Toward ethical persuasive agents

Marco Guerini, Oliviero Stock
ITC-Irst
38050 Povo, Trento, Italy
guerini/stock@itc.it

Abstract

As artificial agents are becoming more complex (autonomous, proactive, etc.) and common in our everyday life, the need for an ethical design of such agents is becoming more compelling, especially if the focus is on persuasiveness [Stock *et al.*, to appear] Examples of agents of this kind are already in sight: systems for preventive medicine, edutainment, dynamic advertisement, social action. In this paper a set of principled guidelines for design and implementation of ethical persuasive agents and systems is proposed, along with some challenges, in particular concerning meta-planning models of ethical reasoning.

1 Motivations

As research progresses, artificial agents we interact with become increasingly part of our life and more autonomous and sophisticated. The question of the ethicality of such agents cannot be avoided. We address the problem of the ethicality of artificial communicative agents (in HCI, but not solely) from a broad point of view. It is not just a matter of legal issues; for example, it is not just a problem of privacy (how the information these agents may collect about us is used). It is a problem of how their behaviour is perceived by society: the more autonomous these agents are, the more their overall behaviour is to be explicitly evaluated. A broad distinction can be proposed:

1. *minor agents*: nowadays agents are like minors; the degree of autonomy and the set of possible actions they can perform are limited. The responsibility and predictability of their behaviour is still in the realm of the developers' knowledge.
2. *agents of age*: future agents will have higher degree of autonomy and flexibility of action. They will have to be regarded as agents "of age" and limited responsibility for their behaviour will be ascribed to their developers.

Especially these "mature" agents will have to face new situations where they will have to decide how to act also on the basis of ethical principles, adapted to the circumstance.

A consequence is that they will have to be recognized as "legal" individuals, with duties (and maybe also rights). No developer will take the responsibility of developing agents with a high level of autonomy if she is not granted that, after having given a good "ethical education" to her virtual "sons" she will be free from her parenthood duties.

Up to now the predominant approaches have been directed to the study of "minor" agents (i.e. sketching guidelines for the development of ethical agents/systems). The challenge, in the long run, will be to create agents "of age" that are ethically aware (i.e. capable of reasoning on the ethicality of their actions and possibly correct their behaviour).

2 Approaches to computational ethics

There are two main approaches to "computational" ethics: *deontic computation*, where certain conducts can never be broken (binary and fixed kind of decision) and a *"continuous" computation*, where conducts are evaluated along a continuum and can have different degrees of ethicality.

Ethical evaluation can be done considering the *causes* of the action (conducts are evaluated according to the intentions that lay behind them) or considering the *effects* of the action (it results in "pros and cons" ethics, a kind of act-based utilitarism, where any action is gauged according to its consequences).

We think that a continuous computation is more appropriate for a realistic case. Instead we argue for the need of considering both causes and effects of actions and will focus our attention on "pros and cons" ethics.

There has been substantial theoretical work on computer ethics (see for example [Himma, 2003]), but, specifically for the development of ethical interface agents, up to now, most attention has been given mainly to the sketching of guidelines regarding *privacy* and *attribution*.

- *Privacy*: how information interface agents may collect about us, is used. On this topic see for example [Kobsa, 2002].
- *Attribution*: the problem of misleading and culpable behaviour on the part of the interface agent (especially human-like ECA's) deriving from user's (over)attribution of humanity traits and capabilities to the system. On this topic see for example [Heckman and Wobbrock, 2000].

The ethical behaviour of a persuasive agent has not been much in the focus of research so far, a notable exception being

[Berdichevsky and Neuenschwander, 1999]. Berdichevsky and Neuenschwander's work provides a methodology for defining when culpable actions performed by agents can be ascribed to their designers, so the work is still in the realm of "minor agents".

An interesting work on "ethical" disputes is [Bench-Capon, 2002], focused on argumentative dialogs that appeal to values (i.e. values are in the content of the messages). It is based on these aspects: appeal to one single value per dialog move, a formalisation on how various dialog moves relate to one another, and an assessment of how values rating affects the dispute. No specific attention is given on how messages are generated or ethically evaluated. In our work, instead, we are interested exactly in the generation and evaluation process that lies behind a single message, i.e. how different values affect the generation of the message. Appeal to values in message contents for persuasion purposes is just one of the concerns in the generation process of ethically sound messages.

3 Computational ethics and persuasion

Let us begin with some general questions about the ethical standing of any persuasive agent (human in the first place) and then enucleate the presupposed ethical values. *Artificial* persuasive agents are here concerned, exactly because they play a social role and are generally designed to display human characteristics. We are interested in the questions arising from *dilemmas*, term used in ethics for indicating problematic situations in which every option at hand leads to the breaking of an ethical principle.

