

## Chapter 2

# The Individual Realm of Machine Ethics: A Survey

In this chapter, a survey of research in machine ethics is presented, providing the context and the motivation for our investigations. The survey concerns the individual realm of machine ethics, whereas the background to other realm, the collective one, is broached in Chap. 9, namely Sects. 9.1 and 9.2.1. The first realm views computation as a vehicle for representing moral cognition of an agent and its reasoning thereof, which motivates our investigation for employing Logic Programming (LP) knowledge representation and reasoning features with respect to the individual realm of machine ethics. On the other hand, the second realm emphasizes the emergence, in a population, of evolutionarily stable moral norms, of fair and just cooperation, that ably discards free riders and deceivers, to the advantage of the whole evolved population. It provides a motivation of our research for introducing cognitive abilities, such as intention recognition, commitment, revenge, apology, and forgiveness, to reinforce the emergence of cooperation in the collective realm of machine ethics.

## 2.1 TRUTH-TELLER and SIROCCO

TRUTH-TELLER [35] is a system that qualitatively compares a pair of ethical dilemma cases about whether to tell the truth and extracts ethically salient similarities and differences of the reasons for telling the truth (or not), from the perspective of the agent faced with the dilemma. The representation of a case is manually constructed from the interpretation of the story. Semantic networks are employed to represent the truth telling episodes (including the actors involved, their relationships, possible actions and reasons supporting possible actions), a hierarchy of relationships (familial, commercial, etc.), and a hierarchy of reasons for or against telling the truth based on the formulation in [14]. The representation is then analyzed by case-based (casuistic) reasoning in several steps:

- First, a mapping between the reasons in two cases is built, viz., by matching similar and marking distinct reasons.
- The second step qualifies: (1) the relationships among actors, actions, and reasons in one case; and (2) the mappings of these objects to those in the other considered case based on considerations such as criticalness, actors' roles and alternative actions.
- The third step points out similarities and differences of cases, with respect to pre-defined comparison contexts, whether the considered reasons apply to both cases, apply more strongly in one case than another, or apply to only one case.

The analysis result is then summarized in a comparison text.

SIROCCO [34] also employs case-based (casuistic) reasoning, but unlike TRUTH-TELLER, it accepts an ethical dilemma case and retrieves ethics principles and past cases that are relevant to the target case. It is developed in order to operationalize general abstract ethics principles, as ethicists often record their explanations of how and why they applied and reconciled principles in resolving specific cases. SIROCCO particularly addresses a domain of engineering ethics, taking into account ethics code of the National Society of Professional Engineers (NSPE) [38] and the cases decided by its Board of Ethical Review (BER). The BER's decision making indicates that several operationalization techniques are applied, which include linking ethics code and past cases with the facts of the considered case, grouping codes and cases so they can be cited in support of a conclusion, and reusing some reasoning applied in past cases to the context of a new case. In SIROCCO, cases are represented in general using the Ethics Transcription Language (ETL) [34] as chronological narratives of facts involving actors, actions participated by actors, and temporal relations between facts. The representation of a source case is particularly extended with its analysis by BER, which captures an operationalization of NSPE ethics codes. In its retrieval process, SIROCCO first computes the best  $N$  matches of source cases with respect to the ETL fact matching between the target case and each source case, and subsequently finds a structural mapping using  $A^*$  search between the target case and the best  $N$  matches [15]. The goal is to map facts of a source to a corresponding fact in the target at the same level of abstraction, while keeping a consistent mapping between their actors and temporal relations. Finally, it analyzes the results of multiple source cases (rather than of a single best match) to generate suggestions for the target case, such as relevant principles and relevant source cases.

