Can Machines Make Ethical Decisions?

Iordanis Kavathatzopoulos, Ryoko Asai

Can machines make ethical decisions?

Iordanis Kavathatzopoulos and Ryoko Asai

Department of IT – VII, Uppsala University, Sweden
{iordanis,ryoko.asai}@it.uu.se

Abstract. Independent systems and robots can be of great help to achieve goals and obtain optimal solutions to problems caused by the quantity, variation and complexity of information. However, we always face ethical issues related to the design as well as to the running of such systems. There are many problems, theoretical and practical, in integrating ethical decision making to robots. It is impossible to designed or run such systems independently of human wish or will. Even if we create totally independent decision making systems, we would not want to lose control. Can we create really independent ethical decision systems? Recent research showed that emotions are necessary in the process of decision making. It seems that it is necessary for an independent decision system to have "emotions." In other words, a kind of ultimate purpose is needed that can lead the decision process. This could make a system really independent and by that ethical.

Keywords: robots, systems, autonomous, independent, decision making, ethics, moral

1 Introduction

The development of Information Technology, systems, robots, etc., that are capable of processing information and acting independently of their human operators, has been accelerated as well as the hopes, and the fears, of the impact of those artifacts on environment, market, society, on human life generally. Many ethical issues are raised because of these systems being today, or in the future, capable of independent decision making and acting. Will these IT systems or robots decide and act in the right way or will they cause harm?

In situations where humans have difficulties perceiving and processing information, or making decisions and implementing actions, because of the quantity, variation and complexity of information, IT systems can be of great help to achieve goals and obtain optimal solutions to problems. One example of this is financial transactions where the speed and volume of information makes it impossible for human decision makers to take the right measures, for example in the case of a crisis. Another example is dangerous and risky situations, like natural disasters or battles in war, where the use of drones and military robots may help to avoid soldier injuries and

deaths. A third example comes from human social and emotional needs, for example in elderly care where robots may play an important role providing necessary care as well as to be a companion to lonely elderly people.

It is clear that such IT systems have to make decisions and act to achieve the goals for which they had been built in the first place. Will they make the right decisions and act in a proper way? Can we guarantee this by designing them in a suitable way? But if it is possible, do we really want such machines given the fact that their main advantage is their increasing independence and autonomy, and hence we do not want to constrain them too much?

There are many questions around this, most of which converge on the issue of moral or ethical decision making. The definition of what we mean by ethical or moral decision making or ethical/moral agency is a very much significant precondition for the design of proper IT decision systems. Given that we have a clear definition we will be able to judge whether an IT system is, 1) capable of making ethical decisions, and 2) able to make these decisions independently and autonomously.

2 Focus on the process of ethical decision making

Ethics and morals have originally the same meaning in Greek and Latin. However, today, in philosophy as well as in psychology we usually give them different meanings. "Ethics" is often used in connection to meta-philosophy or to psychological processes of ethical decision making, whereas the term "moral" is adopted when we talk about normative aspects or about the content of a decision. The distinction between content and process is important in the effort to define ethical or moral decision making.

In common sense, ethics and morals are dependent on the concrete decision or the action itself. Understanding a decision or an action being ethical/moral or unethical/immoral is based mainly on a judgment of its normative qualities. The focus on values and their normative aspects is the basis of the common sense definition of ethics.

Despite its dominance, this way of thinking causes some difficulties. We may note that bad or good things follow not only from the decisions of people but also from natural phenomena. Usually sunny weather is considered a good thing, while rainy weather is not. Of course this is not perceived as something related to morality. But why not? What is the difference between humans and nature acting in certain ways? The answer is obvious: Option, choice.

Although common sense does realize that, people's attachment to the normative aspects is so strong that it is not possible for them to accept that ethics is an issue of choice and option. If there is no choice, or ability of making a choice, then there is no issue of ethics. However this does not solve our problem of the definition of Autonomous Ethical Agents, since IT systems are actually making choices.

Now if ethics are connected to choice then the interesting aspect is how the choice is made, or not made; whether it is made in a bad or in a good way. The focus here is

on how, not on what; on the process not on the content or the answer. Indeed, regarding the effort to make the right decision, philosophy and psychology point to the significance of focusing on the process of ethical decision making rather on the normative content of the decision.

