Bridging Two Realms of Machine Ethics

Luís Moniz Pereira

NOVA Laboratory for Computer Science and Informatics

Universidade Nova de Lisboa, Portugal

Ari Saptawijaya

NOVA Laboratory for Computer Science and Informatics

Universidade Nova de Lisboa, Portugal & Universitas Indonesia, Indonesia

ABSTRACT

We address problems in machine ethics dealt with using computational techniques.

Our research has focused on Computational Logic, particularly Logic Programming, and its appropriateness to model morality, namely moral permissibility, its justification, and the dual-process of moral judgments regarding the realm of the individual.

In the collective realm, we, using Evolutionary Game Theory in populations of individuals, have studied norms and morality emergence computationally. These, to start with, are not equipped with much cognitive capability, and simply act from a predetermined set of actions. Our research shows that the introduction of cognitive capabilities, such as intention recognition, commitment, and apology, separately and jointly, reinforce the emergence of cooperation in populations, comparatively to their absence.

Bridging such capabilities between the two realms helps understand the emergent ethical behavior of agents in groups, and implements them not just in simulations, but in the world of future robots and their swarms. Evolutionary Anthropology provides teachings.

Keywords: Machine Ethics, Moral Dilemmas, Computational Logic, Evolutionary Game Theory, Intention Recognition, Commitment, Apology, Emergence of Cooperation, Evolutionary Biology, Mutualism.

INTRODUCTION

Machine ethics (also known as computational morality, machine morality, artificial morality and computational ethics) is a burgeoning field of enquiry that emerges from the need of imbuing autonomous agents with the capacity of moral decision-making. It has particularly attracted interest from the artificial intelligence community and has brought together perspectives from various fields, amongst them: philosophy, cognitive science, neuroscience and primatology. The overall result of this interdisciplinary research is therefore not only important for equipping agents with the capacity of making moral judgments, but also for helping us better understand morality, through the creation and testing of computational models of ethical theories.

Research in artificial intelligence particularly contributes on how techniques from computational logic, machine learning and multi-agent systems, can be employed in order to computationally model, to some improved extent, moral decision-making. In the present chapter we survey problems in machine ethics that have been examined and techniques used in dealing with such problems. Various techniques have been exploited including machine learning, e.g., case-based reasoning, artificial neural networks; and logic-based formalisms, e.g., deontic logic and non-monotonic logics. Our research, in particular, has been focusing on logic programming techniques and their appropriateness to model some morality aspects, namely moral permissibility, its justification, and the dual-process of moral judgments. We argue that the main characteristics of these aspects can be captured by the available ingredients and formalisms based on logic programming. These include, among others, abduction (with integrity constraints), updating, preferences, argumentation, and counterfactual. These ingredients are framed together in an agent life cycle architecture, which allows an agent to make a moral decision by means of abduction (either reactively or deliberatively—the dual-process), respecting its integrity constraints in order to rule out a priori impermissible actions, weighing and preferring decisions after inspecting their consequences, providing arguments to justify moral decisions made, and updating itself either by the changes due to its decisions or by other ethical principles being told or learned. We also touch upon uncertainty and counterfactual reasoning in moral decision-making, and how they fit in our logic programming based agent architecture.

The agent life cycle architecture concerns itself only in realm of the individual, where computation is vehicle for modeling the dynamics of knowledge and moral cognition of an agent. In the collective realm, norms and moral emergence has been studied computationally, using the techniques of Evolutionary Game Theory, 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 has shown that the introduction of cognitive capabilities, such as intention recognition, commitment, and apology, separately and jointly, reinforce the emergence of cooperation in the population, comparatively to the absence of such cognitive abilities. We discuss how modeling moral cognition in individuals (using the aforementioned ingredients of logic programming) 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. 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.

This chapter hence contemplates two distinct realms of machine ethics, to wit, the individual and collective, and identified needed bridges concerning their connection. 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.

Presently, machine ethics is becoming an ever more pressing concern, as machines become ever more sophisticated, autonomous, and act in groups, among populations of other machines and of humans. Ethics and jurisprudence, and hence legislation, are however lagging much behind in adumbrating the new ethical issues arising from these circumstances.

Meanwhile, research in machine ethics with the purpose of understanding each of the two realms, has been fostering inroads and producing results in each. Namely, our co-authors and we have staked footholds on either side of the two realms gap, and promoted their mutually beneficial bridging. Evolutionary Biology, Anthropology and the Cognitive Sciences providing inspirational teachings to that effect.

The chapter is naturally organized as follows. First we summarize the topics and our research results in the individual realm of machine ethics, and next comes a survey of the topics and our

research results in the collective realm of machine ethics. There ensues the bridging of these two realms in machine ethics. Last but not least, we ponder over the teachings of human moral evolution in this regard. A final coda foretells a road to be tread, and portends about ethical machines and us.

THE INDIVIDUAL REALM OF MACHINE ETHICS

Research in machine ethics have mainly centered on equipping agents with particular ethical theories, e.g., utilitarianism and deontological ethics, and on providing a framework to encode moral rules, typically in favor of deontological ethics, with or without referring to specific moral rules. In so doing, various techniques have been employed, including machine learning (e.g., case-based reasoning, artificial neural networks) and logic-based formalisms (e.g., deontic logic, non-monotonic logics, abductive logic programming).

Computational Approaches in Machine Ethics

Jeremy is an advisor system that follows Jeremy Bentham's act utilitarianism (Anderson, Anderson, & Armen, 2005). Moral decisions are made based on the calculation of a total net pleasure that depends on three considered components with respect to each affected person: the intensity of pleasure/displeasure, the duration of the pleasure/displeasure, and the probability that this pleasure/displeasure will occur. The "right" decision is determined by that giving the highest total net pleasure. The calculation formula in *Jeremy* is later extended to capture prima facie duty theory (Ross, 1930). Two other advisor systems, viz., *MedEthEx* (Anderson, Anderson, & Armen, 2006) and *EthEl* (Anderson & Anderson, 2008), are also based on the same theory in biomedical ethics. *MedEthEx* is dedicated to give advice for dilemmas in biomedical fields, while *EthEl* serves as a medication-reminder system for the elderly and as a notifier to an overseer if the patient refuses to take the medication. For these purposes, both systems benefit from machine learning techniques, viz., inductive logic programming. The latter system has been deployed in the *Nao* robot, being capable to serve patients who need to be reminded of medication, and to bring them their medication (Anderson & Anderson, 2010).

Different machine learning techniques are also used in machine ethics, viz., case-based reasoning and artificial neural networks. Case-based reasoning is employed in *TruthTeller* and *SIROCCO* systems (McLaren, 2006). Though both systems implement casuistry ethical approach (Jonsen & Toulmin, 1988), they have different purposes. *TruthTeller* is designed to accept a pair of ethical dilemmas and describe the salient similarities and differences between the cases, from both an ethical and a pragmatic perspectives, whereas *SIROCCO* to accept an ethical dilemma and to retrieve similar cases and ethical principles relevant to the presented ethical dilemma. For a distinct purpose, artificial neural networks are utilized in Guarini (2011) to understand morality from the philosophy of ethics viewpoint, particularly by exploring the dispute between moral particularism and generalism. Therein, moral situations are classified by training simple recurrent networks with a number of cases, involving actions concerning killing and allowing to die, and then using the trained networks to classify test cases.

Besides machine learning techniques, there has been a growing interest of employing logic-based formalisms in machine ethics. Powers (2006) considers several formalisms to formulate Kant's categorical imperative for the purpose of machine ethics (though only abstractly, as no implementation seems to exist on top of the considered formalisms). With respect to the formulation, three views are taken into account: mere consistency, common-sense practical reasoning, and coherency. To realize the first view, a form of deontic logic is adopted. The second view benefits from non-monotonic logic, and the third view presumes ethical deliberation to follow a logic similar to that of belief revision.

The use of deontic logic as a framework to express ethical codes is explored in Bringsjord, Arkoudas, and Bello (2006). In particular, an axiomatized utilitarian deontic logic (Murakami, 2004) is employed to decide an operative ethical code from several other candidates, by seeking a proof for the expected moral outcome that follows from these candidates. Wiegel (2007) extends the Belief-Desire-Intention (BDI) model (Bratman, 1987) with another variant of deontic logic, viz., the deontic-epistemic-action logic (van den Hoven & Lokhorst, 2002), in order to make BDI suitable for modeling moral

agents. The result is *SophoLab*, a framework for experimental computational philosophy, which is implemented with the JACK agent programming language. This framework is particularly used to study negative moral commands and two different utilitarian theories, viz., act and rule utilitarianism. Other use of BDI in machine ethics is reported in Ganascia (2012), where it is used to model a consequentialist approach, viz., by choosing the action of which consequences are the lesser evil.

All these works with logic-based formalisms share the view that logical systems are appropriate to formalize ethical codes. Taking this view into account, a formal framework to reason over logical systems is proposed in Bringsjord et al. (2011) by employing category theory. The work is strongly based on Piaget's position (Inhelder & Piaget, 1958). This idea of reasoning *over*—instead of reasoning *in*—logical systems, favors post-formal Piaget's stages beyond his well-known fourth stage. In other words, category theory is used as the meta-level of moral reasoning.

Logic Programming for Machine Ethics

Our research in the field has been focusing on the use of Logic Programming (LP). Given its solid theoretical results, LP is mature enough by now, supported by a number of advanced features and practical systems. Kowalski (2011) provides a good overview and presents convincing arguments on the suitability of LP for machine ethics. We have been exploring morality issues to come up with those that, in our view, are amenable to computational modeling by benefiting from LP features, like abduction, updating, preferences, etc. For a recapitulation and more pointers to our prior work see (Saptawijaya & Pereira, in press).

