PROCEEDINGS OF THE TENTH INTERNATIONAL CONFERENCE

Living, Working and Learning Beyond Technology

ETHICOMP 2008*

University of Pavia, Mantua, Italy
24 to 26 September 2008

EDITED BY

TERRELL WARD BYNUM
MARIACARLA CALZAROSSA
IVO DE LOTTO
SIMON ROGERSON

University of Pavia
De Montfort University
Southern Connecticut State University
East Tennessee State University

Title	Living, Working and Learning Beyond Technology
Edited by	Terrell Ward Bynum, Maria Carla Calzarossa, Ivo De Lotto, Simon Rogerson
ISBN	978-88-902869-9-5
Local	Mantova, Italy
Date	September 2008
Publisher	Tipografia Commerciale, Mantova (Italy)

Papers will subsequently appear in the ETHICOMP electronic journal

No responsibility is accepted for the accuracy of the information contained in the text or illustrations. The opinions expressed in the papers are not necessarily those of the editors or the publisher

IS COMPUTER ETHICS COMPUTABLE?

Gaetano Aurelio Lanzarone and Federico Gobbo

Abstract

We discuss aims and difficulties of formalizing ethical knowledge and reasoning by using either 'axiomatic' or 'situated' methods. We then present an approach, based on Machine Learning and Computational Logic techniques, to combine abstract ethical rules with empirical knowledge acquired by concrete experience, and to represent reasoning with rules that allow exceptions. A provisional conclusion of the paper is that, while computer ethics hardly seems to be computable in general, at least part of it may be amenable to a suitable computational formalization, so that the nature of the un-computable 'residue' might appear more clearly.

1. Introduction

Attempting to formalize ethical knowledge and reasoning serves two purposes: understanding human ethics and designing computer ethics. While the former is descriptive, subject to the intricacies of human behavior and scarcely prone to systematic experimentation, the latter is prescriptive, can be experimented with few limitations and has to do with 'ideal' ethical behavior, which restricts the class of models we are interested in. The need to instill ethical guidance into artificial agents, besides its speculative interest, is related to the practical problems arising from the building of (semi-)autonomous intelligent robots, to be deployed not only in special environments inaccessible to humans but also living within the human environment. Logical and computational formalization of ethics could be useful both for artificial agents and for the human agents designing them.

Roughly, two main approaches are possible. In the 'axiomatic' approach, a set of rules is established, from which an agent derives ethical behavior. In the 'situated' approach, the agent is immersed within an environment from which it informally absorbs good behavior. Both approaches have advantages and pitfalls. In the former, the human *desiderata* can be expressed, but their context-dependent application is far from guaranteed; the dangers of hard coding behavioral rules are well known (and will be recalled in the following). In the latter approach, creating ethical robots that learn from scratch is difficult, since such learning does not scale up from simple to more complex capabilities; ethical rules cannot be entirely dispensed with, otherwise un-principled behaviors might emerge.

It is often understood, however, that the two approaches can usefully coexist, and it may be useful to adopt a mixed approach. In fact, this would correspond to an epistemological stance advanced by several relevant authors; to mention one: *Thus truly ethical behaviour does not arise from mere habit or from obedience to patterns or rule. Truly expert people act from extended inclinations, not from precepts, and thus transcend the limitations inherent in a repertoire of purely habitual responses* (Varela, 1999) (pp. 30-31).

Isaac Asimov's Laws of Robotics are celebrated as the first and full-fledged attempts at defining rules of robot ethics. Among several authors, Roger Clarke (Clarke, 1993-94) has discussed, in guise of a thought experiment, how Asimov's robot stories, started in 1940 and continued over the following 45 years, explore the implications of these laws, which are appealing for their simplicity but whose application encounters several difficulties. Clarke's conclusion is that serious doubts arise about the possibility of devising a set of rules that provide reliable control over machines.

Two main problems have been evidenced. The first is related to ambiguities in the language used, so that the robot does what it was told, but not what was intended. To abide by the First Law: "A robot may not injure a human being, or, through inaction, allow a human being to come to harm", the robot has to interpret the vague term 'injure' in a given circumstance to the extent that, for example, it has to take into account psychological injury as well as physical.