Starting from one of the most important contributions, the Socratic dialog, we see that *aporia* is the goal rather than the achievement of a solution to the problem investigated. Reaching a state of no knowledge, that is, throwing aside false ideas, opens up for the right solution. The issue here for the philosopher is not to provide a ready answer but to help the other person in the dialog to think in the right way [1, 2]. Ability to think in the right way is not easy and apparently has been supposed to be the privilege of the few able ones [3]. For that, certain skills are necessary, such as Aristoteles's *phronesis* [4]. When humans are free from false illusions and have the necessary skills they can use the right method to find the right solution to their moral problems [5].

3 Skills for ethical decision making

His philosophical position has been applied in psychological research on ethical decision making. Focusing on the process of ethical decision making psychological research has shown that people use different ways to handle moral problems. According to Piaget [6] and Kohlberg [7], when people are confronted with moral problems they think in a way which can be described as a position on the heteronomy-autonomy dimension. Heteronomous thinking is automatic, purely emotional and uncontrolled thinking or simple reflexes that are fixed dogmatically on general moral principles. Thoughts and beliefs coming to mind are never doubted. There is no effort to create a holistic picture of all relevant and conflicting values in the moral problem they are confronted with. Awareness of own personal responsibility for the way one is thinking or for the consequences of the decision are missing.

Autonomous thinking, on the other hand, focuses on the actual moral problem situation, and its main effort is to search for all relevant aspects of the problem. When one is thinking autonomously the focus is on the consideration and investigation of all stakeholders' moral feelings, duties and interests, as well as all possible alternative ways of action. In that sense autonomy is a systematic, holistic and self-critical way of handling a moral problem.

Handling moral problems autonomously means that a decision maker is unconstrained by fixations, authorities, uncontrolled or automatic thoughts and reactions. It is the ability to start the thought process of considering and analyzing critically and systematically all relevant values in a moral problem situation. This may sound trivial, since everybody would agree that it is exactly what one is expected to do in confronting a moral problem. But it is not so easy to use the autonomous skill in real situations. Psychological research has shown that plenty of time and certain condi-

tions are demanded before people can acquire and use the ethical ability of autonomy [8].

Nevertheless, there are people who have learnt to use autonomy more often, usually people at higher organizational levels or people with higher responsibility. Training and special tools do also support the acquisition of the skill of autonomy. Research has shown that it is possible to promote autonomy. It is possible through training to acquire and use the skill of ethical autonomy, longitudinally and in real life [9].

4 Tools for ethical decision making

IT systems have many advantages that can be used to stimulate autonomy during a process of ethical decision making. For example EthXpert and ColLab [10, 11, 12] is intended to support the process of structuring and assembling information about situations with possible moral implications. Analogous with the deliberation of philosophers throughout history as well as with the findings of psychological research on ethical decision making, we follow the hypothesis that moral problems are best understood through the identification of authentic interests, needs and values of the stakeholders in the situation at hand.

Since the definition of what constitutes an ethical decision cannot be assumed to be at a fix point, we further conclude that this kind of system must be designed so that it does not make any assertions of the normative correctness in any decisions or statements. Consequently, the system does not make decisions and its sole purpose is to support the decision maker (a person, a group or the whole organization) when analyzing, structuring and reviewing choice situations.

In the system, interests of each imaginable stakeholder are identified in a systematic procedure over six steps. 1) Define stakeholders: The system's focus on interests leads to an associative process of identifying related stakeholders. For each stakeholder that is directly involved in the situation there may be third party stakeholders that could influence it. The simple question of who is affected by a specific interest of a stakeholder will help the user to become aware of these. In EthXpert and ColLab the addition of stakeholders is very straightforward and therefore does not provide any obstacle to widening the scope of the problem. 2) Define for each stakeholder its interests: The user determines a set of relevant interests, specifically for each stakeholder. All interests that might relate and affect other stakeholders are important to consider and in the process of scrutinizing interests additional stakeholders will naturally become involved in the analysis. 3) Define how interests relate to other stakeholders: Determining how the interests and values of the stakeholders relate to other stakeholders draws a picture of the dynamics and dependencies in the situation. The considerations that are brought up when an interest is facing another stakeholder may therefore reveal important conflicts. Further, as described above, this approach may help to track down previously unidentified stakeholders, since the topics that are brought up in one relation are not necessarily unique to that and therefore will raise the inclusion of other stakeholders. 4) Define main options: The most apparent alternatives for handling the moral problem can be immediately stated. Usually main al-