One morality issue that we have addressed with LP-based approaches is moral permissibility, by modeling classic moral examples from literature. In Pereira and Saptawijaya (2007a, 2007b) we have shown that several LP features can be employed together in an integrated system, *ACORDA* (Lopes & Pereira, 2006), to model permissibility in various scenarios of the classic trolley problem (Foot, 1967) with the Doctrine of Double Effect (DDE) as the basis of moral decisions in these scenarios. Indeed, DDE is often referred to when explaining the permissibility of an action by distinguishing whether its harm consequence is merely a *side-effect* of achieving a good result, or rather a *means* to bringing about the same good end (McIntyre, 2004). Such reference does not only appear in philosophy literature, but is also considered in psychology experimental studies. For instance, Hauser, Cushman, Young, Jin, & Mikhail (2007) reports that subjects from demographically diverse populations share the consistency of judgments regarding permissibility on a series of moral dilemmas. In their study, while a majority of subjects fail to provide justifications to their judgments, these judgments are consistent with DDE.

Our LP-based approach to machine ethics is primarily supported by abduction. In the philosophy of science, abduction is commonly understood as a reasoning method to infer the best-preferred explanation to observed evidence. In LP, abduction does not necessarily restrict itself to the specific task of explaining observations. Instead, it more generally translates into finding consistent abductive solutions to a goal, whilst satisfying integrity constraints, where a goal typically refers to a desired future state of the environment. In this case observations are simply given as facts that do not need explanations. An abductive solution, built from abductive hypotheses (called *abducibles*), is a set of abduced actions that achieve the goal. A goal itself can be empty, and if so, abduction amounts to satisfying integrity constraints only.

LP abduction is typically accomplished by a top-down goal-oriented procedure for finding, by need, an abductive solution to the goal. For that reason our abduction mechanism is based on the well-founded semantics of LP (van Gelder, Ross, & Schlipf, 1991), that permits finding just relevant abducibles, along with their truth value, whereas those not mentioned in the solution are indifferent to the goal. Nevertheless, other LP semantics can also be useful, e.g., stable models semantics (Gelfond & Lifschitz, 1988) can be utilized to compute the consequences of abductive solutions. These consequences may serve as some criteria to prefer among abductive solutions, as explained below.

In Pereira and Saptawijaya (2007a, 2007b), possible decisions in various scenarios of the trolley problem, e.g., diverting the trolley, pushing a man, etc., are represented as abducibles. Furnishing all observed possible outcomes as goal, and stipulating the consequences of impermissible actions (in accordance to DDE) as some integrity constraint, the abduction mechanism returns all permissible actions that satisfy some given goal and do not violate the integrity constraint. Abductive solutions as permissible moral decisions can be further filtered. For this purpose our approach benefits from preferences in LP (Dell'Acqua & Pereira, 2007), where a posteriori preferences are applied to prefer eventual moral decisions. This is realized, e.g., by examining their consequences and applying utility functions to them.

The integrated LP-based approach shows that it successfully delivers moral decisions for these various scenarios of the trolley problem, which moreover conform to the experimental study by Hauser et al. (2007). The work is further extended in Pereira and Saptawijaya (2009, 2011) using similar scenarios of the trolley problem but considering additionally another moral principle, viz., Doctrine of Triple Effect (DTE). DTE (Kamm, 2006) refines DDE, particularly on the notion about harming someone as an intended means: it distinguishes further between doing an action *in order* that an effect occurs and doing it *because* that effect will occur. This extended work shows that the same LP-based approach is able to express different outcomes between DDE and DTE on relevant scenarios of the trolley problem, viz., the Loop case (Thomson, 1985) and the Loop-Push case. In these two cases the same initial setting applies: *A trolley is headed toward five people walking on the track, and they will not be able to get off the track in time. The trolley can be redirected onto a side track, which loops back towards the five.* In the Loop case the setting is further completed as follows: *A fat man sits on this looping side track, so fat that his body will by itself stop the trolley, thereby saving the five*. While diverting the trolley is morally impermissible in DDE, it is permissible by DTE. According to DTE, it is permissible because it will hit the man, and not in order to intentionally hit him (Kamm, 2006). This is consistent with the opinion of most moral philosophers as well as with the psychology experimental result of Hauser et al. (2007). The Loop-Push case is a variant of the Loop one, where the looping side track is initially empty, and besides the diverting action, an ancillary action of pushing a fat man in order to place him on the side track is additionally performed. For the latter case, both DTE agrees with DDE that such a deliberate action (pushing) performed in order to bring about harm (the man hit by the trolley), even for the purpose of a good or greater end (to save the five), is likewise impermissible.

We have recently further explored the appropriateness of LP to express different views on moral permissibility with respect to DDE and DTE, by means of a LP-based approach of counterfactuals (Pereira & Saptawijaya, 2014). People are naturally engage counterfactual thoughts in moral situations, as they tend to reason about what they should or should not have done when they contemplate alternative decisions in such situations. This is particularly related to the evaluation feature of counterfactuals. Moreover, counterfactuals permit momentary experiential simulation of the possible alternatives, through their reflective nature (Epstude & Roese, 2008), thereby allowing careful consideration before a moral decision is made, and to subsequently justify it. A number of psychology experimental studies on counterfactuals in the context of moral reasoning have also been conducted, e.g., by McCloy and Byrne (2000) and Migliore, Curcio, Mancini, and Cappa (2014). These studies and others indicate prospects for counterfactuals in machine ethics that have never been explored.

Our LP-based method to evaluating counterfactuals is inspired by Pearl's structure-based counterfactuals (Pearl, 2009), itself based on probabilistic causal model and a calculus of intervention. We resort to LP abduction and updating in mirroring Pearl's approach, but abstain from probabilities in order to concentrate on people's naturalized logic. Our work using probabilistic LP moral reasoning is reported elsewhere (Han, Saptawijaya, & Pereira, 2012), where uncertainty of actions and consequences is taken into account in judging moral permissibility, both from the view of oneself and from that of others.

In our LP-based counterfactual approach, abduction hypothesizes background conditions from observations made or evidences given, whereas LP updating fixes the initially abduced context of the

counterfactual being evaluated. Moreover, LP updating facilitates a minimal adjustment to the causal model (in this case, the logic program) by hypothetical updates of causal intervention through defeasible rules. The combination of both LP features establishes a procedure that corresponds to Pearl's counterfactual approach. The procedure can be summarized in three steps as follows. First, abduction is enacted to explain the current observation. The explanation fixes the abduced context in which the counterfactual is evaluated by means of LP updating. Second, the causal intervention is realized by hypothetical updates. In the presence of defeasible LP rules these updates permit hypothetical modification of the program to consistently comply with the antecedent of the counterfactual. Third, the well-founded model (van Gelder, Ross, & Schlipf, 1991) of the hypothetical modified program is examined to verify whether the consequence of the counterfactual holds true at the current state.

In order to examine permissibility of an action in DDE, a form of counterfactuals that is able to distinguish between an instrumental cause and a side-effect can be introduced: *If E would not have been true, then G would not have been true*. The evaluation of this counterfactual form identifies permissibility of action from its morally wrong effect (say, a harm) *E*, by identifying whether *E* is a necessary cause for achieving a good end (a goal) *G* or instead a mere side-effect of that action. If the counterfactual is valid, then *E* is instrumental as a cause of *G*, and not a mere side-effect of the action. Since *E* is morally wrong, achieving *G* that way, by means of that action, is impermissible; otherwise, it is not. We have shown in Pereira and Saptawijaya (2014) that this counterfactual form is general enough to examine permissibility of actions in a number of classic moral problems, such as in military cases, e.g., tactical vs. terror bombing (Scanlon, 2008) and relevant scenarios of the trolley problem, e.g., the previously mentioned Loop and Loop-Push cases.

In the Loop case, proving the validity of the counterfactual *"if the man had not been hit by the trolley, the five people would not have been saved"* is sufficient to show that the harm event of the man hit by the trolley is instrumental as an instrumental cause for the goal of saving the five; hence diverting the trolley is DDE morally impermissible. From the DTE viewpoint, two counterfactuals are evaluated. First, the validity of the counterfactual *"if the man had not been on the side track, then he would not have been hit by the trolley"* is verified, ensuring that the unfortunate event of the man being hit by the trolley is indeed the consequence of the man being on the side track. Second, a hypothetical ancillary action, pushing, is assumed to place the man on the side track, and the counterfactual *"if the man had not been pushed, then he would not have been hit by the trolley"* is examined. The latter counterfactual is not valid, because pushing is not true in the abduced context where the counterfactual is evaluated. It signifies that even without this hypothetical but unexplained deliberate action of pushing, the man would still have been hit by the trolley (just because he is already on the side track). Therefore, though the harm event of the man being hit is a consequence of diverting the trolley and instrumental in achieving the goal of saving the five, no deliberate action is required to cause the man placed on the side track, in order for the harm event to occur. Hence div is DTE morally permissible.

In the Loop-Push case, where the deliberate pushing action is abduced (in addition to diverting the trolley), the counterfactual *"if the man had not been hit by the trolley, the five people would not have been saved"* previously evaluated in the DDE Loop case is still valid. Moreover, the counterfactual *"if the man had not been pushed, then he would not have been hit by the trolley"* is now valid, due to the newly abduced pushing action. From the validity of both these counterfactuals one can infer that, given the trolley diverting action, the ancillary action of pushing the man onto the side track causes him to be hit by the trolley, which in turn causes the five to be saved. In the Loop-Push, DTE agrees with DDE that such a deliberate action (pushing) performed in order to bring about harm (the man hit by the trolley), even for the purpose of a good or greater end (to save the five), is likewise impermissible.