The second problem has to do with conflicts among the laws and even within a single law. The prioritization of the ethical rules may lead to exceptions being invoked because one value is deemed more important than another. For a single rule, while it has to be followed *prima facie*, exceptions need to be considered in practical circumstances. For example, telling the truth is right and deception is wrong, except when lying is acceptable behavior, e.g. lying to avoid causing another person damage. Exceptions to exceptions can also arise; for example, hurting other people is wrong, except when acting in self-defense, but only unless the self-defensive reaction is not disproportioned.

Artificial Intelligence has developed a rich set of methods of knowledge representation and reasoning, which can be considered to be adopted in ethics. In this paper, we examine some of them and discuss their application to mitigate the problems encountered in the axiomatic approach.

In Section 2 we present a method to combine ethical rules with empirical knowledge acquired by concrete experience, in order to cope with linguistic ambiguity. Obeying ethical rules being similar to abiding by laws, we consider an approach developed for the interpretation of open-textured terms in legal rules, based on precedent cases, and an extension of this method which includes abstraction through taxonomies and categorization. We also consider analogical reasoning as a means to interpret an unclear situation by mapping it into a known situation, related to the first by a certain degree of similarity.

In Section 3 we argue that ethical rules can hardly be considered absolute laws, therefore it is often necessary to determine to what extent the rules apply to exceptional situations. This consideration leads us to select the so-called 'common-sense reasoning' formalization as the most correspondent to taking decisions in uncertain real-world circumstances.

Concluding this analysis, in Section 4, after commenting some technical aspects of the examples shown in the previous sections, we finally propose a provisional answer to the question raised in the title.

2. Open-textured terms

The notion of open-textured, or vague, terms, rooted in philosophical investigations of language, was introduced in jurisprudence by Hart (Hart, 1962). Hart discusses (Chapter 7) the open structure of law, where two main devices are employed to communicate criteria of good behavior: legislation, which makes the maximum use of linguistic terms, and precedent, which minimizes the use of terms. He exemplifies the two approaches as follows. A parent says to his son before entering a church: “All men and boys must take off their hats before entering church”. Another, as he uncovers his head while entering a church, says: “Look, this is how one must behave on these occasions”. Hart argues that both ways are uncertain, since they have an open structure which is unavoidable *because we are men, not gods.*

Several researchers, concerned with building formal models of law and knowledge-based legal systems, have addressed the problem of open-textured terms and its importance for statutory interpretation, i.e. *the process of trying to determine the meaning of a legal rule by analyzing its terms and then applying it to a particular set of facts* (Rissland and Skalak, 1989). We consider here ethical rules as having similar character as legal rules, and discuss how approaches developed to formalize legal knowledge can usefully be considered in the context of the representation and use of ethical rules.

Given the impossibility of defining terms for all possible meanings under all possible circumstances, a more viable approach is to interpret a vague term by learning from experience. In a computational setting (machine learning) there are two main approaches: learning based on explanation (deductive learning) and learning based on similarity (inductive or analogical learning). We will sketchily consider both in what follows.

2.1 Explanation-Bases Learning

Explanation-Based Learning (EBL) is a machine-learning technique, by which an intelligent agent can learn through the observation of examples (see (Ellman, 1989) for an overview). Differently from other methods of concept learning, EBL creates generalizations of given examples on the basis of background domain knowledge (thus it is also called Explanation-Based Generalization). For our purposes, we consider EBL's domain knowledge as corresponding to ethical rules, and EBL's training examples as corresponding to precedent cases. By making the interpretation of vague terms as guided by precedents, we use EBL as an effective process, capable of creating a link between terms appearing as open-textured concepts in ethical rules, and terms appearing as ordinary language wording for stating the facts of a previous experience.

EBL is usually presented in the literature as follows (Mitchell et al., 1986).

Given in input:

1. a Target Concept (TC): a predicate representing the concept to be learned;
2. a Training Example (TE): a set of facts constituting an example of the target concept;
3. a Domain Theory (DT): a set of facts and rules representing background knowledge about the domain;
4. an Operationality Criterion (OC): a set of predicates in terms of which TC has to be defined;

EBL gives in output, using DT, a definition of TC which generalizes TE and satisfies OC.

The EBL algorithm consists of two stages:

1. Construct an explanation in terms of the domain theory that shows how the training example satisfies the target concept definition. Each branch of the explanation structure must terminate with an expression that satisfies the operationality criterion.