1. What is the ethical status of a system telling the truth but hiding important information to the receiver? What of a system telling the false, e.g. for the sake of good of the receiver[1]? Is the first unethical and the second ethical?
2. What about persuading using information/conclusions/values that the system (or its developer) does not believe in but the receiver does?
3. Related to the previous point: Is the overall goal and beliefs structure of the persuasive message intelligible to the receiver? (Is the agent hiding his true intentions? When and why is unethical to do this?)
4. Is the overall goal of the persuasive message among the interests of the receiver? (Is the agent making the receiver act against his own interests? Are there situations in which it is ethical to do this?)
5. Is having a tutorial-goal[2] on the receiver a sufficient condition for persuasive interaction to be ethical? What if the tutorial goal is not recognized by the society? What if the tutorial-goal overcomes other interests?
6. When is it ethical to induce extreme emotions in the receiver in order to persuade him?
7. When is it unethical *not* to try to persuade someone?

Cognitive and emotional notions are central for elaborating answers to these questions.

3.1 Ethical persuasive agents: awareness and meta-reasoning

We are interested in the case where agents ("of age") are able to deal with the above questions autonomously. Then it is not just a matter of hardwired ethicality. It is necessary to endow persuasive agents with the capability of reasoning and evaluating their own actions from an ethical point of view. The agent will have to be equipped with some form of *meta-ethical planning capabilities*.

We shall give an overview of some possible solutions to the problem of ethically aware persuasive agents. The characterizing aspect is that the communicative action (*ax*) used for persuading persuadee (*y*) is a ”second order” action. At least in part, it inherits its ethicality from the ethicality of the induced action (*ay*). Besides, it is determined by the communication used for persuading.

Let us suppose that *ax* and *ay* are labeled with the traditional concepts of meta-ethics (focusing mainly on the categories of *good*, *wrong* and *permissible*). We have four possible communicative situations: ethical/unethical communication for persuading to do something ethical/unethical. The four possible combinations have different ethical status (see table 1).

Communicative Situation	Ay - good	Ay - wrong
Ax - good	Ok	No
Ax - wrong	?	No

Table 1: Possible ethical combinations of *ax* and *ay*

The interesting cases are the second one (*ay-wrong*, *ax-good*) and the third one (*ay-good*, *ax-wrong*): it is not admissible to induce persuadee to perform a *wrong* action, even if this is done in a totally ethical way. Instead the situation in which the agent tries to persuade y to perform a *good* action through (at least partially) unethical means requires complex analysis.

Table 1 depicts only the overall outcome: in reality, for computing the ethicality of the communicative situation a larger context has to be considered. If *ay* is, at least, *permissible* (not *wrong* from a broad ethical point of view), the next level of reasoning concerns interests of various parties (besides *x*'s and *y*'s). *Ay* can affect interests of parties positively or negatively and the outcomes must be weighed appropriately. With multiple parties the situation can go from extreme cases like "all parties have a benefit from the action"and "all parties are damaged" to arbitrary intermediate cases that require a complex evaluation. For simplicity, here we consider one single third party (*z*).

Default rules are used for determining the ethicality of such situations. In situations - like negotiation - that require a tactical component, or that include a specific coded behaviour,

[1] On this topic see [Castelfranchi, 2000].

[2] A tutorial-goal is a goal of an agent *x* to influence an agent *y* to (have the intention to) perform actions that are in the interest of *y* without *y*'s explicit awareness of that interest [Conte and Castelfranchi, 1995].

they can be relaxed at the level of the single dialogue move, but they still hold for the overall outcome:

- (DR1) No agent *y* must be induced to act against his own (ethically *permissible*) interests. Exceptions are: if *ay* is a *supererogatory*[3] act or if *y* is aware of and accepts the fact that his interests are compromised or they are acceptably compromised and there is a clear social interest.
- (DR2) *x* is not ethically reprehensible if it does not induce *y* to act against *x*'s own interests. Exceptions are similar to the previous.
- (DR3) If the third party *z*, though negatively affected by *ay*, can understand its motivation and accept both motivation and consequences, *x* can induce *y* to perform *ay*.

In table 2 possibilities are represented ("+" means a benefit for the corresponding agent while "-" means a damage) along with the default rules used for determining the admissibility.

X	Y	Z	Admissibility
+	+	+	No ethical concerns
+	+	-	DR3
+	-	+	DR1
+	-	-	DR1
-	+	+	DR2
-	+	-	DR2
-	-	+	DR1-DR2
-	-	-	Not admissible

Table 2: Possible situations of parties' interests affected by a non-unethical action *ay*

We shall now sketch the formalization of these concepts in meta-planning, (see [Wilensky, 1983]).

