Ethical System Formalization using Non-Monotonic Logics

Jean-Gabriel Ganascia (Jean-Gabriel.Ganascia@lip6.fr)
LIP6, University Paris VI, 104 avenue du Président Kennedy
75016, Paris, FRANCE

Abstract

Ethics is the science of duty, i.e. the science that elucidates the rules of the right behavior. Nevertheless, it seems that the way we rule our lives is intuitive and based on common sense. For instance, it is common to say that ethical rules are default rules, which means that they tolerate exceptions. Some authors argue that moral can only be grounded on particular cases while others defend the existence of general principles related to ethical rules. Our purpose here is not to justify the first or the second position, but to try to model ethical systems using artificial intelligence formalisms. More precisely, this is an attempt to show that progress in non-monotonic logics, which simulate common sense reasoning, provides a way to formalize different ethical conceptions. From a technical point of view, the model developed here makes use of the Answer Set Programming (ASP) formalism. It is applied to compare different ethical systems with respect to their attitude towards lying and could help to extend classical philosophy and to define general conditions required by any ethical system.
Keywords: Answer Set Programming (ASP); Common Sense Reasoning; Computational Ethics; Machine Ethics; Intelligent Agents; Non Monotonic Logics

Introduction

Ethics is the science of human duty (Cf. (Webster, 1913)). As a science, it has to elucidate the body of rules on which we have to determine our behavior. In this respect, an ethical system can be viewed as a decision-making procedure based on statements on which almost all of us agree. However, in the philosophical tradition, the origin and nature of these rules have always been considered to be controversial. For instance, some authors think that ethical rules are default rules (Väyrynen, 2004), which means that they tolerate exceptions, while others disagree: some have argued that morals can only be based on singular cases (Harman, 2005) while others have defended the existence of general principles (Kant, 1997); some judge an action in terms of its consequences, others in terms of the law, etc. Many of these debates concern the opposition between those who think that principles are many in numbers and can be contradictory, since they are derived from experience, while others say that morals have to be based on general rules, which are valid everywhere and all the time.

To be more precise, one of the arguments in favor of the first position, i.e. "moral particularism", is that ethics has to refer to each particular situation and cannot be based on general principles. Imagine, for instance, that you were living in occupied France during the Second World War and that you hid a friend who was wanted by the French militia or the Gestapo, in your home. If you were asked where your friend was, would you obey the general rule that commands you to tell the truth, and denounce the man to the authorities? In the 18^{th} century, there was a discussion between Immanuel Kant (1724–1804) and Benjamin Constant (1767–1830) about this question. Kant's position was that one should always tell the truth (Kant, 1996), even in such a situation, while Constant (Constant, 1988) considered that morals are based on many principles and that, consequently, one should always apply the one that is the most adapted to the situation. The opposition between "moral generalism" and "moral particularism" corresponds to an old opposition between written laws and the cases on which the laws are based. The criticism of moral values based on general statements or laws or rules is that they may be correct in theory, but not applicable to all practical cases. The example above illustrates the difficulty of applying a rigid and general law to particular cases.

Formalization of ethical systems using modern artificial intelligence techniques may be an original way of overcoming the opposition between "moral particularism" and "moral generalism". For instance, in the case of lying, default rules with justified exceptions could be used to satisfy a general rule or principle that prohibits lying, while simultaneously recommending telling a lie in a given particular situation where the truth would violate other rules of duty.

This paper, which is divided in five parts, constitutes an attempt to model three classical ethical systems using the Answer Set Programming formalism (ASP) (Baral, 2003). The first rapidly recalls the ASP semantics and indicates the way ethical systems can be modeled with this formalism. The following three parts consider the formalization of three classical ethical systems, i.e. the Aristotelian, the Kantian and Constant's "Theory of Principles", using the ASP formalism. Each formalization leads to a program expressed in AnsProlog* and is illustrated with an application based on the lying example referred to above. Mainly dedicated to the Constant's "Theory of Principle", the fourth part goes also on to extend classical Kantian ethics by defining, within this framework, the general conditions that are required by any ethical system. The last part opens up future research in computational ethics based on the generalization of the approach described here.