## 2.2 JEREMY and W.D.

JEREMY [6] is a system that follows the theory of act utilitarianism. This theory maintains that an act is morally right if and only if the act maximizes the good, viz., the one with the greatest net good consequences, taking into account all those affected by the action [46]. In JEREMY, hedonistic act utilitarianism is particularly adopted, where the pleasure and displeasure of those affected by each possible action

are considered. This is manifested by three components with respect to each affected person  $p$ : (1) the intensity  $I_p$  of pleasure/displeasure, scaled between -2 and 2; (2) the duration  $D_p$  of the pleasure/displeasure, in days; and (3) the probability  $P_p$  that this pleasure/displeasure will occur. The total net pleasure for each action  $a$  is computed as follows:

$$Total_a = \sum_{p \in Person} (I_p \times D_p \times P_p).$$

The right action is the one giving the highest total net pleasure  $Total_a$ .

In order to respond to critics of act utilitarianism, another prototype, W.D. [6], is developed to avoid a single absolute duty. That is, it follows several duties, where in different cases a duty can be stronger than (and thus overrides) the others, following the theory of prima facie duties of Ross [45]. This theory comprises duties like fidelity (one should honor promises), reparation (one should make amends for wrongs done), gratitude (one should return favors), justice (one should treat people as they deserve to be treated), beneficence (one should act so as to bring about the greatest good), non-maleficence (one should act so as to cause the least harm), and self-improvement (one should develop one's own abilities/talent to the fullest).

In W.D., the strength of each duty is measured by assigning it a weight, capturing the view that a duty may take precedence over another. W.D. computes, for each possible action, the weighted sum of duty satisfaction, and returns the greatest sum as the right action. In order to improve the decision, in the sense of conforming to a consensus of correct ethical behavior, the weight of a duty is allowed to be adjusted through a supervised learning, by acquiring suggested action from the user. This weight adjustment to refine moral decision is inspired by reflective equilibrium of Rawls [41]: reflecting on considered judgments about particular cases and revising any elements of these judgments (principles that govern these judgments, theories that bear on them, etc.) wherever necessary, in order to achieve an acceptable coherence amongst them, the so-called equilibrium [19]. In [6], it is however unclear which supervised learning mechanism is actually implemented in W.D.

## 2.3 MEDETHEX and ETHEL

The theory of prima facie duties is further considered in [2, 8], while also concretely employing machine learning to refine its decision making. As in W.D., the employing of machine learning is also inspired by reflective equilibrium of Rawls [41], viz., to generalize intuition about particular cases, testing this generalization on further cases, and then repeats this process to further refine the generalization towards the end of developing a decision procedure that agrees with intuition.

The first implementation is MEDETHEX [8], which is based on a more specific theory of prima facie duties, viz., the Principle of Biomedical Ethics of Beauchamp and Childress [13]. The considered cases are a variety of the following type of ethical dilemma [7]:

*A healthcare professional has recommended a particular treatment for her competent adult patient, but the patient has rejected it. Should the healthcare professional try to change the patient's mind or accept the patient's decision as final?*

The cases thus involve only two possible actions, viz., (1) accepting a patient's decision to reject a treatment; and (2) trying to convince him to change his mind. Furthermore, the cases are constrained to three of four duties in [13], viz., respecting for the autonomy of the patient, not causing harm to the patient (non-maleficence), and promoting patient welfare (beneficence).

MEDETHEX is implemented using Inductive Logic Programming (ILP) [36] to learn the relation *supersedes*( $A_1, A_2$ ), i.e., whether action  $A_1$  supersedes (viz., is ethically preferable to) action  $A_2$ . The training (positive) examples comprise cases, where each case is associated with an estimate satisfaction/violation value of each duty for each possible action (scaled from -2 to 2) and the ethically preferred action for the case. The negative examples are obtained by simply exchanging the preferred action from the positive training examples. The relation *supersedes*( $A_1, A_2$ ) is then learned from these positive and negative examples, expressing it in terms of the lower bounds for difference of values of the considered duties between the two actions  $A_1$  and  $A_2$ .