2. Determine a set of sufficient conditions under which the explanation holds, stated in terms that satisfy the operationality criterion.

To explain how the process goes, let us consider the following so-called 'suicide' example, from (DeJong and Mooney, 1986)). It is represented in Prolog: each rule (Horn clause), containing first-order predicates, is a logical implication, with the conclusion on the left of the reversed implication symbol ':-' and a conjunction (symbol ','), possibly empty, of conditions on the right; if the conjunction is empty, the ':-' symbol is dropped, and the rule is called a fact.

The Domain Theory is given by the following rules:

```
kill(A,B):-hate(A,B),possess(A,C),weapon(C).
hate(W,W):-depressed(W).
possess(U,V):-buy(U,V).
weapon(Z):-gun(Z).
```

The Training Example is a set of facts:

```
depressed(john).
buy(john,obj1).
gun(obj1).
```

The Operational Predicates are defined by means of the following facts:

```
operational(depressed).
operational(buy).
operational(gun).
```

The Target Concept is the binary predicate 'kill'. Applying the EBG algorithm to the goal: kill(john,john), EBG generates the clause:

```
kill(X,X):-depressed(X),buy(X,Y),gun(Y).
```

which represents a generalization of the Training Example, usable for further cases. Notice that the new rule represents the concept of committing suicide, though no such name predicate is explicitly introduced.

In our context, we interpret non-operational predicates as open-textured terms in rules, and operational predicates as concrete wording in given situations. The role of the operationality criterion is crucial for EBL. It requires that the final concept definition be expressed in terms of the predicates used to describe the training example, or in terms of a selected set of easily evaluated predicates from the domain theory. This means that learned concept definitions need not only be correct, but also in a form usable by a particular agent for a specific task. Which predicates to consider operational is debated in the EBL literature, and EBL systems usually rely on various operationality heuristics. The simplest one is to consider some predicates as operational *a priori*; a more sophisticated approach is to state conditional operationality in rules, which can therefore be reasoned about themselves. In any case, the concept of operationality constitutes a link between the abstract terms in which ethical or other stipulating rules are expressed, and ordinary language terms in which the facts of a situation are phrased.

The standard application of EBL, as shown above, is purely deductive, having mainly the purpose of transforming a chain of inferences into a single rule, to be retrieved and used more efficiently in similar cases. In the Machine Learning field this procedure is called 'knowledge compilation' (see (Cadoli and Donini, 1998) for a survey). As such, it requires a complete, correct and tractable domain theory. In many real-world situations, however, only an incomplete domain theory is available. This is the case when a predicate appearing in conditions of rules is not defined by further rules, leaving the interpretation of some predicates completely open. In such cases, examples may again be taken as guiding the interpretation, but this time provided that they are used together with some additional knowledge.

One simple yet useful heuristics is to abstract single instances provided by cases to higher classes of an abstraction hierarchy. In the above example, 'gun' is a subclass of the class 'weapon', abstracting to which also e.g. 'machine-pistol' is accommodated. 'sleeping pill' does not belong to the class 'weapon' but may be used for suicide; a super-class of both may be expressed by the predicate 'dangerous object', and so on. The extension dealing with taxonomies or hierarchy trees allows climbing the hierarchy up to the most abstract node compatible with the constraints. These are represented by means of assertions that use the binary predicate 'compatible', expressing the fact that a class is compatible with its siblings and therefore with its super-class as a whole. In the suicide example, the taxonomy is expressed by means of facts with unary predicates:

```
weapon(gun).
weapon(machine_pistol).
toxic_substance(poison).
toxic_substance(sleeping_pill).
dangerous_object(weapon).
dangerous_object(toxic_substance).
```

We add the following facts to the Training Example:

```
    belongs_to(gun,weapon).
(*) compatible(weapon,dangerous_object).
```

In such a situation, an extended EBL process generates the final clause:

```
kill(X,X):-depressed(X),buy(X,Y),dangerous_object(Y).
```

which can be used to prove suicide with all the dangerous substances included in the hierarchy and not only with the gun of the previous Training Example. On the contrary, the lack of a 'compatible' assertion means that the abstraction is not possible. For instance, if fact (*) is not given, the generated clause is simply:

```
kill(X,X):-depressed(X),buy(X,Y),weapon(Y).
```

and does not generalize up to 'dangerous_object', but only to 'weapon'.