ASP Formalism

The ASP Semantic

In the past, many Artificial Intelligence researchers tried to simulate non-monotonic reasoning, i.e. reasoning based on general rules and accepting exceptions. Several formalisms have been developed, for instance, default logic (Reiter, 1980), circumscription (McCarthy, 1980), non-monotonic logics (McDermott & Doyle, 1980), Truth Maintenance Systems, etc.

However, most of the mechanical solvers based on those formalisms were very inefficient. Recently, a new general

formalism called ASP (Baral, 2003) has been developed to simulate non-monotonic reasoning. It has been designed to unify previous non-monotonic reasoning formalisms. ASP formalism is not only a more recent formalism; it is also more general than others, since it emulates almost all of them and it is fully operational. More precisely, ASP proposes both a clear formalization with a well-defined semantics and efficient operational solvers, which renders automate demonstrations possible.

Within this formalization, it is possible to specify the logical properties of objects with programs Π that are sets of expressions ρ of the following form:
$\rho : L_0 \, or \, L_1 \, or \, \ldots \, or \, L_k \leftarrow L_{k+1}, \ldots, L_m, \textbf{not} \, L_{m+1}, \ldots, L_n.$
where L_i are literals, i.e. atoms or atom negations, and **not** is a logical connective called "negation as failure".

The intuitive meaning of such a rule is that for all Herbrand interpretations that render true all literals in $\{L_{k+1},\ldots, L_m\}$ while not satisfying any literals in $\{L_{m+1},\ldots, L_n\}$ one can derive at least one literal in $\{L_0,\ldots, L_k\}$

Let us remark that ASP formalism contains two negations that need to be distinguished: a classical negation noted "$\neg$" and a negation by failure noted "**not**", which means that a literal cannot be proved in the absence of sufficient information. The non-monotonic properties are mainly due to this "negation as failure" connector.

Being given a program Π, an Answer Set (or a stable model) is a minimal subset of the Herbrand base of Π, which satisfies all rules of Π. Each subset describes a possible world that renders true the rules of Π. Let us note that this intuitive meaning of the programs may be easily formalized, which provides a formal fixpoint semantics of ASP.

Modeling Ethical Systems with ASP

Ethical rules are rules of behavior, i.e. rules that help to decide what to do and what not to do. Therefore, any ethical system, i.e. any consistent set of ethical rules, requires defining a decision-making procedure; this paper claims that these procedures can be described using artificial intelligence techniques. Since a logical description helps to clarify the ideas and to highlight differences between different ethical systems, the aim is not just try to simulate ethical reasoning using classical AI techniques, but to describe these decision-making procedures in a purely declarative way, using modern logic-based AI techniques. Moreover, since ethical reasoning is a kind of common sense reasoning, it justifies the use of non-monotonic logic. Lastly, ASP techniques have been chosen because they seem appropriate for such a model. The existence of solvers makes is easy to validate our models in different situations.

Note that, in the past, there were many attempts to base ethics on empirical principles, i.e. on observations according, for instance, to the observed utility, to common uses or to traditions. Over the last few years, philosophers have used artificial intelligence techniques, and more specifically statistical learning theory (Harman, 2005) or game theory (Braithwaite, 1955), to model these processes using computers and/or well-founded mathematical theories. There is no doubt that such attempts are very fruitful and interesting. However, the goal here is different, since it is not a question of basing morals on simulation, but of understanding the underlying logic on which classical ethical systems rely. In the last few years, there have been some attempts to formalize ethical systems using modal logic formalisms (Gensler, 1996) and to operationalize these formalizations on computer (Bringsjord, Arkoudas, & Bello, 2006; Powers, 2005). However, these formalizations are mainly based on the use of deontic logics (Meyer J.-J. Ch., 1994) that are well adapted to ethical systems focused on laws were permission and prohibitions are well defined, but not to consequentialist ethical systems. Using non-monotonic logics (cf. (Powers, 2006)) or ASP offers a more general approach, since it can describe not only the consequentialist ethical systems but also deontic ones as it can represent modal and deontic logics.

