.2 \text{ till } 4.8 \rightarrow 10 \rightarrow 11$$

- The individual realm of machine ethics via logic programming (LP):

$$1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 11$$

- Focusing on abduction and preferences over abductive scenarios:

$$(3) \rightarrow 4.1 \rightarrow 4.2 \rightarrow 4.3 \rightarrow 7.1 \rightarrow 8.1$$

- Focusing on abduction and probabilistic LP:

$$(3) \rightarrow 4.1 \rightarrow 4.2 \rightarrow 4.4 \rightarrow 7.2 \rightarrow 8.2$$

- Focusing on abduction, updating, and tabling:

$$(3) \rightarrow 4.1 \rightarrow 4.2 \rightarrow 4.5 \rightarrow 4.7 \rightarrow 5 \rightarrow 7.3 \rightarrow 8.3$$

- Focusing on counterfactuals:

$$(3) \rightarrow 4.1 \rightarrow 4.2 \rightarrow 4.5 \rightarrow 4.6 \rightarrow 6 \rightarrow (7.3) \rightarrow 8.3.2$$

- The collective realm of machine ethics via evolutionary game theory:

$$1 \rightarrow 9 \rightarrow 10 \rightarrow 11$$

- General LP engineering techniques:

- Abduction with tabling: $4.1 \rightarrow (4.2) \rightarrow (4.7) \rightarrow 5.1$
- Updating with tabling: $4.1 \rightarrow (4.5) \rightarrow (4.7) \rightarrow 5.2$
- Counterfactuals: $4.1 \rightarrow (4.6) \rightarrow 6$

Audience

The primary audience for this book are the researchers and postgraduate students in the cognitive sciences, artificial intelligence, robotics, philosophy, and ethics. The secondary audience for this book, taking into account the above reading suggestions, are the curious academic and general publics; undergraduates looking for topics of research; science journalists; science and society forums; legislators and the military concerned with machine ethics.

Lisbon, Portugal
Jakarta, Indonesia

Luís Moniz Pereira
Ari Saptawijaya

Programming Machine Ethics

Pereira, L.M.; Saptawijaya, A.

2016, XIX, 175 p. 5 illus., Hardcover

ISBN: 978-3-319-29353-0