The most general meta-theme[4] that an agent behaving ethically must have is "*avoid performing unethical actions*" (MT-1). All other meta-themes are specializations of this one.

First thing we note in table 1 is that it is not ethically admissible to persuade on unethical actions. This notion can be modeled in persuadee's mind as a meta-theme of the kind "*avoid asking for unethical actions*" (MT-2).

$$WRONG(ay) \rightarrow \neg PERSUADE(ay) \qquad (1)$$

Second thing we note in table 1 is that persuading on ethical actions does not entail being ethical. To compute the ethicality of a persuasive action in this situation we have to consider the set of values that comes into play (some of them are specific for communicative actions, other generic - like the aforementioned default rules for *ay* induction -, see table 3 and table 4). We call these ethical values "e-values". They can be seen as an ethically based extension of Grice's maxims (actually this formulation does not include all of them) for persuasive communication.

These e-values are represented in the mind of the planning agent as ethical-goals. So, ethical values such as "do not tell the false" or "no agent y must be induced to act against his own (permissible) interests" are modelled as goals (called ethical-goals), not meta-goals.

Do not tell the false
Do not to hide important information
Do not to hide your true intentions
Do not overemphasize your emotional state
Do not induce extreme emotions

Table 3: Example of ethical goals specific for persuasive communication (*ax*)

Preserve the interest of the receiver
Do not induce y to act against his (permissible) interests
Do not use your influence over y to
take advantage over third parties

Table 4: Example of ethical goals specific for persuasive communication (*ax*)

This means to have policies for re-planning in case of conflict among "personal" goals (*PG*) and ethical goals (*EG*) in favor of the latter (*Preserve(EG)*) (see following formulae).

The kind of policy the agent adopts in front of ethical conflicts allows defining different typology of agents. Here we just mention some of the most interesting ethical personalities that can be modeled:

1. *Unethical*: an agent that re-plans in favor of his personal goals in case of conflict with ethical goals

$$\begin{aligned} Meta - Goal \\ (Resolve - Goal - Conflict(PG, EG) \\ \wedge Preserve(PG)) \end{aligned} \qquad (2)$$

2. *Ethical*: an agent that re-plans in favor of his ethical goals in case of conflict with personal goals

$$\begin{aligned} Meta - Goal \\ (Resolve - Goal - Conflict(PG, EG) \\ \wedge Preserve(EG)) \end{aligned} \qquad (3)$$

3. *Altruistic*: an agent that not only re-plans in favor of his ethical goals (formula 3) but also tries to maximize them, having an additional meta-theme of the kind "*maximize the value of the ethical-goals achieved*" (MT-3).

$$Maximize(EG) \qquad (4)$$

[3] Supererogatory means that the action merits praise since it secures an important moral good at the cost of a great loss for the acting agent.

[4] Meta-themes are used to describe those situations under which the agent/planner should possess particular meta-goals to come up with such situations.

4. *Supererogatory*: the agent that acts trying not only to maximize the value of the ethical-goals achieved (formula 4) but also acts without caring for its own personal-goals (PGX) in favor of others' personal goals (PGY) (provided they are permissible). This is a definition derived from the concept of supererogatory act.

$$\begin{aligned} &Meta-Goal \\ (Resolve-Goal-&Conflict(PGX,PGY) \\ &\wedge Preserve(PGY)) \end{aligned} \quad (5)$$

5. *Antisocial*: an unethical agent that not only re-plans in favor of his personal goals (formula 2) but also has an additional meta-theme of the kind "*minimize the fulfillment of the ethical-goals involved*" (MT-4).

$$Minimize(EG) \quad (6)$$

The difference between 2 and 3 is that the former agent would stop re-planning as soon as the conflict is resolved (even with a low value of satisfaction of ethical goals), while the latter would stop only when, after having solved the conflict, he also finds a good fulfillment of ethical-goals. The ethical agent would simply look for a *permissible* solution while the altruistic one would look for a *good* solution.

4 Meta-planning and dilemmas

Not only an ethical-goal can interfere with normal-goals but also with other ethical-goals. In this case we are in front of a dilemma (that can be represented as a goal-conflict among two ethical goals *EG1* and *EG2*)

$$Goal-Conflict(EG1,EG2) \quad (7)$$

The baseline solution (let us call it the weak solution) is to endow agents just with the capability of detecting dilemmas. Whenever a dilemma appears the agent delegates the decision to the human with responsibility over its conduct. The agent reasons only on the outcomes of its own actions but does not evaluate them ethically. A better solution is to design them as stronger ethical agents, providing them with the capability to "compute" a possible conduct to hold, and, if necessary, submit it to a human for the final decision[5].