The extended EBL process exemplified above has been formalized within a meta-logic approach in (Bertarello et al., 1994) and applied to the open-textured terms problem in (Costantini and Lanzarone, 1995). This method is viable for taking specific situations into account, in order to transform general rules possibly containing vague terms into ready-to-use rules worded in ordinary terms. It recalls a cognitive significance suggested by Varela: *Through appropriate extension and attention and by training over time we have transformed these actions into embedded behaviour* (Varela 1999) (p. 35).

2.2 Analogical Learning

A second way of learning from experience is analogical learning. In Machine Learning, several authors (see (Leishman, 1990) and (Hall, 1989) for surveys) have pointed out that analogy is mainly concerned with a mapping between two domains. As, for instance, Winston puts it (Winston 1980), *analogy is based on the assumption that if two situations are similar in some respect, then they may be similar in other respects as well.* Thus, an analogy is a mapping of knowledge from a known 'source' domain into a novel 'target' domain.

The simplest approach to analogy is to transfer properties from source to target by using rules like "if the source is sufficiently similar to the target and has a property of interest, then ascribe that property to the target". (Winston, 1980) discusses analogy as a replacement of the source object with the target object on the basis of the following kind of inference: knowing that from premises A conclusion B follows, and that A' corresponds to A, analogically conclude B'. Let's consider again the 'suicide' example, restated as follows.

Source Theory S:

```
kills(X,Y):-hates(X,Y),has_weapon(X)
hates(john,george)
has_weapon(john)
```

Target Theory T:

```
hates(anne,joe)
hates(X,X):-depressed(X)
has_weapon(anne)
has_weapon(bill)
depressed(bill)
hates(marc,bruce)
violent(marc)
beats(john,carl)
```

Since *kills(john,george)* can be concluded in S and because of the first two clauses in T, then *kills(anne,joe)* can analogically be concluded in T by the correspondences of terms (john,anne) and (george,joe). Notice that also *kills(bill,bill)* can analogically be concluded in T by the correspondences (john,bill) and (george,bill).

Not only terms but also relations can be transferred. If we add:

```
analogous('kills','beats')
analogous('has_weapon,','violent')
```

beats(marc,bruce) can analogically be concluded.

A formalization of this method, again in computational meta-logic, is described in (Costantini and Lanzarone, 1991), (Costantini et al., 1995)

Quoting again from (Varela, 1999) (p. 27): According to Mencius, people actualize virtue when they learn to extend knowledge and feelings from situations in which a particular action is considered correct to analogous situations in which the correct action is unclear.

3. Absolute rules vs. rules with exceptions

The philosophical problems arising from considering absolute ('categorical') rules are well represented by the notorious Emmanuel Kant vs. Benjamin Constant debate. In the brief essay "On a Supposed Right to Lie Out of Love for Man" (first published in a Berlin magazine in 1797), Kant claimed that it is wrong to tell a lie even to save a friend from possible murder. The rebuttal by Constant was: *The moral principle: it is a duty to speak the truth, if taken unconditionally and in isolation, would make all society an impossibility*, and most later critics considered untenable Kant's refusal to admit exceptions to obeying a duty (see e.g. (Benton, 1982)).

Artificial Intelligence researchers have developed several approaches to dealing with rules with exceptions and common-sense (or, more technically, non-monotonic) reasoning (see e.g. (Ginsberg, 1987). When considering rules as propositions expressed in a rigorous language, logic is involved, but classical logic is inadequate and new logics have been devised.

In classical logic, statements are 'categorical', i.e. they are intended to always hold, with no exception (e.g.: all humans are mortal). Deductive reasoning is monotonic, that is, adding new axioms, the set of conclusions previously derivable cannot decrease.

In common-sense reasoning, statements are 'typical', i.e. they are intended to usually hold, but there can be exceptions. For example, vehicles are prohibited in a public park; the rule, however, does not apply to fire trucks, maintenance cars, vehicles for disabled people and so on. Since it is hardly feasible to specify all the possible exceptions, a default (also called 'closed-world') assumption is taken, that is, the typicality conclusion is derived about a case if it cannot be proved to be an exceptional case. But if information is added about that case being an exception to the typicality rule, then the conclusion is retracted (hence the term non-monotonic reasoning: adding new information, the set of previously derivable conclusions can decrease).