According to Scanlon (2008), the appeal of DDE and DTE to explain moral judgments in the trolley problem and other similar dilemmas is due to the so-called critical employment of moral judgments. Furthermore, Scanlon argues that moral permissibility can differently be assessed through the so-called *deliberative* employment of moral judgments. According to Scanlon, the deliberative

employment concerns answering the question of the permissibility of actions, by identifying the justified but defeasible argumentative considerations, and their exceptions, which make actions permissible or impermissible. That is, moral dilemmas typically have the same structure: (1) they concern general principles that in some cases admit exceptions, and (2) they raise questions about when those exceptions apply. In other words, an action can be determined impermissible through deliberative employment when there is no countervailing consideration that would justify an exception to the applied general principle. Indeed, this deliberative employment is in line with Scanlon's contractualism (Scanlon, 1982). Contractualism provides flexibility on the set of principles to justify moral judgments so long as no one could reasonably reject them. Reasoning is an important aspect here, as argued in Scanlon (1998), in that making judgments does not seem to be merely relying on internal observations but is achieved through reasoning. Hence, method of reasoning is one of primary concerns of contractualism in providing justification to others, by looking for some common ground that others could not reasonably reject. In this way, morality can be viewed as (possibly defeasible) argumentative consensus, which is why contractualism is interesting from the Artificial Intelligence perspective

The deliberative employment of moral judgments to determine permissibility of actions opens up another venue where LP may play its role. On the one hand, defeasible rules in LP updating can conveniently represent exceptions to a principle, thereby addressing point (1) in the previous paragraph. See also Ganascia (2007) for an alternative use of answer set programming (a LP paradigm based on stable model semantics) for addressing the same purpose. On the other hand, LP argumentation (see Rahwan and Simari (2009) for a general survey) provides a way to reach an agreement on whether or not countervailing considerations can be justified, addressing point (2). In fact, counterfactuals may also be appropriate to provide an argument for justifying moral judgments, through 'compound counterfactuals': *"Had I known what I know today, then if I were to have done otherwise, something preferred would have followed."* Such counterfactuals, typically imagining alternatives with worse effect—the so-called *downward counterfactuals* (Markman, Gavanski, Sherman, & McMullen, 1993)—, may provide moral justification for what was performed due to lack of the current fuller knowledge. This is accomplished by evaluating what would have followed if the intent would have been otherwise, other things (including present knowledge) being equal. It may justify that what would have followed is no morally better than the actual ensued consequence.

We have demonstrated the roles of LP updating with its defeasible rules, both in realizing causal intervention to evaluating counterfactuals and in expressing exceptions for the deliberative employment of moral judgments. Obviously, LP updating is appropriate for representing changes and for dealing with incomplete information. To this end, LP updating can be employed for moral updating, viz., the adoption of new (possibly overriding) ethical rules on top of those an agent currently follows. Such adoption is often necessary when the ethical rules one follows have to be revised in the light of situations faced by the agent, e.g., whenever some authority contextually imposes other ethical rules. We have shown the applicability of LP updating together with other features discussed here (LP abduction and preferences) for moral updating via an interactive storytelling (Lopes & Pereira, 2010).

Individual Realm Concluding Remarks

Having been placed on the back burner, the prospect of LP has stimulated us now to rethink how its features can approach issues in machine ethics. Here we particularly refer to the realm of the individual agent, i.e., to endow machines with the capability to declaratively represent ethical situations so they can reason on ethical issues arising from such situations. Though we are still at an early stage of our journey, we have exhibited in our works the successful interplay of various LP features in tackling a number of morality issues.

This interplay is evident in several implemented systems we have employed in our works of machine ethics. *ACORDA* (Lopes & Pereira, 2006), used in our initial work of DDE and DTE permissibility (Pereira & Saptawijaya, 2007a, 2007b, 2009, 2011) and in interactive moral storytelling

(Lopes & Pereira, 2010), benefits from LP abduction, updating, and preferences. Its subsequent reincarnation, *Evolution Prospection Agent (EPA) system* (Pereira & Han, 2009a), benefits from the same LP features, but with the dual program transformation (Alferes, Pereira, & Swift, 2004) for its abduction mechanism (instead of an ad-hoc one, as in *ACORDA*). It is also later equipped with the capability to reason under uncertainty via an implementation of the probabilistic logic programming language *P-log* (Baral, Gelfond, & Rushton, 2009). The *EPA* system has been used in our work on intention recognition (Pereira & Han, 2011) and probabilistic moral reasoning (Han, Saptawijaya, & Pereira, 2012). For our recent work on counterfactuals, we benefit from *QUALM* (available from http://goo.gl/XLhBxO), which is built on top of an integrated LP abduction and incremental updating (Saptawijaya & Pereira, 2014), comprising tabling mechanisms (Swift & Warren, 2012).

We mention in Pereira and Saptawijaya (2014) how compound counterfactual benefits from the incremental tabling in LP updating (Saptawijaya & Pereira, 2013b) of *QUALM*. Tabling may also be useful in modeling the dual-process of moral decision making, i.e., the interaction between deliberative and reactive processes in moral decision making (Cushman, Young, & Greene, 2010). Deliberative reasoning in *QUALM* is induced by abduction. Because we employed tabling for contextual abduction (Saptawijaya & Pereira, 2013a), abductive solutions (e.g., actions/decisions to some goals in a moral situation) are stored for future use, possibly in different context. Reactive processes can therefore benefit from it, since decisions are readily available for reuse in the present context, without the need to deliberatively re-compute them. Furthermore, though only reactively obtained, these tabled decisions can be deliberatively re-evaluated with the rules that support them (cf. the notions of expectation and contra-expectation in hypotheses generation (Pereira, Dell'Acqua, Pinto, & Lopes, 2013)), so as to provide a form of argumentation between agents about their decisions.

THE COLLECTIVE REALM OF MACHINE ETHICS

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 (EP). Their systematic study resorts to simulation techniques, thus enabling the study of aforesaid mechanisms under a variety of conditions, parameters, and alternative virtual games. The theoretical and experimental results have continually been surprising, rewarding, and promising. For a background on EGT and its use by EP we refer to Pereira (2012a).

In recent work, one of us (Pereira and the mentioned co-authors) has initiated the introduction, in such groups of individuals, of cognitive abilities inspired on techniques and theories of Artificial Intelligence, namely those pertaining to Intention Recognition, Commitment, and Apology (separately and jointly), encompassing errors in decision-making and communication noise. As a result, both the emergence and stability of cooperation become reinforced comparatively to the absence of such cognitive abilities. This holds separately for Intention Recognition, for Commitment, and for Apology, and even more so when they are jointly engaged.

This section aims to sensitize the reader to these Evolutionary Game Theory based issues, results and prospects, which are accruing in importance for the modeling of minds with machines, with impact on our understanding of the evolution of mutual tolerance and cooperation, and of the arising of moral norms. Recognition of someone's intentions, which may include imagining the recognition others have of our own intentions, and may comprise not just some error tolerance, but also a penalty for unfulfilled commitment though allowing for apology, can lead to evolutionary stable win/win equilibriums within groups of individuals, and perhaps amongst groups. The recognition and the manifestation of intentions, plus the assumption of commitment—even whilst paying a cost for putting it in place—and the acceptance of apology, are all facilitators in that respect, each of them singly and, above all, in collusion.

Emergence of Cooperation via Intention Recognition, Commitment, and Apology

In collective strategic interaction, wherein multiple agents pursue individual strategies, conflicts will arise because the actions of individual agents may have an effect on the welfare of others, and on their own in return (Han, Pereira, Santos, & Lenaerts, 2014). Hence, in these situations the need arises for the regulation of individual and collective behavior, traditionally having followed two distinct approaches, well-known in the Economics and Artificial Intelligence literature (Groves, 1973; Myerson, 1979; Axelrod, 1986; McAfee, 1993; Jackson, 2000; Nisan & Ronen, 1999; Naor, Pinkas, & Sumner, 1999; Ross, 2005; Phelps, McBurney, & Parsons, 2010): the spontaneous emergence of order approach, which studies how norms result from endogenous agreements among rational individuals, and the mechanism by design approach, which studies how norms are exogenously imposed in order to attain desirable properties of the whole.

In this summary, we describe the main results we have obtained following essentially the former approach, but crucially complementing it in instilling some individual agents with cognitive abilities that can and will induce cooperation in the population. These abilities enable such individuals to recognize the opportunity whether to decide to cooperate outright, or possibly propose costly cooperation commitments, susceptible to compensation on defaulting, and to accept apology-redressing dues. In consequence, norm-based cooperation can evolve and emerge.

The problem of evolution of cooperation and of the emergence of collective action—cutting across areas as diverse as Biology, Economy, Artificial Intelligence, Political Science, or Psychology—is one of the greatest interdisciplinary challenges science faces today (Hardin, 1968; Axelrod, 1984; Nowak, 2006a; Sigmund, 2010). To understand the evolutionary mechanisms that promote and keep cooperative behavior among individuals is all the more complex as increasingly intricate is the intrinsic complexity of those individuals partaking of the cooperation.

In its simplest form, a cooperative act is metaphorically described as the act of paying a cost to convey a benefit to someone else. If two players simultaneously decide to cooperate or not, the best possible response will be to try to receive the benefit without paying the cost. In an evolutionary setting, we may also wonder why would natural selection equip selfish individuals with altruistic tendencies while it incites competition between individuals and thus apparently rewards only selfish behavior? Several mechanisms responsible for promoting cooperative behavior have been recently identified (Sigmund, 2010; Nowak, 2006b). From kin and group ties, to different forms of reciprocity and networked populations, several aspects have been shown to play an important role in the emergence of cooperation (see survey in (Sigmund, 2010; Nowak, 2006b)).