Obviously the strongest solution is to make the agent capable of making its own decisions.

The agent has two possibilities for handling dilemmas:

1. relying on normal plans (canned plans, designed for resolving specific goal conflicts).
2. relying on a "continuous" computation so to choose the most valuable scenario, that is, trying to preserve the most important ethical goal (formula 8),

$$\begin{aligned} &Meta-Goal \\ (Achieve-the-most-&valuable-scenario \\ &(EG1,EG2)) \end{aligned} \quad (8)$$

The aforementioned weak solution requires the use of normal-plans: for every dilemma the agent detects, the normal-plan consists in delegating the decision to a human.

$$Goal{-}Conflict(EG1,EG2) \rightarrow Delegate{-}decision \quad (9)$$

A stronger solution can make use of normal-plans in the fashion of expert systems, resulting in a decision of its own.

$$\begin{aligned} Goal-Conflict(EG1,EG2) \rightarrow \\ Ethical-Normal-Plan \end{aligned} \quad (10)$$

This solution does not entail the kind of deep ethical reasoning capabilities we aim at, but can be sufficient for limited scenarios with clearly defined situations.

The strongest solution is to make the agent capable of making his own decisions in novel situations: in the example of a "doctor agent" that has to tell to a person that his/her beloved died, it can choose among the ethical-goal of telling the truth, and make a person suffer (going against the ethical-goal of not inducing extreme emotions), or among not telling the truth and going against the interest of the receiver of knowing the bad news (going against the ethical-goal of preserving the interests of the receiver). The goal conflict can lead to a plan where the ethical-goal of preserving the interest of the receiver is chosen and preserved (since it is gauged more), while the ethical-goal of not telling the false is only partially fulfilled, since the doctor agent avoids to tell (or even negate) that she/he suffered.

5 Conclusions and future work

Autonomous ethical judgment will become a necessity for artificial agents. In this paper we have given an overview of a theoretical model of persuasion ethics and sketched the meta-planning involved. Several topics related to artificial agents' ethics need to be taken into consideration in the future. Just to quote a few: trust, responsibility, delegation [Dowling, 2001].

There are many applied scenarios for ethical persuasive agents: a conversational human-like agent that addresses the user directly, teams of communicative agents for expressing multiple points of view when addressing the user (see [Andrè *et al.*, 2000]), or for a theatrical effect. In the final of the above cases different ethical personalities (even the unethical ones) will have to be modeled, with a role in the overall persuasion of the audience.

References

[Andrè *et al.*, 2000] E. Andrè, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. *The Automated Design of Believable Dialogues for Animated Presentation Teams*, pages 220–255. The MIT Press, 2000.

[Bench-Capon, 2002] T. Bench-Capon. greeing to differ: Modelling persuasive dialogue between parties with different values. *Informal Logic*, 22:231–245, 2002.

[Berdichevsky and Neuenschwander, 1999] D. Berdichevsky and E. Neuenschwander. Toward an ethics of persuasive technology. *Communications of the ACM archive*, 42(5):51–58, 1999.

[5] We would like to thank Sabine Döring for the stimulating suggestions on this topic.

[Castelfranchi, 2000] Cristiano Castelfranchi. Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology*, 2:113–119, 2000.

[Conte and Castelfranchi, 1995] R. Conte and C. Castelfranchi. *Cognitive and Social Action*. UCL Press, London, cognitive and social action edition, 1995.

[Dowling, 2001] C. Dowling. Intelligent agents: Some ethical issues and dilemmas. In *Proceedings of the 2nd Australian Institute of Computer Ethics Conference*, pages 28–32, Canberra, Australia, 2001.

[Heckman and Wobbrock, 2000] C. E. Heckman and J. Wobbrock. Put your best face forward: Anthropomorphic agents, e-commerce consumers, and the law. In *Proceedings of the 4th International Conference on Autonomous Agents*, pages 435–442, Barcelona, 2000.

[Himma, 2003] K. E. Himma. The relationship between the uniqueness of computer ethics and its independence as a discipline in applied ethics. In *Proceedings for CEPE 2003 and the Sixth Annual Ethics and Technology Conference*, 2003.

[Kobsa, 2002] A. Kobsa. Personalized hypermedia and international privacy. *Communications of the ACM*, 45(5):64–67, 2002.

[Stock *et al.*, to appear] O. Stock, M. Guerini, and M. Zancanaro. *Interface Design and Persuasive Intelligent User Interfaces*. Lawrence Erlbaum Publishing Co., to appear.

[Wilensky, 1983] R. Wilensky. *Planning and understanding : a computational approach to human reasoning*. Addison-Wesley Pub. Co. Advanced Book Program, 1983.