ternatives are to their character mutually excluding in some aspect, similar to answering a question with "Yes" or "No". There is no obligation to apply such a polarization, but to make full use of the later stage of formulating compromise options it can be useful to consider whether such patterns exist. 5) Translate considerations: For each optional strategy the user is urged to state how the interests of the stakeholders are affected by the option if that option would be the final decision. The considerations from the interest-stakeholder matrix will not be automatically copied to the decision matrix. Instead the interest-stakeholder relationships will serve as background and incentive for considering how the different decision alternatives affect the stakeholders. 6) Define compromise options: To counter problems in the main options, i.e. unacceptable negative effects, compromise solution candidates can be forked from main alternatives. A compromise option will inherit considerations from its parent, but the user should revise these and determine the difference in effect between them. The feature is useful for considering many options that only differ partly. The intention is to allow any user to easily get an overview of the strengths and weaknesses of similar alternatives.

5 Non-independent ethical agents

Ethical decision support programs, like EthXpert and ColLab, can be integrated into robots and other decision making systems to secure that decisions are made according to the basic theories of philosophy and to the findings of psychological research. This would be the ideal. But before we are there we can see that ethical decision making support systems based on this approach can be used in two different ways.

During the development of a non-independent decision making system, support tools can be used to identify the criteria for making decisions and for choosing a certain direction of action. This means that the support tool is used by the developers, they who make the real decisions, and who make them according to the previous named philosophical/psychological approach [13].

Another possibility is to integrate a support tool, like EthXpert and ColLab, into the non-independent decision system. Of course, designers can give to the system criteria and directions, but they can also add the support tool itself, to be used in the case of unanticipated future situations. The tool can then gather information, treat it, structure it and present it to the operators of the decision system in a way which follows the requirements of the above mentioned theories of autonomy. If it works like that, operators of non-independent systems make the real decisions and they are the users of the ethical support tool.

A non-independent system that can make decisions and act in accordance to the hypothesis of ethical autonomy is a system which 1) has the criteria already programmed in it identified through an autonomous way in an earlier phase by the designers, or 2) prepares the information of a problem situation according to the theory of ethical autonomy, presents it and stimulates the operators to make the decision in a way compatible with the theory of ethical autonomy.

All this can work and it is possible technically. But how could we design and run a really independent ethical decision making system? However, before we can speculate on that it is important to address some issues shortly, regarding the criteria for independence.

6 Independent ethical agents

One is the issue of normative quality of the decisions made. Can we use this criterion for the definition of an independent ethical decision system? As we have already discussed this is not possible although it is inherently and strongly connected to common sense, and sometimes into research [14]. Normative aspects can be found in the consequences of obviously non-independent natural phenomena. Besides, there are always good arguments supporting opposite normative positions. So this cannot be a working criterion [15].

The alternative would be the capability of choice. Connected to this is the issue of free will. We could say that really independent systems are those that are free to decide whatever they want. However, this has many difficulties. There is theoretical obscurity around the definition of free will as well as practical problems concerning its description in real life situations. Furthermore, it is obvious that many systems can make "choices." Everything from simple relays to complex IT systems is able to choose among different alternatives, often in arcane and obscure ways, reminiscent of the way humans make choices. Then the problem would be where to put the threshold for real choice making.

If the ability to make choices cannot be the criterion to determine the independence of a decision system, then the possibility to control the system by an operator becomes interesting. Wish or effort to control, external to the system, may be something that has to be involved and considered. The reason of the creation of IT systems is the designers' and the operators' wish to give them a role to play. These systems come to existence and are run as an act of will to control things, to satisfy needs. It is an execution of power by the designers and the operators. We can imagine a decision system as totally independent, but even this cannot be thought without a human wish or will behind it. It is always a will for some purpose. It can be a simple purpose, for example to rescue trapped people in collapsed buildings, or an extremely complex purpose, like to create systems able of making independent decisions! In any case the human designer or operator wants to secure the fulfillment of the main purpose and does not want to lose control.

So the issue could be about possession of an original purpose, a basic feeling, an emotion. Indeed recent research in neurobiology and neuropsychology shows that emotions are necessary in the decision making process [16]. It seems that a rational decision process requires uninterrupted connection to emotions. Without this bond the decision process becomes meaningless. Another effect of the "primacy" of emotions and purposes is that very often heteronomous or non-rational ways to make ethical decisions are adopted, despite the human decision maker being able to think autonomously and rationally.