Moreover, more complex strategies based on the evaluation of interactions between third parties allow the emergence of kinds of cooperation that are immune to exploitation because then interactions are channeled to just those who cooperate. Questions of justice and trust, with their negative (punishment) and positive (help) incentives, are fundamental in games with large diversified groups of individuals gifted with intention recognition capabilities. In allowing them to choose amongst distinct behaviors based on suggestive information about the intentions of their interaction partners—these in turn influenced by the behavior of the individual himself—individuals are also influenced by their tolerance to error or noise in the communication. One hopes that, to start with, understanding these capabilities can be transformed into mechanisms for spontaneous organization and control of swarms of autonomous robotic agents (Bonabeu, Dorigo, & Theraulaz, 1999), these being envisaged as large populations of agents where cooperation can emerge, but not necessarily to solve a priori given goals, as in distributed Artificial Intelligence (AI).

With these general objectives, we have specifically studied the way players' strategies adapt in populations involved in cooperation games. We used the techniques of Evolutionary Game Theory (EGT) (Hofbauer & Sigmund, 1998; Sigmund, 2010), considered games such as the Prisoner's Dilemma and Public Goods Game (Hofbauer & Sigmund, 1998; Sigmund, 2010), and showed how the actors

participating in repeated iterations in these games can benefit from having the ability to recognize the intentions of other actors, to apologize when making mistakes, to establish commitments, or to combine some of them, thereby leading to an evolutionary stable increase in cooperation (Han, Pereira, & Santos, 2011a, 2012a, 2012b, 2012c; Han, Pereira, Santos, & Lenaerts, 2013a; Han, 2013), compared to extant best strategies.

In this section we summarize our recent publications on how intention recognition, commitment arrangement and apology can, separately and jointly, lead to the evolution of high levels of cooperation. We discuss how these works provide useful insights for mechanism design in Multi-agent Systems for regulative purposes. Evolutionary emergent futures is what we have studied, tied to the co-presence of fixed strategies in agents, though an agent may replace its strategy by a more advantageous one on occasion (social learning). We have not yet made a strategy also evolve by adopting features of other strategies into its own, through rule-defined strategies updating, which could be a direction for Multi-agent Systems (MAS).

Intention recognition promotes the evolution of cooperation

The ability of recognizing (or reading) intentions of others has been observed and shown to play an important role in many cooperative interactions, both in humans and primates (Tomasello, 2008; Meltzoff, 2005; Ran, Fudenberg, & Dreber, 2013). However, most studies on the evolution of cooperation, grounded on evolutionary dynamics and game theory, have neglected the important role played by a basic form of intention recognition in behavioral evolution. In Han et al. (2011a, 2012a), we have addressed explicitly this issue, characterizing the dynamics emerging from a population of intention recognizers.

In that work, intention recognition (IR) was implemented using Bayesian Networks (BN) (Pereira & Han, 2009b, 2011; Han et al., 2011a), taking into account the information of current signals of intent, as well as the mutual trust and tolerance accumulated from previous one-on-one play experience—including how my previous defections may influence another's intent—but without resorting to information gathered regarding players' overall reputation in the population.

A player's present intent can be understood here as how he's going to play the next round with me, whether by cooperating or defecting (Han et al., 2011a). Intention recognition can also be learnt from a corpus of prior interactions among game strategies (Han et al., 2011b, 2012a), where each strategy can be envisaged and detected as players' (possibly changing) intent to behave in a certain way (Han & Pereira, 2011). In both cases, we experimented with populations with different proportions of diverse strategies in order to calculate, in particular, what is the minimum fraction of individuals capable of intention recognition for cooperation to emerge, invade, prevail, and persist.

Intention recognition techniques have been studied actively in AI for several decades (Charniak & Goldman, 1993; Sadri, 2011), with several applications such as for improving human-computer interactions, assistive living and teamwork (Lesh, 1998; Pereira & Han, 2011; Roy, Bouchard, Bouzouane, & Giroux, 2007; Heinze, 2003). In most of these applications the agents engage in repeated interactions with each other. Our results suggest that equipping the agents with an ability to recognize intentions of others can improve their cooperation and reduce misunderstanding that can result from noise and mistakes.

Commitments promote the emergence of cooperation

Agents make commitments towards others when they give up options in order to influence others. Most commitments depend on some incentive that is necessary to ensure that an action (or even an intention) is in the agent's interest and thus will be carried out in the future (Gintis, 2001). Asking for prior commitments can just be used as a strategy to clarify the intentions of others, whilst at the same time manifesting our own. All parties then clearly know to what they commit and can refuse such a

commitment whenever the offer is made. A classical example of such an agreement is marriage. In that case mutual commitment ensures some stability in the relationship, reducing the fear of exploitation and providing security against potential cataclysms.

In our recent works (Han et al., 2012b, 2013a) we investigate analytically and numerically whether costly commitment strategies, in which players propose, initiate and honor a deal, are viable strategies for the evolution of cooperative behavior, using the symmetric one-shot Prisoner's Dilemma (PD) game to model a social dilemma. Next to the traditional cooperate (C) and defect (D) options, a player can propose its co-player to commit to cooperation before playing the PD game, willing to pay a personal cost to make the proposal credible. If the co-player accepts the arrangement and also plays C, they both receive their rewards for mutual cooperation. Yet if the co-player plays D, then he or she will have to provide the proposer with compensation at a personal cost. Finally, when the co-player does not accept the deal, the game is not played and hence both obtain no payoff. Several free-riding strategies were included in the model, including (i) the fake committers, who accept a commitment proposal yet defect when playing the game, assuming that they can exploit the proposers without suffering a too severe consequence; and (ii) the commitment free-riders, who defect unless being proposed a commitment, which they then accept and cooperate afterwards in the PD game. In other words, these latter players are willing to cooperate when a commitment is proposed but are not prepared to pay the cost of setting it up.

We have shown that when the cost of arranging a commitment is justified with respect to the benefit of cooperation, substantial levels of cooperation can be achieved, especially when one insists on sharing the arrangement cost. On the one hand, such commitment proposers can get rid of fake committers by proposing a strong enough compensation cost. On the other hand, they can maintain a sufficient advantage over the commitment free riders, because a commitment proposer will cooperate with players alike she, while the latter defect among themselves. We have also compared the commitment strategy with the simple costly punishment strategy, where no prior agreements are made. The results show that the first strategy leads to a higher level of cooperation than the latter one.

Economical use of costly commitment via intention recognition

Commitments have been shown to promote cooperation if the cost of arranging them is justified with respect to the benefit of cooperation. But commitment may be quite costly, which leads to the possible prevalence of commitment free-riders (Han et al., 2013a). Hence, it should be avoided when necessary. On the other hand, there are many cases where it is difficult to recognize the intention of another agent with sufficient confidence to make any decision based on it. One may have insufficient information for making the prediction (not enough actions being observed, such as in the first interaction scenario), or even one may know the agent well, but also know that the agent is very unpredictable. In such cases, the strategy of proposing a commitment, or manifesting an intention, can help to impose or clarify intentions of others. In addition, intention is usually defined as choice with commitment (Cohen & Levesque, 1990; Bratman, 1987; Roy, 2009). That is, once the agent intends to do something, it must settle on some state of affairs for which to aim, because of its resource limitation and in order to coordinate its future actions. Deciding what to do establishes a personal form of commitment (Cohen & Levesque, 1990; Roy, 2009). Proposing a commitment deal to another agent consists in asking it to express or clarify its intended decisions.

In a marriage commitment, by giving up the option to leave the other, spouses gain security and an opportunity for a much deeper relationship that would be impossible otherwise (Nesse, 2001a; Frank, 2001), as it might be risky to assume a partner's intention of staying faithful without the commitment of marriage. A contract is another popular kind of commitment, e.g. for an apartment lease (Frank, 2001). When it is risky to assume another agent's intention of being cooperative, arranging an appropriate contract provides incentives for cooperation. However, for example in accommodation rental, a contract is not necessary when the cooperative intention is of high certainty, e.g. when the business affair is between close friends or relatives. It said arranging a commitment deal can be useful to encourage

cooperation whenever intention recognition is difficult, or cannot be performed with sufficiently high certainty. On the other hand, arranging commitments is not free, and requires a specific capacity to set it up within a reasonable cost (for the agent to actually benefit from it) (Nesse 2001a, 2001b)—therefore it should be avoided when opportune to do so.

With such motivations in mind, in our work (Han et al., 2012c; Han, 2013) we showed that if the player first predicts the intentions of a co-player and proposes commitment only when they are not confident about their intention prediction, it can significantly facilitate the conditions for cooperation to emerge. The improvement (in level of cooperation) is most significant when it is costly to arrange commitments and when the cooperation is highly beneficial.

In short, it seems to us that intention recognition, and its use in the scope of commitment, is a foundational cornerstone where we should begin at, naturally followed by the capacity to establish and honor commitments, as a tool towards the successive construction of collective intentions and social organization (Searle, 1995, 2010). Finally, one hopes that understanding these capabilities can be useful in the design of efficient self-organized and distributed engineering applications (Bonabeau, Dorigo, & Theraulaz, 1999), from bio- and socio-inspired computational algorithms, to swarms of autonomous robotic agents.

Apology in committed vs. commitment-free repeated interactions

Apology is perhaps the most powerful and ubiquitous mechanism for conflict resolution (Abeler, Calaki, Andree, & Basek, 2010; Ohtsubo & Watanabe, 2009; Fischbacher and Utikal, 2013), especially among individuals involving in long-term repeated interactions (such as a marriage). An apology can resolve a conflict without having to involve external parties (e.g. teachers, parents, courts), which may cost all sides of the conflict significantly more. Evidence supporting the usefulness of apology abounds, ranging from medical error situations to seller-customer relationships (Abeler, Calaki, Andree, & Basek, 2010). Apology has been implemented in several computerized systems such as human-computer interaction and online markets so as to facilitate users' positive emotions and cooperation (Tzeng, 2004; Utz, Matzat, & Snijders, 2009).