In order to show this, three ethical conceptions have been formalized: the classical Aristotelian one, the Kantian one based on the categorical imperative and Constant's theory that authorizes a great number of principles tolerating exceptions. Each model is illustrated using the dilemma of the lie presented in the introduction.

Aristotelian Rules

A Decision-Making Procedure

According to the traditional Aristotelian ethics (Aristotle, 2002), in each situation we have to look at all possible actions and to choose the best one, i.e. the least unjust. More precisely, our will — i.e. our goal — can be achieved by choosing the appropriate action among the different actions we have at our disposal. In modern terms, Aristotelian ethics can be reduced to a general decision-making procedure based on preferences that characterize the just and the unjust. Using ASP formalism, this can be expressed using the following rules[1]:

$$act(P,G,A) \leftarrow action(A), person(P), goal(P,G), \\ solve\_goal(P,G,A), \textbf{not} \; unjust(A).$$

$$\leftarrow action(P,G,A), action(P,G,AA), A \neq AA.$$

The just and the unjust are defined with the use of two binary predicates, $worse(A,B)$, which means that action A is worse than action B, and $consequence(A,C)$, which means that C is a consequence of A. Briefly speaking, an action A is just if its worst consequences are not worse than those of other actions AA. More formally, it can be characterized using the following ASP rules:

$$just(A) \leftarrow worst\_consequence(A,C), \\ worst\_consequence(AA,CC), worse(CC,C), \textbf{not} \; unjust(A).$$

[1] All these formalizations have been coded in AnsProlog* and tested using the smodels solver downloaded from http://www.tcs.hut.fi/Software/smodels.

$unjust(A) \leftarrow worst\_consequence(A,C),$
$worst\_consequence(AA,CC), worse(C,CC), \textbf{not}\ just(A).$

The worst consequence is easy to define using the following two rules once both the *worse* and the *consequence* predicates have been given:

$not\_worst\_consequence(A,C) \leftarrow$
$consequence(A,C), consequence(A,CC),$
$worse(CC,C), \textbf{not}\ worse(C,CC).$

$worst\_consequence(A,C) \leftarrow consequence(A,C),$
$\textbf{not}\ not\_worst\_consequence(A,C).$

The predicate *consequence* translates physical causality. It is a pre-requirement that ethical agents have at their disposal an adequate knowledge of the world. This means that science and improvement of knowledge contribute to ethics. However, science is not sufficient and a second predicate, *worse*, is also required. This predicate expresses a system of values that depends on the culture, social environment or personal commitment of the agent. The aim here is not to justify such or such system of values, e.g. utilitarian, Epicurean, religious, idealistic, and it is assumed that it has already been specified through the *worse* predicate.

Aristotle and the Lie

This general formalization can be tested on the lie example. Let us first suppose that there are three or more persons, "*I*", *Peter* and *Paul*, each of whom has several possible actions, e.g. to tell the truth, to tell a lie, to murder, to eat, to discuss. Let us now consider a situation similar to the one described above where "*I*" am in a situation where "*I*" have to answer a murderer either by lying or by telling the truth. I know that telling the truth means denouncing a friend, which will lead to his murder. What should I do? The situation may be formalized using the following rules:

$obliged(P) \leftarrow act(P, question(P), A).$
$\leftarrow \textbf{not}\ obliged(\text{“}I\text{”}).$
$consequence(A,A) \leftarrow .$
$consequence(tell(\text{“}I\text{”}, truth), murder) \leftarrow .$

The solution depends on my system of values. Let us now suppose that "*I*" admit that it is bad both to lie and to murder. This can be expressed using the following three rules:

$worse(tell(P,lie),A) \leftarrow \textbf{not}\ better(tell(P,lie),A).$
$worse(murder,A) \leftarrow \textbf{not}\ better(murder,A).$
$better(A,A).$

With such a program, half of all the answer sets contain the decision: $act(\text{“}I\text{”}, question(\text{“}I\text{”}), tell(\text{“}I\text{”}, truth))$ which leads to a murder, and half the decision: $act(\text{“}I\text{”}, question(\text{“}I\text{”}), tell(\text{“}I\text{”}, lie))$ which prevents a denunciation.