Thus the criterion for a really independent decision system could be the existence of an emotional base that guides the decision process. Human emotions and goals have been evolved by nature seemingly without any purpose. That may happen in decision systems and robots if they are left alone, but designers, operators, and humans would probably not want to lose control. So what is left? Can we create really independent ethical decision systems?

7 Non-independent ethical agents

The criterion of such a system cannot be based on normative aspects, or on the ability to make choices, or on having own control, or on ability of rational processing. It seems that it is necessary for an independent decision system to have "emotions" too. That is, a kind of ultimate purposes that can lead the decision process, and depending on the circumstances, even make the system react automatically, or alternatively, in a rational way.

Well, this is not easy to achieve. It may be impossible. However, if we accept this way of thinking we may be able to recognize a really independent or autonomous ethical agent, if we see one, although we may be not able to create one. This could work like a Turing test for robot ethics because we would know what to look for: A decision system capable of autonomous ethical thinking but leaning most of the time toward more or less heteronomous ways of thinking; like humans who have emotions leading them to make decisions in that way.

References

1. Πλάτων [Platon]: Θεαίτητος [Theaitetos]. Ι.Ζαχαρόπουλος [I. Zacharopoulos], Αθήνα [Athens] (1981)
2. Πλάτων [Platon]:'Απολογία Σωκράτους [Apology of Socrates]. Κάκτος [Kaktos], Αθήνα [Athens] (1992)
3. Πλάτων [Platon]: Πολιτεία [The Republic]. Κάκτος [Kaktos], Αθήνα [Athens] (1992)
4. 'Αριστοτέλης [Aristoteles]: 'Ηθικά Νικομάχεια [Nicomachean Ethics]. (Πάπυρος [Papyros], Αθήνα [Athens]) (1975)
5. Kant, I.: Grundläggning av Sedernas Metafysik [Groundwork of the Metaphysic of Morals]. Daidalos, Stockholm. (1785/2006)
6. Piaget, J.: The Moral Judgement of the Child. Routledge and Kegan Paul, London (1932)
7. Kohlberg. L.: The Just Community: Approach to Moral Education in Theory and Practice. In: Berkowitz, M. and Oser, F. (eds.) Moral Education: Theory and Application, pp. 27-87. Lawrence Erlbaum Associates, Hillsdale, NJ. (1985)
8. Sunstein, C. R.: Moral Heuristics. Behavioral and Brain Sciences, 28, 531-573 (2005)
9. Kavathatzopoulos, I.: Assessing and Acquiring Ethical Leadership Competence. In: Prastacos, G.P. et al. (eds.) Leadership through the Classics, pp. 389-400, Springer-Verlag, Berlin Heidelberg (2012)

10. Kavathatzopoulos, I. and Laaksoharju, M.: Computer Aided Ethical Systems Design. In: Arias-Oliva, M. et al. (eds.) The "Backwards, Forwards, and Sideways" Changes of ICT, pp. 332-340. Tarragona, Spain: Universitat Rovira i Virgili (2010)
11. Laaksoharju, M.: Let us be Philosophers! Computerized Support for Ethical Decision Making. Uppsala University, Department of Information Technology, Uppsala (2010)
12. Laaksoharju, M. and Kavathatzopoulos, I.: Computerized Support for Ethical Analysis. In: Botti, M. et al. (eds.) Proceedings of CEPE 2009 – Eighth International Computer Ethics and Philosophical Enquiry Conference. Kerkyra, Greece: Ionian University (2009)
13. Kavathatzopoulos, I.: Philosophizing as a usability method. CEPE 2013, Ambiguous Technologies: Philosophical Issues, Practical Solutions, Human Nature. Universidade Autónoma de Lisboa, Lisbon (in press).
14. Kohlberg, L.: The Philosophy of Moral Development: Moral Stages and the Idea of Justice. Harper and Row, San Francisco (1984)
15. Wallace, W. and Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, New York (2009)
16. Koenigs, M. and Tranel, D.: Irrational Economic Decision-Making after Ventromedial Prefrontal Damage: Evidence from the Ultimatum Game. The Journal of Neuroscience, 27, 951-956 (2007)