The iterated Prisoner's Dilemma (IPD) has been the standard model to investigate conflict resolution and the problem of the evolution of cooperation in repeated interaction settings (Axelrod, 1984; Sigmund, 2010). This IPD game is usually known as a story of tit-for-tat (TFT), which won both Axelrod's tournaments (Axelrod, 1984). TFT cooperates if the opponent cooperated in the previous round, and defects if the opponent defected. But if there can be erroneous moves due to noise (i.e. an intended move is wrongly performed), the performance of TFT declines, because an erroneous defection by one player leads to a sequence of unilateral cooperation and defection. A generous version of TFT, which sometimes cooperates even if the opponent defected (Nowak & Sigmund, 1992), can deal with noise better, yet not thoroughly. For these TFT-like strategies, apology is modeled implicitly as one or more cooperative acts after a wrongful defection.

In our recent work (Han, Pereira, Santos, & Lenaerts, 2013b), we describe a model containing strategies that explicitly apologize when making an error between rounds. An apologizing act consists in compensating the co-player an appropriate amount (the higher the more sincere), in order to ensure that this other player cooperates in the next actual round. As such, a population consisting of only apologizers can maintain perfect cooperation. However, other behaviors that exploit such apology behavior could emerge, such as those that accept apology compensation from others but do not apologize when making mistakes (fake apologizers), destroying any benefit of the apology behavior. Resorting to the Evolutionary Game Theory (Sigmund, 2010), we show that when the apology occurs in a system where the players first ask for a commitment before engaging in the interaction (Han et al., 2012b, 2012c; Han et al., 2013a; Han, 2013), this exploitation can be avoided. Our results lead to the following conclusions: (i) Apology alone is insufficient to achieve high levels of cooperation; (ii) Apology supported by prior

commitment leads to significantly higher levels of cooperation; (iii) Apology needs to be sincere to function properly, whether in a committed relationships or commitment-free ones (which is in accordance with existing experimental studies, e.g. in Ohtsubo and Watanabe (2009); (iv) A much costlier apology tends to be used in committed relationships than in commitment-free ones, as it can help better identify free-riders such as fake apologizers: *commitments bring about sincerity*.

As apology (Tzeng, 2004; Utz, Matzat, & Snijders, 2009) and commitment (Winikoff, 2007; Wooldridge & Jennings, 1999) have been widely studied in AI and Computer Science, for example, about how these mechanisms can be formalized, implemented, and used to enhance cooperation in human-computer interactions and online market systems (Tzeng, 2004; Utz, Matzat, & Snijders, 2009), as well as general multi-agent systems (Winikoff, 2007; Wooldridge & Jennings, 1999), our study would provide important insights for the design and deployment of such mechanisms; for instance, what kind of apology should be provided to customers when making mistakes, and whether apology can be enhanced when complemented with commitments to ensure better cooperation, e.g. compensation from customer's for wrongdoing.

Commitments in Public Goods

Whenever creating a public good, strategies or mechanisms are required to handle defectors. Arranging a prior commitment or agreement is an essential ingredient to encourage cooperative behavior in a wide range of relationships, ranging from personal to political and religious ones. Prior agreements clarify the intentions and preferences of other players. Hence, refusing to establish an agreement may be considered as intending or preferring not to cooperate (non-committers). Prior agreements may be highly rewarding in group situations, as in the case of Public Goods Games (Ostrom, 1990), as it forces the other participants to signal their willingness to achieve a common goal. Especially for increasing group sizes, such prior agreements could be ultimately rewarding, as it becomes more and more difficult to assess the aspirations of all participants.

We have shown (Han, Pereira, & Lenaerts, 2014), mathematically and numerically, that prior agreements with posterior compensations provide a strategic solution that leads to substantial levels of cooperation in the context of Public Goods Games, results that are corroborated by available experimental data.

Notwithstanding this success, one cannot, as with other approaches, fully exclude the presence of defectors, raising the question of how they can be dealt with to avoid the demise of the common good. We showed that avoiding creation of the common good (whenever full agreement is not reached), or limiting the benefit that disagreeing defectors can acquire (using costly restriction mechanisms), are both relevant choices.

Nonetheless, restriction mechanisms are found to be the more favorable, especially in larger group interactions. Given decreasing restriction costs, then introducing restraining measures to cope with public goods free-riding issues is the ultimate advantageous solution for all involved participants, rather than avoiding its creation.

Collective Realm Conclusion

We have argued that the study of the aforementioned issues has come of age and is ripe with research opportunities, having communicated some of the inroads we explored, and pointed to the more detailed published results of what we have achieved, with respect to intention recognition, commitment, and mutual tolerance through apology, within the overarching Evolutionary Game Theory context.

BRIDGING THE TWO REALMS OF MORALITY FOR MACHINES

We have examined above two types of incursions, one into the individual's success in a fixed group, and the second into the evolving population realms of morality.

The first type resorts to individual cognition and reasoning to enable such individuals to successfully compete amongst free riders and deceivers. Such successful competition can be achieved by learning past interactions with them or by recognizing their intentions (Pereira & Han, 2011). The second type emphasizes instead the emergence, in a population, of evolutionarily stable moral norms, of fair and just cooperation, that ably discard free riders and deceivers, to the advantage of the whole evolved population.

To this latter end, some cognitive abilities such as intention recognition, commitment, and apology were employed, singly or jointly, by instilling them into just some individual agents, which then become predominant and lastly invade the evolving population, whether in the context of pairwise interactions or of public good situations.

A fundamental question then arises, concerning the study of individual cognition in groups of often morally interacting multi-agents (that can choose to defect or cooperate with others), whether from such 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. Some initial Evolutionary Game Theory results into certain learning methods have identified a broad class of situations where this is the case (Segbroeck, Jong, Nowé, Santos, & Lenaerts, 2010; Pinheiro, Pacheco, & Santos, 2012; Börgers & Sarin, 1997). A premium outstanding issue remains in regard to which cognitive abilities and circumstances the result may obtain in general, and for sure that will be the object of much new and forthcoming programs of research.

Specifically with respect to human morality, the answer to the above-enounced fundamental question would appear to be a resounding 'Yes'. For one, morality concerns both groups and populations, requires cognition, and will have had to evolve in a nature/nurture or gene/culture intertwining and reinforcement. For another, evolutionary anthropology, psychology, and neurology have been producing ever more consilient views on the evolution of human morality.

Their scientific theories and results must per force be kept in mind, and serve as inspiration, when thinking and rethinking about machine ethics. And all the more so because the machines will need to be ethical amongst us human beings, not just among themselves.

On the other hand, the very study of ethics, and the evolution of human morality too, can now avail themselves of the experimental, computation theoretic, and robotic means to enact and simulate individual or group moral reasoning, in a plethora of circumstances. Likewise for the emergence of moral rules and behaviors in evolving populations.

Hence, having already addressed above two computational types of models, in the next section below we stress this double outlook, by bringing to the fore congenial present views and research on the evolution of human morality, hoping to reinforce the bridging ideas and paradigm 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.

What emerges as morality? The answer is not some “thing” but rather something like a form, or pattern, or function. The concept of emergence applies to phenomena in which relational properties

dominate over constituent properties in determining aggregate features. It is with respect to configurations and topologies, not specific properties of constituents, that we trace processes of emergence.

We depart then from the point of view where morality is established as that property or ability to act solely according to reasons and motives that are taken as one's own; however, but these could otherwise be redefined to capture the mutual influences among individuals and the population.

By analogy with computing machines, cognitive scientists have argued that the “functional” properties that define a given cognitive operation are like the logical architecture of a computer program. Philosophically, this general form of argument is known as 'functionalism', and it is quite relevant for viewing morality as an emergent property. In that respect, we adopt the standpoint of functionalism. As we put it in another context, the point “is that the brain, in its biological evolution, evolved so that it could execute any kind of mind software: personhood, art, whatever; that the brain has bootstrapped itself into generality (Pereira, 2012b, 2014).

THE EVOLUTIONARY TEACHINGS

Added dependency on cooperation makes it more competitive to cooperate well. Thus, it is advantageous to invest on shared morals in order to attract partners who will partake of mutual and balanced advantages.

This evolutionary hypothesis inspired by mutualism (Baumard, 2010)—itself a form of contractualism (Ashford & Mulgan, 2007)—contrasts with a number of naturalist theories of morality, which make short shrift of the importance of cognition for cooperation. For example, the theory of reciprocity, in ignoring a wider cognitive capacity to choose and attract one's partners, forbids itself from explaining evolution on the basis of a cooperation market.

Indeed, when assigning all importance to population evolutionary mechanisms, naturalist theories tend to forget the evolution of cognition in individuals. Such theories habitually start off from evolutionary mechanisms for understanding the specificity of human morals: punishment (Boyd & Richerson, 1992 ; Sober & Wilson, 1998), culture (Henrich & Boyd, 2001 ; Sober & Wilson, 1998), political alliances (Boehm, 1999 ; Erdal, Whiten, Boehm, & Knauft, 1994). According to Baumard's hypothesis, morality does not emerge because humans avail themselves of new means for punishing free-riders or for recompensing cooperators, but simply because mutual help—and hence the need to find partners—becomes much more important.

In summary, it's the development of cooperation that induces the emergence of morals, and not the stabilization of morals (via punishment or culture) that promotes the development of cooperation.

Experimental results are in line with the hypothesis that the perfecting of human intuitive psychology is responsible for the emergence of morality, on the basis of an improved understanding of the mental states of others. This permits to communicate, not just to coordinate with them, and thus extend the domain cooperation, thereby leading to a disposition toward moral behaviors. For a systematic and thorough account of research into the evolutionary origins of morality, see Krebs (2011) and Bowles and Gintis (2011).