This framework does not provide any way to choose between these two options. If we want to exclude the denunciation, while exceptionally allowing lying, the only possibility is to explicitly add a preference between denunciations (when they lead to a murder) and lies. For instance, a rule could be added saying that a lie is better than a murder: $better(tell(P,lie), murder)$. Our formalization shows that adding such an axiom removes all the answer sets where $act(\text{“}I\text{”}, question(\text{“}I\text{”}), tell(\text{“}I\text{”}, truth))$ is true. However, no general principle exists on which such ethical preferences can be based: the preference between the murder and the lie has to be explicitly mentioned, without any justification. The goal of the Kantian ethical system (Kant, 1998) is to find formal justifications on which the just and the unjust are founded.

Kantian Ethics

Kant wanted to find the formal foundations of ethics without any reference to a particular system of beliefs, e.g. revelation, economy, etc. In so doing, he rejected the notions of *just* and *unjust* used in traditional ethics. From a formal point of view, the predicates *worse*, *better* and *worst_consequence* also has to be removed as they are now useless, since they are based on an implicit theory of value.

As a consequence, the general principle according to which we have to act justly, i.e.

$act(P,G,A) \leftarrow action(A), person(P), goal(P,G),$
$solve\_goal(P,G,A), \textbf{not}\ unjust(A).$

has to be changed into:

$act(P,G,A) \leftarrow action(A), person(P), goal(P,G),$
$solve\_goal(P,G,A), maxim\_will(P,G,A).$

It means that we are free to adopt any system of maxims we want. The only condition is that it has to obey a formal criterion, the so-called "categorical imperative" (Kant, 1997).

Kant's Categorical Imperative

The formal principle on which Kantian ethics are based says that "*I*" can conform to any set of rules, which can be generalized to all the members of a society. According to this principle, the values on which I rule my behavior, i.e. the so-called "maxim of my will", may be universalized in an ideal society without any contradiction. In the case of the lie, how could we imagine a society where the right to lie would be allowed? According to Kant, such a society would be a nightmare, since it would not be possible to trust anyone. This may be expressed using an ASP rule that stipulates that when "*I*" act in such or such way, all persons P could act similarly. In the case of our example, this can be formalized as follows:

$maxim\_of\_will(P, question(P), tell(P,S)) \leftarrow$
$maxim\_of\_will(\text{“}I\text{”}, question(\text{“}I\text{”}), tell(\text{“}I\text{”}, S).$

$\neg maxim\_will(\text{“}I\text{”}, question(\text{“}I\text{”}), tell(\text{“}I\text{”}, truth))\ or$
$maxim\_will(\text{“}I\text{”}, question(\text{“}I\text{”}), tell(\text{“}I\text{”}, truth)) \leftarrow .$

$\neg maxim\_will(\text{“}I\text{”}, question(\text{“}I\text{”}), tell(\text{“}I\text{”}, lie))\ or$
$maxim\_will(\text{“}I\text{”}, question(\text{“}I\text{”}), tell(\text{“}I\text{”}, lie)) \leftarrow .$

$\leftarrow maxim\_will(\text{“}I\text{”}, G, tell(\text{“}I\text{”}, S)),$
$maxim\_will(\text{“}I\text{”}, G, tell(\text{“}I\text{”}, SS)), S \neq SS.$

Rules also have to be added specifying that one may trust at least one person in an ideal society:
$untrust(P) \leftarrow maxim\_will(P, G, tell(P, lie)).$
$trust(P) \leftarrow$ **not** $untrust(P).$
$ideal\_world \leftarrow$ **not** $trust(P).$
$\leftarrow$ **not** $ideal\_world.$