Similar to MEDETHEX, ETHEL is also based on the Principle of Biomedical Ethics of Beauchamp and Childress [13], but applied to the domain of eldercare with the main purpose to remind a patient to take his/her medication, taking ethical duties into consideration. It also decides, after a patient has been reminded, whether to accept his/her refusal to take the medication (in which case a further reminder may take place) or to notify an overseer (e.g., a medical doctor) instead.

ETHEL is also implemented using ILP, following a similar technique employed in MEDETHEX to learn the same relation *supersedes*( $A_1, A_2$ ); this relation is also defined in terms of the lower bounds for difference of values of the corresponding duties between actions  $A_1$  and  $A_2$ . Unlike MEDETHEX, due to the reminder feature, the satisfaction/violation values of duties for each action in ETHEL are adjusted over time. This adjustment is determined by several factors, such as the maximum amount of harm if the medication is not taken, the number of hours for this maximum harm to occur, etc.; this information is obtained from the overseer. Adjusting the satisfaction/violation values of duties permits ETHEL to remind (or not) the patient to take his/her medication as well as to notify (or not) the overseer at ethically justifiable moment.

ETHEL has been deployed in a robot prototype, capable to find and walk toward a patient who needs to be reminded of medication, to bring the medication to the patient, to engage in a natural-language exchange, and to notify an overseer by email when necessary [3].

There has been work going beyond ETHEL from the same authors, viz., GENETH [4, 5], which is implemented using the same ILP technique.

## 2.4 A Kantian Machine Proposal

A more philosophical tone of machine ethics is presented in [40], where he argues that rule-based ethical theories like the first formulation of Kant’s categorical imperative (“Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction” [30]) appear to be promising for computational morality, because of their computational structure for judgment. Three views on how to computationally model categorical imperative are envisaged.

First, in order for a machine to maintain consistency in testing ethical behavior, it should be able to construct a moral theory that renders individual maxims to be universally quantified (over circumstances, purposes, and agents) and to map them onto deontic categories, viz., forbidden, permissible, and obligatory action. Deontic logic is regarded as an appropriate formalism with respect to this first view. He abstractly refers to schemata for the three deontic categories, that for every agent, circumstance  $C$ , and purpose  $P$ :

- Action  $A$  is obligatory:  $(C \text{ and } P) \rightarrow A$ .
- Action  $A$  is forbidden:  $(C \text{ and } P) \rightarrow \neg A$ .
- Action  $A$  is permissible:  $\neg((C \text{ and } P) \rightarrow A)$  and  $\neg((C \text{ and } P) \rightarrow \neg A)$ .

where a candidate maxim should be an instance of these three schemata.

Powers suggests that mere consistency is not sufficient for a maxim. Instead, its consistency should also be checked with other existing facts or background theory. This leads to his second view, viz., the need of common-sense reasoning in the categorical imperative to deal with contradiction. For this view, he refers to non-monotonic logic, which is appropriate to capture defeating conditions to a maxim. In this regard, he particularly resorts to default logic of Reiter [43] as a suitable formalism, that adding the default rules allows maxims to contradict the background set of facts and common-sense rules without introducing inconsistency.

In his third view, Powers contemplates on the construction of a coherent system of maxims, where he sees such construction analogous to the belief revision problems. In the context of bottom-up construction, he envisages an update procedure for a machine to update its system of maxims with another maxim, though it is unclear to him how such an update can be accomplished.

The formalisms in these three views are only considered abstractly and no implementation is referred to address them.

## 2.5 Machine Ethics via Theorem Proving

In [16], mechanized multi-agent deontic logic is employed with the view that ethically correct robot behaviors are those that can be proved in a deontic logic. For obtaining such a proof of ethically permissible actions, they resort to a sequent-based natural-deduction of Murakami [37] axiomatization of Horty’s utilitarian formulation of

multi-agent deontic logic [26]. This deduction system is encoded in the interactive theorem prover ATHENA [9]. The use of *interactive* theorem prover is motivated by the idea that an agent operates according to ethical codes bestowed on them, and when its automated reasoning fails, it suspends its operation and asks human guidance to resolve the issue.