Let us consider the following rules and facts.

```
wrong_behavior(Agent):-performs(Agent,Action),
                       causes(Action,Effect),bad(Effect),
                       not exceptional(Agent,Action).
exceptional(Agent,Action):-reaction(Agent,Cause),
                           acceptable_motivation(Cause, Reason),
                           not disproportioned(Reason, Cause).
performs(agent_1,injurying).
performs(agent_2,injurying).
```

```
performs(agent_3,injurying).
performs(agent_4,lying).
performs(agent_5,lying).
causes(injurying,damage).
causes(lying,damage).
bad(damage).
reaction(agent_1,gentle_invitation).
reaction(agent_2, physical_attack).
reaction(agent_3, verbal_offense).
reaction(agent_4,valid_accusation).
reaction(agent_5,false_accusation_of_friend).
acceptable_motivation(physical_attack,violent_self_defence).
acceptable_motivation(false_accusation_of_friend, save_life_of_friend).
disproportioned(violent_self_defence,verbal_offense).
```

The connective 'not' appearing in rules above is not a classical but a non-monotonic negation, i.e., according to the default assumption, its argument is considered not to hold if it cannot be proved (from the given clauses) to hold. Therefore, in the example, the predicate 'wrong_behavior' holds for: agent_1, agent_4 (not exceptions), agent_3 (an exception of an exception), and doesn't hold for: agent_2 and agent_5 (exceptions).
Notice that, should we for instance add for agent_4 the information:

```
acceptable_motivation(valid_accusation, save_life_of_friend).
```

then the previous conclusion would no longer hold, hence the non-monotonicity of this kind of inference.

4. Concluding remarks

4.1 About formalisms

In summarizing what we have presented in this paper, a remark is in order with respect to the mentioned techniques. To give a flavour of the approaches involved and at the same time to avoid technicalities unnecessary in the present context, we have represented examples in Prolog, as probably the simplest and most understandable logical formalism. It is not, however, the most appropriate to the treated topics, for reasons whose discussion goes beyond the purpose of this paper.

On the other hand, more powerful and 'neat' formalisms are either too specialized or not amenable to computational use (the topic of the paper being 'computability' of ethics). For instance, deontic logics could be employed to represent the concepts of 'permitted' and 'prohibited', central in legal or ethical discourse. But these would not be appropriate to represent explanation-based and analogical learning. Similarly, other non-monotonic formalisms (such as those appearing in (Ginsberg, 1987)) would not be suitable to learning, not to mention that most of them are computationally intractable.

A general approach that has proved, in the experience of the first author (see an epistemological account in (Lanzarone 2003), better suited to encompass all the different techniques presented in this paper, is meta-logic programming, developed in (Costantini and Lanzarone, 1994a) and applied to EBL and analogy in the previously mentioned papers and to non-monotonic reasoning in (Costantini and Lanzarone, 1994b); the reader is referred to these papers for the formal details. In fact, most of the involved concepts are meta-theoretical in nature: analogy is a mapping between different domains, hence logical theories; common-sense reasoning is based on the meta-level 'closed world assumption', and learning in general implies working with changing axioms ('theory revision').

4.2 A provisional conclusion

We have supported the epistemological stance underlying the shown approaches with the work of two authors (Hart and Varela) that seem to converge to similar positions though starting, in different times, from very different backgrounds.

The axiomatic and the situated approaches are thus reconciled. On the one hand, without empirical experience, rule-based ethical systems cannot determine whether open-textured terms in rule antecedents match the current situation to be decided and acted upon. On the other hand, without the guidance of general rules, precedent cases are only fragmented knowledge, unsuitable to being carried on to new situations; analogical reasoning and abstraction principles are needed to fill in the knowledge gaps, for example, by noting similarities and considering more general classes encompassing the concepts of both rules and cases.

A provisional answer to the question raised in the title of this paper is that, while computer ethics does not seem amenable to finite axiomatizability and therefore is un-computable in general, at least part of it could be computed by supplementing ethical rules with empirical experience, gained and employed by learning. Studying how far the frontier of the computable part of computer ethics can be pushed, the nature might appear more clearly of the un-computable 'residue', as Alan Turing called it in the following passage (where by 'discipline' he means finite rules and by 'initiative' he means learning): *But discipline is certainly not enough in itself to produce intelligence. That which is required in addition we call initiative. This statement will have to serve as a definition. Our task is to discover the nature of this residue as it occurs in man, and try to copy it in machines* (Turing, 1948).