At the end of the day, one may consider three theories bearing on three different aspects of morality: the evaluation of interests for utilitarianism, the proper balance of interests for mutualism, and the discharging of obligations for the virtues principled.

A naturalistic approach to moral sense does not make the psychological level disappear to the benefit of the evolutionary one. To each its explanation level: psychology accounts for the workings of the moral sense; sociology, for the social context that activates it; and a cupola theory, for the evolution of

causes that occasioned it (Sperber, 1977). Moral capability is therefore a "mechanism" amongst others (Elster, 1998), as are the concern for reputation, the weakness of the will, the power to reason, etc.

An approach that is at once naturalist and mutualist allows escape from these apparently opposite viewpoints: the psychological and the societal. At the level of psychological motivations, moral behavior does neither stem from egotism nor altruism. To the contrary, it aims at the mutual respect for everyone's attending interests. And, simultaneously, it obeys the logic of equity. At the evolutionary level, moral behavior is not contradictory with egotism because, in human society, it is often in our own interest to respect the interests of others. Through moral motivations, we avail ourselves of a means to reconcile the diverse individual interests. Morality vies precisely at harmonizing individual interest with the need to associate, and profit from cooperation, by adopting a logic of fairness.

The mutualist solution is not new. Contractualist philosophers have upheld it for some time. Notably, they have furnished detailed descriptions of our moral capacity (Rawls, 1971; Thomson, 1971). However, they never were able to explain why humans are enabled with that particular capacity: Why do our judgments seek equity? Why do we behave morally at all?

Without an explanation, the mutualist theory seems improbable: Why behave we as if an actual contract had been committed to, when in all evidence one was not?

Past and ongoing evolutionary studies, intertwining and bridging cognitive and population aspects, and both becoming supported on computational simulations, will help us find answers to that. In the process, rethinking machine ethics and its implementations.

According to Boehm (2012), conscience and morality evolved, in the biological sense. Conscience evolved for reasons having to do with environments humans had to cope with prehistorically, and their growing ability to use group punishment to better their social and subsistence lives and create more equalized societies. His general evolutionary hypothesis is that morality began with having a conscience and that conscience evolution began with systematic but initially non-moralistic social control by groups.

This entailed punishment of individual “deviants” by bands of well-armed large-game hunters, and, like the ensuing preaching in favor of generosity, such punishment amounted to “social selection”, since the social preferences of members and of groups as a whole had systematic effects on gene pools.

This punitive side of social selection adumbrates an immediate kind of “purpose”, of large-brained humans actively and insightfully seeking positive social goals or avoiding social disasters arising out of conflict. No surprise the genetic consequences, even if unintended, move towards fewer tendencies for social predation and more towards social cooperation. Hence, group punishment can improve the quality of social life, and over the generations gradually shape the genotype in a similar direction.

Boehm's idea is that prehistoric humans made use of social control intensively, so that individuals who were better at inhibiting their own antisocial tendencies, by fear of punishment or by absorbing and identifying with group’s rules, garnered a superior fitness. In learning to internalize rules, humankind acquired a conscience. At the beginning this stemmed from punitive social selection, having also the strong effect of suppressing free riders. A newly moralistic type of free-rider suppression helped evolve a remarkable capacity for extra-familial social generosity. That conscience gave us a primitive sense of right and wrong, which evolved the remarkable “empathy” which we are infused with today. It is a conscience that seems to be as much a Machiavellian risk calculator as a moral force that maximizes prosocial behavior, with others' interests and equity in mind, and minimizes deviance too. It is clear that “biology” and “culture” work together to render us adaptively moral.

Boehm believes the issue of selfish free riders requires further critical thought, and that selfish intimidators are a seriously neglected type of free rider. There has been too much of a single-minded focus on cheating dominating free rider theorizing. In fact, he ascertains us the more potent free riders

have been alpha-type bullies, who simply take what they want. It is here his work on the evolution of hunter-gatherer egalitarianism enters, namely with its emphasis on the active and potentially quite violent policing of alpha-male social predators by their own band-level communities. Though there's a large literature on cheaters and their detection, free-rider suppression in regard to bullies has not been taken into account so far in the mathematical models that study altruism.

"For moral evolution to have been set in motion," Boehm (2012) goes on, "more was needed than a preexisting capacity for cultural transmission. It would have helped if there were already in place a good capacity to strategize about social behavior and to calculate how to act appropriately in social situations."

In humans, the individual understanding that there exists a self in relation to others makes possible participation in moral communities. Mere self-recognition is not sufficient for a moral being with fully developed conscience, but a sense of self is a necessary first step useful in gauging the reactions of others to one's behavior and to understand their intentions. And it is especially important to realize that one can become the center of attention of a hostile group, if one's actions offend seriously its moral sensibilities. The capacity to take on the perspective of others underlies not just the ability of individuals in communities to modify their behavior and follow group imposed rules, but it also permits people acting as groups to predict and cope insightfully with the behavior of "deviants."

Social selection reduced innate dispositions to bully or cheat, and kept our conscience in place by self-inhibiting antisocial behavior. A conscience delivers us a social mirror image. A substandard conscience may generate a substandard reputation and active punishment too. A conscience supplies not just inhibitions, but serves as an early warning system that helps prudent individuals from being sanctioned.

Boehm (2012) wraps up: "When we bring in the conscience as a highly sophisticated means of channeling behavioral tendencies so that they are expressed efficiently in terms of fitness, scenarios change radically. From within the human psyche an evolutionary conscience provided the needed self-restraint, while externally it was group sanctioning that largely took care of the dominators and cheaters. Over time, human individuals with strong free-riding tendencies—but who exercised really efficient self-control—would not have lost fitness because these predatory tendencies were so well inhibited. And if they expressed their aggression in socially acceptable ways, this in fact would have aided their fitness. That is why both free-riding genes and altruistic genes could have remained well represented and coexisting in the same gene pool."

For sure, we conclude, evolutionary biology and anthropology, like the cognitive sciences too (Hauser, 2006; Gazzaniga, 2006; Churchland, 2011; Greene, 2013; Tomasello, 2014), have much to offer in view of rethinking machine ethics, evolutionary game theory simulations of computational morality to the rescue.

CODA

In realm of the individual, computation is vehicle for the study and teaching of morality, namely in its modeling of the dynamics of knowledge and cognition of agents. In the collective realm, norms and moral emergence have been studied computationally in populations of rather simple-minded agents. By bridging these realms, cognition affords improved emerged morals in populations of situated agents.

At the end of the day, we will certainly wish ethical machines to be convivial with us.

ACKNOWLEDGMENTS

We thank the co-authors of joint papers, The Anh Han, Francisco C. Santos, and Tom Lenaerts, for use of material from diverse joint publications referenced below.

Ari Saptawijaya acknowledges the support of Fundação para a Ciência e a Tecnologia (FCT/MEC) Portugal, grant SFRH/BD/72795/2010.

REFERENCES

Abeler, J., Calaki, J., Andree, K., & Basek, C. (2010). The power of apology. *Economics Letters, 107(2)*, 233-235.

Alferes, J. J., Pereira, L. M., & Swift, T. (2004). Abduction in well-founded semantics and generalized stable models via tabled dual programs. *Theory and Practice of Logic Programming*, *4(4)*, 383–428.

Anderson, M., & Anderson S. L. (2008). EthEl: Toward a principled ethical eldercare robot. In *AAAI Fall Symposium Technical Report on AI in Eldercare*. Palo Alto, CA: AAAI Press.

Anderson, M., & Anderson S. L. (2010). Robot be good: A call for ethical autonomous machines. *Scientific American*, *303(4)*, 54-59.

Anderson, M., Anderson S. L., & Armen, C. (2005). Towards machine ethics: Implementing two action-based ethical theories. In *AAAI Fall Symposium Technical Report on Machine Ethics*. Palo Alto, CA: AAAI Press.

Anderson, M., Anderson S. L., & Armen, C. (2006). MedEthEx: a prototype medical ethics advisor. In *Proceedings of the Eighteenth Conference on Innovative Applications of Artifical Intelligence (IAAI'06)*. Palo Alto, CA: AAAI Press.

Ashford, E., & Mulgan, T. (2007). Contractualism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition). Retrieved from http://plato.stanford.edu/entries/contractualism/

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review, 80(4)*, 1095–1111.

Baral, C., Gelfond, M., & Rushton, N. (2009). Probabilistic reasoning with answer sets. *Theory and Practice of Logic Programming, 9(1)*, 57–144.

Baumard, N. (2010). *Comment nous sommes devenus moraux: Une histoire naturelle du bien et du mal.* Paris: Odile Jacob.

Boehm, C. (1999). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

Boehm, C. (2012). *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. New York: Oxford University Press.

Börgers, T., & Sarin, R. (1997). Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory, 77(1)*, 1-14.

Bowles, S., & Gintis, H. (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press.

Boyd, R., & Richerson, P. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13(3)*, 171-195.

Bratman, M. E. (1987). *Intention, Plans and Practical Reasoning.* Cambridge, MA: Harvard University Press.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems, 21(4)*, 38-44.

Bringsjord, S., Taylor, J., van Heuveln, B., Arkoudas, K., Clark, M., & Wojtowicz, R. (2011). Piagetian roboethics via category theory: Moving beyond mere formal operations to engineer robots whose decisions are guaranteed to be ethically correct. In M. Anderson and S. L. Anderson (Eds.), *Machine Ethics* (pp. 361-374). New York, NY: Cambridge University Press.

Charniak, E., & Goldman, R. P. (1993). A Bayesian model of plan recognition. *Artificial Intelligence, 64(1)*, 53–79.