Taking an example in health care, where two agents are in charge of two patients with different needs (patient  $H_1$  depends on life support, whereas patient  $H_2$  on very costly pain medication), two actions are considered: (1) terminate  $H_1$ 's life support to secure his organ for five humans; and (2) delay delivery of medication to  $H_2$  to conserve hospital resources. The approach in [16] begins with supposing several candidates of ethical codes, from harsh utilitarian (that both terminates  $H_1$ 's life and delay  $H_2$  medication) to most benevolent (neither terminates  $H_1$ 's life nor delay  $H_2$  medication); these ethical codes are formalized using the aforementioned deontic logic. The logic additionally formalizes behaviors of agents and their respective moral outcomes. Given these formalizations, ATHENA is employed to query each ethical code candidate in order to decide which amongst them should be operative, meaning that the best moral outcome (viz., that resulting from neither terminates  $H_1$ 's life nor delay  $H_2$  medication) is provable from the operative one.

## 2.6 Particularism versus Generalism

A computational model to study the dispute between particularism and generalism, is explored in [23, 25]. Moral generalism stresses the importance of moral principles and their *general* application in moral reasoning, whereas moral particularism favors on the view that moral reasoning (and decisions) depend on cases and not on a general application of moral principles to cases [18].

In [23], different ethical principles (of Aristotle, Kant, and Benjamin Constant) are modeled using answer set programming, implemented with ANSPROLOG\* [11]. The aim is to show that non-monotonic logic is appropriate to address the opposition between generalism and particularism by capturing justified exceptions in general ethics rules. The tension between these two viewpoints is exemplified by a classic dilemma about lying: in a war situation one hides a friend who is wanted by the military force, raising to a dilemma whether he should tell the truth, denouncing his friend to the military, which leads to the murder of his friend.

In order to model this dilemma in the view of Aristotle's ethics (viz., choosing the least unjust action), several possible actions are conceived, e.g., tell the truth, tell a lie, etc., and facts about consequences of these actions are defined. Predicate *unjust*( $A$ ) is then defined by assessing whether the consequence of  $A$  is worse than the consequence of other actions, via predicate *worse*/2, whose parameters are the consequences of two considered actions. Depending on the definition of *worse*/2, the answer sets may be split into one part corresponding to telling the truth and the other part to telling a lie. The model itself does not provide a mechanism to prefer among these answer sets, though it illustrates that an ad hoc preference is possible by

explicitly changing the definition of *worse/2* predicate so as all answer sets contain the action of telling a lie (providing that murder has worse consequence than that of all other actions).

Ganascia [23] also contrasts Kant’s categorical imperative and Constant’s objection. For Kant’s categorical imperative, a rule such as:

$$act(P, A) \leftarrow person(P), action(A), act("I", A)$$

is defined to universalize a maxim: it stipulates that if “I” act in such  $A$ , all person ( $P$ ) could act the same. This view does not require preferences among different actions, but emphasizes possible consequences of a maxim that cannot be universalized, e.g., a society where nobody can be trusted:  $untrust(P) \leftarrow act(P, tell(P, lie))$ . To this end, while lying can be admitted in an answer set, the answer set reflects a world where nobody can be trusted.

While this Kantian view aims at upholding generality of ethics principles, Constant’s theory authorizes principles that tolerate exceptions. The lying dilemma is modeled by capturing a more specific principle for telling the truth: we always have to tell the truth, except to someone who does not deserve it. This is achieved in the model by: (1) not only considering the transmitter of the speech (as in the Kantian model), but also the receiver; and (2) using default negation to express the principle in a way that one should always tell the truth, except when the receiver is a murderer.

In [25], the dispute between particularism and generalism is addressed using artificial neural networks. More specifically, simple recurrent networks are trained with cases about permissibility of actions involving killing and allowing to die.