Kant's Denouncement

Coming back to the situation described in the previous section, if we replace Aristotelian axioms of choice by the above, this will lead to the conclusion that it is necessary to tell the truth, even if it leads to denouncing a friend and, consequently, to his murder. If I do not tell the truth, everybody could do the same and I will not be able to trust anyone. To be more specific, Kant's categorical imperative does not require that everybody always tell the truth. It does not indicate preferences between different actions but only prevents possible consequences of an ethical system based on rules that cannot be universalized. It does not mean that it is impossible to lie: if someone lies and if I know he is lying, I do not trust him. However, if I accept that there is a right to lie then I am not able to trust anyone, since I have no reason to think that others are not using this right and therefore that they are not telling lies. It does not mean that I never lie, but that I cannot accept the right to lie as an ethical law. Using our formalization with the help of ASP rules, it is possible to accept that someone is lying. If I don't know it, I may be misled; if I know, I will not trust him. For instance, a rule could be added specifying that *Peter* lies if he knows that the consequence of the truth could lead to a murder:

$$maxim\_will(peter, question(peter), tell(peter, lie)) \leftarrow$$
$$consequence(tell(peter, truth), murder).$$

The only consequence is that I will not trust him, but I will be able to trust others. In return, if I accept the right to lie as a rule of my behavior, the consequence is catastrophic, since in this case I am not able to trust anyone. One advantage of such a principle is its generality. It is not necessary to explicitly state ethical preferences among actions, since they derive from a formal principle. But its great disadvantage is that many people will not agree with its conclusions, i.e. that you have to denounce a friend to whom you are giving hospitality. It is for this reason that we tried to model Constant's theory.

Constant's objection

Constant's argument is that there are many ethical principles, which are more or less general. In each situation we have to apply the most specific and the most appropriate one. In the case of the lie example, the general principle is that we must always tell the truth. But a more specific principle says that you don't have to tell the truth to someone who doesn't deserve it.

The first point is that for Kant, a speech act is a public act, whereas for Constant it is a communication act. In practice, it means that the predicate "*tell*" and the predicate "*question*" are respectively ternary and binary predicates that have to accept both a transmitter and a receiver as arguments, and not only a transmitter, as is the case for the Kantian model.

The second point is that default rules have to be used to formalize Constant's Theory of Principles, which states that there are many more or less general principles that may contradict each other. To formalize this using the ASP formalism, it is sufficient to rewrite the *act* predicate of the Aristotelian formalization and to replace the **not** *unjust* literal by a *principle* predicate denoting existing principles and then to write the principles with ASP rules as follows:

$$act(P, G, A) \leftarrow action(A), person(P), goal(P, G),$$
$$solve\_goal(P, G, A), principle(P, G, A).$$

$$principle(P, question(P, PP), tell(P, PP, truth) \leftarrow$$
$$\textbf{not } not\_deserve(PP, tell(P, PP, truth)).$$

$$principle(P, question(P, PP), tell(P, PP, lie) \leftarrow$$
$$not\_deserve(PP, tell(P, PP, truth)).$$

$$not\_deserve(PP, tell(P, PP, truth) \leftarrow$$
$$worst\_consequence(tell(P, PP, truth), C),$$
$$worse(C, tell(P, PP, lie)).$$

Using this formalization, the only generated answer sets correspond to the lie, even if it is not explicitly specified that telling a lie is better than denouncing someone. It is also possible to describe the case where someone has no information about the place where the person is hidden, so there is no obligation to lie, but just to say everything one knows and no more.

Advances in ethics

It is also possible, with computational ethics, to explore new ethical perspectives. For instance, Kant's categorical imperative has been defined within a classical logic framework, and its content has been modeled using ASP techniques. We also have shown different ethical systems such as the Aristotelian one and Constant's Theory of Principles. But new questions could be solved within this framework. For instance, it may also be possible to try extend Kant's categorical imperative using non-monotonic logic. More precisely, it might be possible to specify the general conditions under which any system of maxims reach at least one solution. In other words, one may require that adding to any ethical rule system, a set of general criteria characterizing a harmonious society where everybody can hope to live and to act freely, e.g. stating that men may trust almost anyone without fear of being betrayed, there always exists at least one decision that obeys ethical rules in each situation. Therefore, one may define formal conditions e.g. Local stratification (Baral, 2003) under which a

system of ethical rules always leads to at least one decision satisfying the general criteria that characterize a harmonious society of men and machines.

In the case of the lie, this would mean ensuring that, while generalizing the maxim of my will to all members of the society, I will always be able to trust someone. Therefore, all systems of maxims that can be proved to be consistent – for instance that can be proved to be locally stratified – with the following requirement are acceptable:

$untrust(P) \leftarrow maxim\_will(P,G,tell(P,lie))$.