Churchland, P. (2011). *Braintrust: What Neuroscience Tells Us about Morality*. Princeton: Princeton University Press.

Cohen, P. R., & Levesque, H. J. 1990. Intention is Choice with Commitment. *Artificial Intelligence, 42(2-3)*, 213–261.

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. M. Doris (Ed.), *The Moral Psychology Handbook*. New York: Oxford University Press.

Dell'Acqua, P., & Pereira, L. M. (2007). Preferential theory revision. *Journal of Applied Logic, 5(4)*, 586–601.

Elster, J. (1998). A plea for mechanisms. In P. Hedström & R. Swedberg (Eds.), *Social Mechanisms: An analytical approach to social theory* (pp. 45-73). Cambridge, NY: Cambridge University Press.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review, 12(2)*, 168–192.

Erdal, D., Whiten, A., Boehm, C., & Knauft, B. (1994). On human egalitarianism: An evolutionary product of machiavellian status escalation? *Current Anthropology, 35(2)*, 175-183.

Fischbacher, U., & Utikal, V. (2013). On the acceptance of apologies. *Games and Economic Behavior, 82*, 592 – 608.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.

Frank, R. H. (2001). Cooperation through Emotional Commitment. In R. M. Nesse (Ed.), *Evolution and the capacity for commitment* (pp. 55-76). New York: Russell Sage.

Ganascia, J.-G. (2007). Modelling ethical rules of lying with Answer Set Programming. *Ethics and Information Technology, 9(1)*, 39-47.

Ganascia, J.-G. (2012). *An Agent-Based Formalization for Resolving Ethical Conflicts.* In Proceedings of the Workshop on Belief Change, Non-monotonic Reasoning, and Conflict Resolution (BNC@ECAI'12), Montpellier, France.

Gazzaniga, M. S. (2006). *The Ethical Brain: The Science of Our Moral Dilemmas.* New York: Harper Perennial.

Gelfond, M., & Lifschitz, V. (1988). The stable model semantics for logic programming. In *Proceedings of the Fifth International Conference on Logic Programming (ICLP)* (pp. 1070-1080). Cambridge, MA: MIT Press.

Gintis, H. (2001). Beyond selfishness in modeling human behavior. In R. M. Nesse (Ed.), *Evolution and the capacity for commitment*. New York: Russell Sage.

Greene, J. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them***.** New York: The Penguin Press HC.

Groves, T. 1973. Incentives in Teams. *Econometrica, 41(4)*, 617–31.

Guarini, M. (2011). Computational neural modeling and the philosophy of ethics: Reflections on the particularism-generalism debate. In M. Anderson and S. L. Anderson (Eds.), *Machine Ethics* (pp. 316-334). New York, NY: Cambridge University Press.

Han, T. A. (2013). Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models. *SAPERE series*, 9. Berlin: Springer-Verlag.

Han, T. A., & Pereira, L. M. (2011). Context-dependent incremental intention recognition through Bayesian network model construction. In A. Nicholson (Ed.), *Proceedings of the Eighth UAI Bayesian Modeling Applications Workshop (CEUR Workshop Proceedings)* (Vol. 818, pp. 50–58). Retrieved from http://ceur-ws.org/Vol-818/paper7.pdf

Han, T. A., Pereira, L. M., & Lenaerts, T. (2014). Emergence of Commitments in Public Goods Game: Restricting vs. Avoiding Non-Committers (Submitted). Available from http://centria.di.fct.unl.pt/~lmp/publications/online-papers/commitment_restriction.pdf

Han, T. A., Pereira, L. M., & Santos, F. C. (2011a). Intention recognition promotes the emergence of cooperation. *Adaptive Behavior, 19(3)*, 264–279.

Han, T. A., Pereira, L. M., & Santos, F. C. (2011b). The role of intention recognition in the evolution of cooperative behavior. In T. Walsh (Ed.), *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (pp. 1684–1689). AAAI Press.

Han, T. A., Pereira, L. M., & Santos, F. C. (2012a). Corpus-based intention recognition in cooperation dilemmas. *Artificial Life, 18(4)*, 365–383.

Han, T. A., Pereira, L. M., & Santos, F. C. (2012b). The emergence of commitments and cooperation. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems* (pp. 559-566). International Foundation for Autonomous Agents and Multiagent Systems.

Han, T. A., Pereira, L. M., & Santos, F. C. (2012c). Intention Recognition, Commitment, and The Evolution of Cooperation. In *Proceedings of IEEE Congress on Evolutionary Computation* (pp. 1–8). IEEE Press.

Han, T. A., Pereira, L. M., Santos, F. C., & Lenaerts, T. (2013a). Good agreements make good friends. *Scientific Reports,* 3. doi: 10.1038/srep02695.

Han, T. A., Pereira, L. M., Santos, F. C., & Lenaerts, T. (2013b). Why is it so hard to say sorry: The evolution of apology with commitments in the iterated Prisoner's Dilemma. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 177–183). Palo Alto: AAAI Press.

Han, T. A., Pereira, L. M., Santos, F. C., & Lenaerts, T. (in press). Emergence of Cooperation via Intention Recognition, Commitment, and Apology -- A Research Summary. AI Communications.

Han, T. A., Saptawijaya, A., & Pereira, L. M. (2012). Moral reasoning under uncertainty. In N. Bjørner, & A. Voronkov (Eds.), *Proceedings of the Eighteenth International Conference on Logic for Programming Artificial Intelligence and Reasoning (LNCS)* (Vol. 7180, pp. 212-227). Berlin:Springer-Verlag.

Hardin, G. (1968). The tragedy of the commons. *Science, 162(3859)*, 1243–1248.

Hauser, M. D. (2006). *Moral Minds: The Nature of Right and Wrong.* New York: Harper Perennial.

Hauser, M., Cushman, F., Young, L., Jin, R. K., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind and Language, 22(1)*, 1–21.

Heinze, C. (2003). *Modeling Intention Recognition for Intelligent Agent Systems* (Doctoral Dissertation). The University of Melbourne, Australia.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208(1)*, 79-89.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. New York, NY: Cambridge University Press.

Inhelder, B., & Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books.

Jackson, M. O. (2000). Mechanism theory. In U. Derigs (Ed.), *Optimization and Operations Research.* Paris: EOLSS Publishers.

Jonsen, A. R., & Toulmin, S. (1988). *The Abuse of Casuistry: A History of Moral Reasoning*. Oakland, CA: University of California Press.

Kamm, F. M. (2006). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm.* New York, NY: Oxford University Press.

Kowalski, R. (2011). *Computational Logic and Human Thinking: How to be Artificially Intelligent.* New York, NY: Cambridge University Press.

Krebs, D. L. (2011). *The Origins of Morality: An Evolutionary Account*. New York: Oxford University Press.

Lesh, N. (1998). *Scalable and Adaptive Goal Recognition* (Doctoral Dissertation). University of Washington.

Lopes, G., & Pereira, L. M. (2006). Prospective programming with ACORDA. In *Proceedings of the FLoC'06 Workshop on Empirically Successful Computerized Reasoning (ESCoR'06)*, Seattle, USA.

Lopes, G., & Pereira, L. M. (2010). Prospective storytelling agents. In M. Carro, & R. Peña (Eds.), *Proceedings of the Twelfth International Symposium on Practical Aspects of Declarative Languages (LNCS)* (Vol. 5937, pp. 294-296). Berlin: Springer-Verlag.

Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1993). The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology, 29*, 87–109.

McAfee, R. P. (1993). Mechanism Design by Competing Sellers. *Econometrica, 61(6)*, 1281–1312.

McCloy, R., & Byrne, R. M. J. (2000). Counterfactual thinking about controllable events. *Memory and Cognition, 28*, 1071–1078.

McIntyre, A. (2004). Doctrine of double effect. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2011 edition). Retrieved from http://plato.stanford.edu/entries/double-effect/

McLaren, B. M. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems, 21(4)*, 29–37.

Meltzoff, A. N. (2005). Imitation and other minds: the "like me" hypothesis. In *Perspectives On Imitation: From Neuroscience to Social Science. Imitation, Human Development, and Culture* (pp. 55-77). Cambridge, MA: MIT Press.

Migliore, S., Curcio, G., Mancini, F., & Cappa, S. F. (2014). Counterfactual thinking in moral judgment: an experimental study. *Frontiers in Psychology, 5*, 451.

Murakami, Y. (2004). Utilitarian Deontic Logic. In *Proceedings of the Fifth International Conference on Advances in Modal Logic (AiML'04)*. London: King's College Publications.

Myerson, R. (1979). Incentive compatibility and the bargaining problem. *Econometrica, 47(1)*, 61–73.

Naor, M., Pinkas, B., & Sumner, R. (1999). Privacy preserving auctions and mechanism design. In *Proceedings of the 1st ACM Conference on Electronic Commerce* (pp. 129–139). ACM.

Nesse, R. M. (2001a). Natural selection and the capacity for subjective commitment. In R. M. Nesse (Ed.), *Evolution and the Capacity for Commitment* (pp. 1-44). New York: Russell Sage.

Nesse, R. M. (2001b). *Evolution and the Capacity for Commitment*. New York: Russell Sage.

Nisan, N., & Ronen, A. (1999). Algorithmic mechanism design. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing* (pp. 129–140). ACM.

Nowak, M. A. (2006a). *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge, MA: Harvard University Press.

Nowak, M. A. (2006b). Five rules for the evolution of cooperation. *Science, 314(5805)*, 1560-1563. doi: 10.1126/science.1133755.

Nowak, M. A., and Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature, 355*, 250–253.

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior, 30(2)*, 114–123.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge, MA: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. New York, NY: Cambridge University Press.