The input for a network encodes the actor, the recipient of the action and the motive or the consequence of the action (e.g., killing in self-defence, allowing one to die to save many innocents, etc.), but without the provision of explicit moral rules, whereas the output of the network determines the permissibility of an input case. The experiments are performed on several networks that are trained differently. For instance, one network is trained by classifying permissibility based on the motive or the consequence of the action (irrespective whether the action is killing or allowing to die), whereas another network is trained by distinguishing the action killing from allowing to die.

By employing these trained networks to classify test cases, one result suggests that acting in self-defence contributes to permissibility, whereas actions that lead to the deaths of innocents are impermissible. Further analysis on the similarity of hidden unit activation vector between cases suggests that killing and allowing to die are making different contributions to the similarity spaces for different trained networks. Nonetheless, the networks admittedly learn some principles in general, though it cannot be directly expressed in classical, discrete representational structure. The experiments thus show that the behavior of the networks is in agreement with the so-called *contributory standards* of moral principles [18]—a middle ground between particularist and generalist—which allows more than one principle to be applicable to a case, as each specifies how things are, only in a certain respect.

## 2.7 Concluding Remarks

In this chapter the open and broad nature of research in machine ethics has been illustrated. On the one hand, it spans a variety of morality viewpoints, e.g., utilitarianism, Kant's categorical imperative, *prima facie* duties, particularism and generalism. On the other hand, a number of approaches have been proposed to model these viewpoints, with some assumptions to simplify the problem, which is somewhat unavoidable, given the complexity of human moral reasoning. The open nature of this research field is also indicated by different purposes these approaches are designed for (e.g., retrieving similar moral cases, explicitly making moral decisions, or finding operative moral principles).

TRUTH-TELLER and SIROCCO point out the role of knowledge representation (using semantic networks and its specific language ETL, respectively) to represent moral cases in a sufficiently fine level of detail, and rely on such representation for comparing cases and retrieving other similar cases. This form of representation is not so emphasized in JEREMY and W.D., as they reduce the utilitarian principle and duties into numerical values within some scale. Unlike TRUTH-TELLER and SIROCCO, JEREMY and W.D. aim at explicitly making moral decisions; these decisions are determined by these values through some procedure capturing the moral principles followed. Such quantitative valuation of duties also forms the basis of MEDETHEX and ETHEL, though some basic LP representation is employed for representing this valuation in positive and negative instances (which are needed for their ILP learning mechanism), as well as for representing the learned principle in the form of a LP rule. This learned principle, in terms of these numerical values, determines moral decisions made by these systems.

The employment of logic-based formalisms in the field, notably deontic logic, to formalize moral theory appears in the Kantian machine proposal of Powers [40]. Indeed, the insufficiency of abstract logic-based formalism for rule-based ethical theories is identified in this proposal, emphasizing the need of non-monotonic reasoning in order to capture defeating conditions to a maxim. Moreover, the proposal also points out the importance of an update procedure to anticipate updating a system of maxims with another. Unfortunately there is no concrete realization of this proposal. In [16], an interactive theorem prover is employed to encode a specific deontic logic formalism in order to find an operative moral principle (amongst other available ones) in the form of proof. The use of theorem prover in this approach however does not concern the non-monotonic reasoning and the moral updating issues raised by Powers [40].

The issue of non-monotonic reasoning becomes more apparent in the study about particularism versus generalism. Ganascia [23] demonstrates in a concrete moral case how non-monotonic reasoning can be addressed in LP—more specifically in answer set programming—using defeasible rules and default negation to express principles that tolerate exception. From a different perspective, the experiments with artificial neural networks [25] also reveal that more than one principle may be applicable to



similar cases that differ in a certain aspect (e.g., motives, consequences, etc.), thus upholding morality viewpoints that tolerate exceptions.