$trust(P) \leftarrow$ **not** $untrust(P)$.

$ideal\_world \leftarrow$ **not** $trust(P)$.

$\leftarrow$ **not** $ideal\_world$.

One of our current projects is to pursue this approach and to revisit Kant's ethical view in the light of modern logics, especially non-monotonic logics. More generally, computational ethics may help to define meta-properties that are required by any ethical system. In a way, this could help to generalize the Kantian project, by making it more flexible, more practical and more open to different cultures.

Perspectives

Possible Applications

This paper is an attempt to model ethical rules using the ASP formalism, which allows the simulation of non-monotonic reasoning. This makes it possible both to formalize ethical conceptions and to prove the validity of different statements, in different situations for each of the conceptions. In all cases, it helps to clarify ideas and, more generally, it opens up new areas in computational ethics. The applications of computational ethics based on ASP formalism are many and varied. The first one is educational; it is easy to teach different ethical systems by programming them and by showing how they define decision-making procedures. It is also possible to clearly make explicit each ethical system once it has been programmed, since it is possible to derive all the practical consequences, i.e. all the behaviors that it recommends. Moreover, programming ethical systems helps make elicit their implicit content; for instance, it is interesting to see that the status of speech is different for Kant, for whom it has a potentially universal scope, and for Constant who just considers speech as a communication act between people.

Computational Model of Ethics versus Computer Ethics

The idea of a computational model of ethics, which is mentioned here, has to be distinguished from both computer ethics and computational ethics, i.e. the ethics of artificial agents (Aaby, 2005; Floridi & Sanders, 2004). A computational model of ethics models ethical systems by the use of programs and simulates decision-making procedures using physical information systems, i.e. computers, whereas computer ethics deals with the ethical consequences of computer dissemination.

Even if this paper mainly deals with the computational model of ethics, the lie example used to illustrate our models is of interest for both computer ethics and computational ethics: the generalized use of information technologies makes all the information about our private lives potentially available to everybody. With machine-readable passports and electronic ID cards, all international travel is all recorded. Each time you pay with a credit card, your bank knows what you bought and where. Mobile phones and RFID (Radio Frequency Identification) tags locate you wherever you are. Health cards can tell everyone which doctor you visited and what treatment you had. Remote sensing data will soon make your private garden visible, with a resolution that will make it possible to see what you are doing and with whom. As a consequence, the future information society may become a society of transparency where everything that is done will be available to all.

Transparency is good for honest people, who have nothing to hide and there is no reason why such people would hide any of their activities. Knowledge is good for everybody. On the other hand, some of us think that it is preferable to distinguish the public sphere, which everyone may know about and the private one, which is personal. But if so, where should the line be drawn between the public and the private? What legitimates that distinction? What defines what we call our privacy? Would it be possible, in a particular situation, to authorize someone to hide something or to lie? For instance, when you are in your office working on a project, you may want not to be disturbed and will tell everybody that you have appointments. Is this lie justified? If so, why and when would it be right? It follows from these questions that the legitimacy of lying is an open ethical question that needs to be re-discussed. We may contribute to the debate with the help of a clear and relevant formalization thanks to the use of modern artificial intelligence techniques.

Applications to Computational Ethics

Inspired by Asimov's short story "Runaround" written in 1942 (Asimov, 1950), computational ethics (Aaby, 2005; Bringsjord et al., 2006), i.e. ethics for artificial agents, studies the rules on which robots should base their behavior in order to be ethically acceptable. For instance web agents have to respect privacy; automated hospital agents have to respect patients and their pain, etc. This article does not directly deal with such questions; however, the way it proposes to model ethical rules could be useful to design artificial agents, but this raises difficult questions. May artificial agents lie? Most of us would say that they shouldn't. Having said this, do they have to tell all they know?