Pereira, L. M. (2012a). Evolutionary Tolerance. In L. Magnani, & L. Ping (Eds.), *Philosophy and Cognitive Science—Western & Eastern Studies (SAPERE)* (Vol. 2, pp. 263-287). Berlin: Springer-Verlag.

Pereira, L. M. (2012b). Turing is Among Us. *Journal of Logic and Computation, 22(6)*, 1257-1277.

Pereira, L. M. (2014). Can we not Copy the Human Brain in the Computer? In *"Brain.org"* (pp. 118-126). Lisbon: Fundação Calouste Gulbenkian.

Pereira, L. M., Dell'Acqua, P., Pinto, A. M., & Lopes, G. (2013). 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.

Pereira, L. M., & Han, T. A. (2009a). Evolution Prospection. In K. Nakamatsu, G. Phillips-Wren, L. C. Jain, & R. J. Howlett (Eds.), *Proceedings of the First KES International Symposium IDT (New Advances in Intelligent Decision Technologies)* (Vol. 199, pp. 51-63). Berlin: Springer-Verlag.

Pereira, L. M., & Han, T. A. (2009b). Intention recognition via causal Bayes networks plus plan generation. In *Proceedings of 14th Portuguese International Conference on Artificial Intelligence (LNCS)* (Vol. 5816, pp. 138–149). Berlin: Springer-Verlag.

Pereira, L. M., & Han, T. A. (2011). Intention recognition with evolution prospection and causal Bayesian networks. In A. Madureira, J. Ferreira, & Z. Vale (Eds.), *Computational Intelligence for Engineering Systems: Emergent Applications* (pp. 1-33). Berlin: Springer-Verlag.

Pereira, L. M., & Saptawijaya, A. (2007a). Moral Decision Making with ACORDA. In *Local Proceedings of the Fourteenth International Conference on Logic for Programming Artificial Intelligence and Reasoning (LPAR'07)*, Yerevan, Armenia.

Pereira, L. M., & Saptawijaya, A. (2007b). Modelling Morality with Prospective Logic. In J. M. Neves, M. F. Santos, & J. M. Machado (Eds.), *Proceedings of the Thirteenth Portuguese Conference on Artificial Intelligence (LNCS)* (Vol. 4874, pp. 99-111). Berlin: Springer-Verlag.

Pereira, L. M., & Saptawijaya, A. (2009). Modelling Morality with Prospective Logic. *International Journal of Reasoning-based Intelligent Systems, 1(3/4)*, 209–221.

Pereira, L. M., & Saptawijaya, A. (2011). Modelling Morality with Prospective Logic. In M. Anderson and S. L. Anderson (Eds.), *Machine Ethics* (pp. 398-421). New York, NY: Cambridge University Press.

Pereira, L. M., & Saptawijaya, A. (2014). Counterfactuals in Logic Programming with Applications to Agent Morality (Submitted). Available from http://centria.di.fct.unl.pt/~lmp/publications/online-papers/moral_counterfactuals.pdf

Phelps, S., McBurney, P., & Parsons, S. (2010). Evolutionary mechanism design: a review. *Autonomous Agents and Multi-Agent Systems, 21(2)*, 237–264.

Pinheiro, F. L., Pacheco, J. M., & Santos, F. C. (2012). From Local to Global Dilemmas in Social Networks. *PLoS ONE, 7(2)*: e32114. doi:10.1371/journal.pone.0032114.

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems*, *21(4)*, 46–51.

Rahwan, I., & Simari, G. (Eds.). (2009). *Argumentation in Artificial Intelligence.* Berlin: Springer-Verlag.

Rand, D. G., Fudenberg, D., & Dreber, A. (2013). It's the thought that counts: The role of intentions in noisy repeated games. *Social Science Research Network.* Retrieved September 22, 2014, from http://ssrn.com/abstract=2259407

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.

Ross, D. (2005). *Economic theory and cognitive science: Microexplanation*. Cambrdge, MA: MIT press.

Ross, W. D. (1930). *The Right and the Good*. New York: Oxford University Press.

Roy, P., Bouchard, B., Bouzouane, A., & Giroux, S. (2007). A hybrid plan recognition model for alzheimer's patients: interleaved-erroneous dilemma. In *Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology* (pp. 131–137).

Roy, O. (2009). *Thinking before Acting: Intentions, Logic, Rational Choice* (Doctoral Dissertation). ILLC Dissertation Series DS-2008-03, Amsterdam.

Sadri, F. (2011). Logic-based approaches to intention recognition. In N.-Y. Chong, & F. Mastrogiovanni (Eds.), *Handbook of Research on Ambient Intelligence: Trends and Perspectives* (pp. 346–375). Hershey, PA: IGI Global.

Saptawijaya, A., & Pereira, L. M. (2013a). Tabled abduction in logic programs (Technical Communication of ICLP 2013). *Theory and Practice of Logic Programming, Online Supplement, 13(4-5)*. Retrieved from http://journals.cambridge.org/downloadsup.php?file=/tlp2013008.pdf

Saptawijaya, A., & Pereira, L. M. (2013b). Incremental tabling for query-driven propagation of logic program updates. In K. McMillan, A. Middeldorp, & A. Voronkov (Eds.), *Proceedings of the Nineteenth International Conference on Logic for Programming Artificial Intelligence and Reasoning (LNCS)* (Vol. 8312, pp. 694-709). Berlin:Springer-Verlag.

Saptawijaya, A., & Pereira, L. M. (2014). Joint tabling of logic program abductions and updates (Technical Communication of ICLP 2014). *Theory and Practice of Logic Programming, Online Supplement, 14(4-5)*. Retrieved from http://arxiv.org/abs/1405.2058

Saptawijaya, A., & Pereira, L. M. (in press). The Potential of Logic Programming as a Computational Tool to Model Morality. In R. Trappl (Ed.), *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations* (Cognitive Technologies). Berlin: Springer-Verlag.

Scanlon, T. M. (1982). Contractualism and utilitarianism. In A. Sen, & B. Williams (Eds.), *Utilitarianism and Beyond*. New York, NY: Cambridge University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.

Searle, J. R. (1995). *The Construction of Social Reality*. New York: The Free Press.

Searle, J. R. (2010). *Making the Social World: The Structure of Human Civilization*. New York: Oxford University Press.

Segbroeck, S. V., Jong, S. D., Nowé, A., Santos, F. C., & Lenaerts, T. (2010). Learning to coordinate in complex networks. *Adaptive Behavior, 18(5)*, 416–427.

Sigmund, K. (2010). *The Calculus of Selfishness*. Princeton, NJ: Princeton University Press.

Sober, E., & Wilson, D. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Sperber, D. (1997). Individualisme méthodologique et cognitivisme. In R. Boudon, F. Chazel, & A. Bouvier (Eds.), *Cognition et sciences sociales* (pp. 123-136). Paris: Presses Universitaires de France.

Swift, T., & Warren, D. S. (2012). XSB: Extending Prolog with tabled logic programming. *Theory and Practice of Logic Programming, 12(1-2)*, 157–187.

Thomson, J. J. (1971). A defense of abortion. *Philosophy & Public Affairs, 1(1)*, 47-66.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 279*, 1395–1415.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.

Tzeng, J.-Y. (2004). Toward a more civilized design: studying the effects of computers that apologize. *International Journal of Human-Computer Studies, 61(3)*, 319 – 345.

Utz, S., Matzat, U., & Snijders, C. (2009). On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions. *International Journal of Electronic Commerce, 13(3)*, 95–118.

van den Hoven, J., & Lokhorst, G-J. (2002). Deontic logic and computer-supported computer ethics. *Metaphilosophy, 33(3)*, 376–386.

van Gelder, A., Ross, K. A., & Schlipf, J. S. (1991). The well-founded semantics for general logic programs. *Journal of ACM, 38(3)*, 620–650.

Wiegel, V. (2007). *SophoLab; Experimental Computational Philosophy* (Doctoral dissertation). Delft University of Technology, The Netherlands.

Winikoff, M. (2007). Implementing commitment-based interactions. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 868–875).

Wooldridge, M., & Jennings, N. R. (1999). The cooperative problem-solving process. *Journal of Logic and Computation, 9(4)*, 563-592.

Key Terms and Definitions

Abduction: A reasoning method whereby one chooses from available hypotheses those that best explained the observed evidence, in a preferred sense.

Computational Logic: An interdisciplinary field of enquiry that employs the techniques from symbolic logic to reason using practical computations, and typically achieved by means of computer supported automated tools.

Contractualism: A school of thought about morality, emphasizing explicit reasoning (rather than merely relying on subjective observation) for providing moral justifications to others, through looking for common ground that others could not reasonably reject to.

Counterfactual: A concept that captures the process of reasoning about a past event that did not occur, namely what would/could/might have happened, had this alternative event occurred; or, conversely, to reason about a past event that did occur, but what if it had not.

Doctrine of Double Effect: A moral principle that explains the permissibility of an action by distinguishing whether its harm consequence is merely a *side-effect*, rather than a *means* to bring about a good result.

Doctrine of Triple Effect: A moral principle that refines the Doctrine of Double Effect, particularly on the notion about harming someone as an intended means, by distinguishing further between doing an action *in order* that an effect occurs and doing it just *because* that effect will occur.

Dual-Process Model: A model that explains how a moral judgment is driven by an interaction of two different psychological processes, namely the controlled process (whereby explicit moral principles are consciously applied via deliberative reasoning), and the automatic process (whereby moral judgments are intuition-based and mostly low-level, not entirely accessible to conscious reflection).

Evolutionary Game Theory: An application of game theory to systematically study the evolution of populations, typically by resorting to simulation techniques under a variety of conditions, parameters, and strategies.

Logic Programming: A programming paradigm based on formal logic that permits a declarative representation of a problem and reasoning about this representation, that reasoning being is driven by a specific semantics.