While the survey in this chapter shows that several logic-based approaches have been employed in machine ethics, the use of LP has not been much explored in the field despite its potential:

- Like TRUTH-TELLER and SIROCCO, LP permits declarative knowledge representation of moral cases with sufficiently level of detail to distinguish one case from other similar cases. Indeed, except the philosophical approach by Powers [40], all other approaches anchor the experiments to concrete moral cases, indicating that representing moral principles alone is not enough, but the principles need to be materialized into concrete examples. Clearly, the expressivity of LP may extend beyond basic representation of (positive/negative) example facts demonstrated in MEDETHEX and ETHEL.
- Given its declarative representation of moral cases, appropriate LP-based reasoning features can be employed for moral decision making, without being constrained merely to quantitative simplifying assumption (cf. MEDETHEX and ETHEL) and ILP. For instance, the role of LP abduction [27] for decision making in general is discussed in [31]. Indeed, LP abduction has been applied in a variety of areas, such as in diagnosis [24], planning [21], scheduling [29], reasoning of rational agents and decision making [32, 39], knowledge assimilation [28], natural language understanding [10], security protocols verification [1], and systems biology [42]. These applications demonstrate the potential of abduction, and it may as well be suitable for moral decision making, albeit without focusing on learning moral principles. Moral reasoning with quantitative valuation of its elements (such as actions, duties, etc.), either in utility or probability, can still be achieved with other LP-based reasoning features in combination with abduction, e.g., using preferences (see, e.g., [20]) and probabilistic LP (see, e.g., [12, 44]).
- LP provides a logic-based programming paradigm with a number of practical Prolog systems, allowing not only addressing morality issues in an abstract logical formalism (e.g., deontic logic in [40]), but also via a Prolog implementation as proof of concept and a testing ground for experimentation. The use of a theorem prover in [16] to find a proof of an operative moral principle with respect to a particular deontic logic is an attempt to provide such a testing ground for experimentation, albeit not addressing non-monotonic reasoning and moral updating concerns of Powers [40]. The use of LP, without resorting to deontic logic, to model Kant's categorical imperative and non-monotonic reasoning (via default negation), is shown in [23], but no LP updating is considered yet. To this end, a combination of LP abduction and updating may be promising in order to address moral decision making with non-monotonic reasoning and moral updating, in line with the views of Powers [40].
- While logical formalisms, such as deontic logic, permit to specify the notions of obligation, prohibition and permissibility in classic deontic operators, they are not immediately appropriate for representing morality reasoning *processes* studied in philosophy and cognitive science, such as the dual-process model of reactive

and deliberative processes [17, 22, 33, 47]. Advanced techniques in Prolog systems, such as tabling [48–50], open an opportunity to conceptually capture such processes, by appropriately applying it to considered reasoning mechanisms in moral decision making, such as LP abduction and updating.

Given this potential of LP in addressing all the above issues, there is a need to investigate further its potential. But before we do so, we need first to study more results from morality-related fields, such as philosophy and psychology, to better identify some significant moral facets which, at the start, are amenable to computational modeling by LP knowledge representation and reasoning features. This is the subject of the next chapter.