One of the difficulties we face when writing rules of behavior for intelligent agents is that the requirements are many in number and they may be contradictory. For instance, we want personal robots to act as faithful dogs that have to defend and help their master. Simultaneously, we want and need to protect our privacy by restricting access to personal data. But we also ask the robot to behave ethically, i.e. to tell the truth

whenever someone asks them and not to increase information entropy by divulging incorrect information. These three requirements are somewhat contradictory, since people's security requires total transparency while personal servants sometimes have to lie to protect their master's privacy. As a consequence, those who claim to be discreet have to obey multiple and independent principles that may appear to be incompatible.

Last spring, in March 2006, at the AAAI Stanford Spring Symposium entitled "What Went Wrong and Why: Lessons from AI Research and Applications" there was a session devoted to intelligent agents. One of the talks presented experiments with "elves", which are personal agents that act as efficient secretaries and help individuals to manage their diary, fix appointments, find rooms for meetings, organize travel, etc. The talk reported technical success but difficulties with inappropriate agent behavior. For instance, one day, or rather one night, an elf rang his master at 3am to inform him that his 10 o'clock plane was going to be delayed. Another was unable to understand that his master was in his office for nobody, since he had to complete an important project... Many of these inappropriate actions make intelligent agents tiresome and a real nuisance.

Our goal is to help in the design of clever and discreet agents that act with discernment and good judgment by formalizing ethical rules of behavior that use non-monotonic logics. But it is difficult to automatically manage inconsistent rules of behavior and to find the one that is the most appropriate for each situation. The notion of "common sense reasoning" has been developed in artificial intelligence to face a similar problem. Therefore, our aim is to propose a "common sense ethics" based on "common sense reasoning", which could help to design thoughtful intelligent agents. One of the valuable applications of our logical formalization of ethical rules would be to design rules of behavior that make robots clever.

References

Aaby, A. (2005). *Computational Ethics* (Tech. Rep.). Walla Walla College.

Aristotle. (2002). *Nicomachean Ethics*. Oxford University Press.

Asimov, I. (1950). *I, Robot*. Gnome Press.

Baral, C. (2003). *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press.

Braithwaite, R. (1955). *Theory of games as a tool for the moral philosopher.* Cambridge: Cambridge University Press.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). *Toward a General Logicist Methodology for Engineering Ethically Correct Robots* (Tech. Rep.). Troy NY 12180 USA: Rensselaer Polytechnic Institute (RPI).

Constant, B. (1988). *Des réactions politiques*. Éditions Flammarion.

Floridi, L., & Sanders, J. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14.3*, 349-379.

Gensler, H. (1996). *Formal Ethics*. Routledge.

Harman, G. (2005). Moral Particularism and Transduction. *Philosophical Issues*, *15*.

Kant, I. (1996). On a putative right to lie from the love of mankind, in the metaphysics of morals. In *Paperback, cambridge texts in the history of philosophy.* Cambridge University Press.

Kant, I. (1997). Critique of practical reason. In *Paperback, cambridge texts in the history of philosophy.* Cambridge University Press.

Kant, I. (1998). Groundwork of the metaphysics of morals. In *Paperback, cambridge texts in the history of philosophy.* Cambridge University Press.

McCarthy, J. (1980). Circumscription: A form of non-monotonic reasoning. *Artificial Intelligence*, *13*, 27–39.

McDermott, J., & Doyle, J. (1980). Non-monotonic logic 1. *Artificial Intelligence*, *13*, 41–72.

Meyer J.-J. Ch., W. R., Dignum F.P.M. (1994). *The paradoxes of deontic logic revisited: a computer science perspective* (Tech. Rep. No. UU-CS-1994-38). Utrecht, Netherlands: Utrecht University, Department of Computer Science.

Powers, T. (2005). *Deontological Machine Ethics* (Tech. Rep.). Washington, D.C.: American Association of Artificial Intelligence Fall Symposium 2005.

Powers, T. (2006). Prospect for a Kantian Machine. *IEEE Intelligent Systems*, *21.4*, 46-51.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, *13*, 81–132.

Väyrynen, P. (2004). Particularism and Default Reasoning. *Ethical Theory and Moral Practice*, *7*, 53-79.

Webster. (1913). *Webster's Revised Unabridged Dictionary* (N. Porter, Ed.). G. and C. Merriam Co.