References

Benton, R.J. (1982), Political Expediency and Lying: Kant vs Benjamin Constant, Journal of the History of Ideas, 43, 1, 135-144, University of Pennsylvania Press.
Bertarello, S., Costantini, S. and Lanzarone, G.A. (1994), Extending Explanation-Based Generalization with Metalogic Programming, in: Alpuente, M., Barbuti, R. and Ramos, I. (eds), Procs. GULP-PRODE Joint Conference on Declarative Programming, Peniscola (Spain), Sept. 19-22, vol. II, 16-31.
Cadoli, M and Donini F.M. (1998), A Survey on Knowledge Compilation, AI Communications, 10, 3-4,137-150.
Clarke, R. (1993-1994), Asimov's Laws of Robotics – Implications for Information Technology, IEEE Computer 26, 12, 53-61 and 27, 1, 57-66.
Costantini, S. and Lanzarone, G.A. (1991), Metalevel Representation of Analogical Inference, in: Ardizzone, E., Gaglio, S. and Sorbello, F. (eds), Trends in Artificial Intelligence, Lecture Notes in Artificial Intelligence, 549, 460-464.
Costantini, S. and Lanzarone, G.A. (1994a), A Metalogic Programming Approach: Language, Semantics and Applications, International Journal of Experimental and Theoretical Artificial Intelligence, 6, 239-287.

Costantini, S. and Lanzarone, G.A. (1994b), Metalevel Negation and Non-Monotonic Reasoning, Journal of Methods of Logic in Computer Science, 1, 111-140.
Costantini, S. and Lanzarone, G.A. (1995), Explanation-Based Interpretation of Open-Textured Concepts in Logical Models of Legislation, Artificial Intelligence and Law, Kluwer Academic Publisher, 3, 3, 191-208.
Costantini, S., Lanzarone, G.A. and Sbarbaro, L. (1995), A Formal Definition and a Sound Implementation of Analogical Reasoning in Logic Programming, Annals of Mathematics and Artificial Intelligence, 14, 17-36.
DeJong, G and Mooney, R. (1986), Explanation-Based Learning: An Alternative View, Machine Learning,1, 145-176.
Ellman, T. (1989), Explanation-Based Learning: a Survey of Programs and Perspectives, ACM Computing Surveys, 21, 2, 163-221.
Ginsberg, M. (ed.) (1987): Readings in Nonmonotonic Reasoning, Morgan Kaufmann.
Hall, R.P. (1989), Computational Approaches to Analogical Reasoning: A Comparative Analysis, Artificial Intelligence, 39, 39-120.
Hart, H.L.A. (1961), The Concept of Law, Clarendon Press, Oxford.
Lanzarone, G.A. (2003), Computational Meta-languages: Theory and Applications, in: Cyrus, L., Feddes, H., Schumacher, F. and Steiner, P. (eds.), Language between Theory and Technology, The Deutscher Universitaets-Verlag, 123-134.
Leishman, D. (1990), An Annotated Bibliography of Works on Analogy, International Journal of Intelligent Systems, 5, 43-81.
Mitchell, T.M., Keller, R.M. and Kedar-Cabelli, S.T. (1986), Explanation-Based Generalization: A Unifying View, Machine Learning 1, 1, 47-80.
Hart, H.L.A. (1961), The Concept of Law, Clarendon Press, Oxford.
Rissland, E.L. and Skalak, D.B. (1989), Combining Case-Based and Rule-Based Reasoning: A Heuristic Approach, Procs. International Joint Conference on Artificial Intelligence, 524-530
Turing, A.M. (1948), Intelligent Machinery, reprinted in: Meltzer, B. and Michie D. (eds.) (1970), Machine Intelligence 5, American Elsevier, and in: Ince, D.C. (ed.) (1992), Collected Works of A.M. Turing: Mechanical Intelligence, North Holland, 87-106.
Varela, F. (1999), Ethical Know-How, Stanford University Press.
Winston, P. (1980), Learning and Reasoning by Analogy, Communication of the Association for Computing Machinery, 23, 12, 689-703.