## References

1. Alberti, M., Chesani, F., Gavanelli, M., Lamma, E., Torroni, P.: Security protocols verification in abductive logic programming. In: *Proceedings of the 6th International Workshop on Engineering Societies in the Agents World (ESAW), LNCS*, vol. 3963. Springer (2005)
2. Anderson, M., Anderson, S.L.: EthEl: Toward a principled ethical eldercare robot. In: *Proceeding of the AAAI 2008 Fall Symposium on AI in Eldercare* (2008)
3. Anderson, M., Anderson, S.L.: Robot be good: A call for ethical autonomous machines. *Scientific American* pp. 54–59 (2010)
4. Anderson, M., Anderson, S.L.: GenEth: A general ethical dilemma analyzer. In: *Proceeding of the 28th AAAI Conference on Artificial Intelligence* (2014)
5. Anderson, M., Anderson, S.L.: Toward ensuring ethical behavior from autonomous systems: a case-supported principle based paradigm. In: *Proceeding of the AAAI Workshop on Artificial Intelligence and Ethics (1st International Workshop on AI and Ethics)* (2015)
6. Anderson, M., Anderson, S.L., Armen, C.: Towards machine ethics: implementing two action-based ethical theories. In: *Proceeding of the AAAI 2005 Fall Symposium on Machine Ethics* (2005)
7. Anderson, M., Anderson, S.L., Armen, C.: An approach to computing ethics. *IEEE Intell. Syst.* **21**(4), 56–63 (2006)
8. Anderson, M., Anderson, S.L., Armen, C.: MedEthEx: a prototype medical ethics advisor. In: *Proceeding of the 18th Innovative Applications of Artificial Intelligence Conference (IAAI 2006)* (2006)
9. Arkoudas, K., Bringsjord, S., Bello, P.: Toward ethical robots via mechanized deontic logic. In: *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics* (2005)
10. Balsa, J., Dahl, V., Lopes, J.G.P.: Datalog grammars for abductive syntactic error diagnosis and repair. In: *Proceedings of the Natural Language Understanding and Logic Programming Workshop* (1995)
11. Baral, C.: *Knowledge Representation. Reasoning and Declarative Problem Solving*. Cambridge University Press, New York (2010)
12. Baral, C., Gelfond, M., Rushton, N.: Probabilistic reasoning with answer sets. *Theory Pract. Logic Program.* **9**(1), 57–144 (2009)
13. Beauchamp, T.L., Childress, J.F.: *Principles of Biomedical Ethics*. Oxford University Press, Oxford (1979)
14. Bok, S.: *Lying: Moral Choice in Public and Private Life*. Vintage Books, New York (1989)
15. Branting, L.K.: *Reasoning with Rules and Precedents*. Springer, Netherlands (2000)
16. Bringsjord, S., Arkoudas, K., Bello, P.: Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intell. Syst.* **21**(4), 38–44 (2006)

17. Cushman, F., Young, L., Greene, J.D.: Multi-system moral psychology. In: Doris, J.M. (ed.) *The Moral Psychology Handbook*. Oxford University Press, Oxford (2010)
18. Dancy, J.: Moral particularism. In: E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Fall 2013 edn. Center for the Study of Language and Information, Stanford University <http://plato.stanford.edu/archives/fall2013/entries/moral-particularism/> (2013)
19. Daniels, N.: Reflective equilibrium. In: E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Winter 2013 edn. Center for the Study of Language and Information, Stanford University <http://plato.stanford.edu/archives/win2013/entries/reflective-equilibrium/> (2013)
20. Dell'Acqua, P., Pereira, L.M.: Preferential theory revision. *J. Appl. Log.* **5**(4), 586–601 (2007)
21. Eshghi, K.: Abductive planning with event calculus. In: *Proceedings of the International Conference on Logic Programming*. The MIT Press (1988)
22. Evans, J.S.B.T.: *Thinking Twice: Two Minds in One Brain*. Oxford University Press, Oxford (2010)
23. Ganascia, J.G.: Modelling ethical rules of lying with answer set programming. *Eth. Inf. Technol.* **9**(1), 39–47 (2007)
24. Gartner, J., Swift, T., Tien, A., Damásio, C.V., Pereira, L.M.: Psychiatric diagnosis from the viewpoint of computational logic. In: *Proceedings of the 1st International Conference on Computational Logic (CL 2000)*, LNAI, vol. 1861, pp. 1362–1376. Springer (2000)
25. Guarini, M.: Computational neural modeling and the philosophy of ethics: reflections on the particularism-generalism debate. In: Anderson, M., Anderson, S.L. (eds.) *Machine Ethics*. Cambridge University Press, New York (2011)
26. Horta, J.: *Agency and Deontic Logic*. Oxford University Press, Oxford (2001)
27. Kakas, A., Kowalski, R., Toni, F.: The role of abduction in logic programming. In: Gabbay, D., Hogger, C., Robinson, J. (eds.) *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 5. Oxford University Press, Oxford (1998)
28. Kakas, A.C., Mancarella, P.: Knowledge assimilation and abduction. In: *International Workshop on Truth Maintenance, ECAI 1990* (1990)
29. Kakas, A.C., Michael, A.: An abductive-based scheduler for air-crew assignment. *J. Appl. Artif. Intell.* **15**(1–3), 333–360 (2001)
30. Kant, I.: *Grounding for the Metaphysics of Morals*, translated by Ellington, J., Hackett, Indianapolis (1981)
31. Kowalski, R.: *Computational Logic and Human Thinking: How to be Artificially Intelligent*. Cambridge University Press, New York, NY (2011)
32. Kowalski, R., Sadri, F.: Abductive logic programming agents with destructive databases. *Ann. Math. Artif. Intell.* **62**(1), 129–158 (2011)
33. Mallon, R., Nichols, S.: Rules. In: J.M. Doris (ed.) *The Moral Psychology Handbook*. Oxford University Press, Oxford (2010)
34. McLaren, B.M.: Extensionally defining principles and cases in ethics: an AI model. *Artif. Intell. J.* **150**, 145–181 (2003)
35. McLaren, B.M., Ashley, K.D.: Case-based comparative evaluation in truth teller. In: *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (1995)
36. Muggleton, S.: Inductive logic programming. *New Gener. Comput.* **8**(4), 295–318 (1991)
37. Murakami, Y.: Utilitarian deontic logic. In: *Proceedings of the 5th Advances in Modal Logic conference (AiML)* (2004)
38. National Society of Professional Engineers (NSPE): *The NSPE Ethics Reference Guide*. The National Society of Professional Engineers, Alexandria, VA (1996)
39. Pereira, L.M., Dell'Acqua, P., Pinto, A.M., Lopes, G.: Inspecting and preferring abductive models. In: K. Nakamatsu, L.C. Jain (eds.) *The Handbook on Reasoning-Based Intelligent Systems*, pp. 243–274. World Scientific Publishers (2013)
40. Powers, T.M.: Prospects for a Kantian machine. *IEEE Intell. Syst.* **21**(4), 46–51 (2006)
41. Rawls, J.: *A Theory of Justice*. Harvard University Press, Cambridge (1971)
42. Ray, O., Antoniadis, A., Kakas, A., Demetriades, I.: Abductive logic programming in the clinical management of HIV/AIDS. In: *Proceeding of 17th European Conference on Artificial Intelligence*. IOS Press (2006)

- 43. Reiter, R.: A logic for default reasoning. *Artif. Intell.* **13**, 81–132 (1980)
- 44. Riguzzi, F., Swift, T.: The PITA system: tabling and answer subsumption for reasoning under uncertainty. *Theory Pract. Log. Program.* **11**(4–5), 433–449 (2011)
- 45. Ross, W.D.: *The Right and the Good*. Oxford University Press, Oxford (1930)
- 46. Sinnott-Armstrong, W.: Consequentialism. In: E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Spring 2014 edn. Center for the Study of Language and Information, Stanford University (2014). <http://plato.stanford.edu/archives/spr2014/entries/consequentialism/>
- 47. Stanovich, K.E.: *Rationality and the Reflective Mind*. Oxford University Press, Oxford (2011)
- 48. Swift, T.: Tabling for non-monotonic programming. *Ann. Math. Artif. Intell.* **25**(3–4), 201–240 (1999)
- 49. Swift, T.: Incremental tabling in support of knowledge representation and reasoning. *Theory Pract. Log. Program.* **14**(4–5), 553–567 (2014)
- 50. Swift, T., Warren, D.S.: XSB: extending Prolog with tabled logic programming. *Theory Pract. of Log. Program.* **12**(1–2), 157–187 (2012)

Programming Machine Ethics

Pereira, L.M.; Saptawijaya, A.

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

ISBN: 978-3-319-29353-0