Independent Agents and Ethics

Iordanis Kavathatzopoulos

Independent agents and ethics

Iordanis Kavathatzopoulos

Department of IT – VII, Uppsala University, Uppsala, Sweden

iordanis@it.uu.se

Abstract. 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. In the present paper it is discussed how ethical decision support pro-grams can be integrated into robots and other relatively independent decision making systems to secure that decisions are made according to the basic theories of philosophy and to the findings of psychological research.

Keywords: agents, automation, autonomy, decision making, ethics, moral, robots, systems

1 Challenges of new technology

Modern technology gives us the chance to automatize many tasks previously done under careful supervision by humans. Routine procedures have been able to be automatized since some time but now can technical systems accomplish many tasks in a relatively independent way. Accomplishment of important tasks independently and automatically has many advantages for persons, organizations and society. It delivers goods and services with higher quality, lower cost and in satisfying or increasing quantities. It helps us to reach important goals and solve many of our basic or special problems: prosperity, wealth, health, for example production of food, fight diseases and other problems that have tormented humanity before.

On the other hand, efficient technologies and automation may cause unemployment. New jobs are created because of the new technologies since this a process of societal and economic restructuring. Still the transition is painful, therefore responsible, innovative and mainly ethical policies by societies, organizations and persons have to be formulated and applied to alleviate any pain and to direct development toward preferred directions.

This is true on a general level. We can easily get aware of all the difficult ethical issues that may arise because of these changes. Particularly, in the area of independent agents and robots ethical concern is necessary. The development of Information Technology, systems, robots, etc., that are capable of processing information and

adfa, p. 1, 2011.

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. 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, capable of making ethical decisions, and able to make these decisions independently and autonomously.

2 Option and ethical decision making

The distinction between moral content or moral statement or moral behavior, and ethical process or way of thinking 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 this, people's attachment to the nor-mative 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 independent agents, since IT systems are actually making choices. 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 doubt is the goal rather than the achievement of a solution to the problem investigated. Reaching a state of no knowledge, that is, throw-ing 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. Ability to think in the right way is not easy and apparently has been supposed to be the privilege of the few able ones [1, 2, 3]. For that, certain skills are necessary, such as Aristotle's virtue of *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].

This 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 conditions 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].

3 Support for ethical competence

The focus of any measures to support the skill of ethical autonomy varies depending on the degree of independence of the system. In designing and using non-independent systems the focus cannot be on the system but exclusively on designers, users, operators and owners. They will develop the system and decide in detail how it will operate, so they need to have the ability to find the ethical solutions while they are thinking, having a dialog or negotiating with each other. In semi-independent systems the focus is again on designers, operators, users and owners like in non-independent systems. In addition to the system itself they have also to design an information-gathering, treating and communicating system which will be integrated into the agent. Its task will be to inform the operators about possible ethical conflicts and let them decide the proper action. However, the information has to be presented in such a way as to block heteronomy and promote autonomy in the thinking of operators. In the case of totally independent systems the focus of support for ethical autonomy should be on all parts involved, not only humans but also the agent itself since it will also have the responsibility of independent choice as well as an own basic interest to take care of (see Table 1).

It is important to keep in mind that measures to support ethical autonomy or philosophizing have to be applied anyway. Either on humans involved in designing and operating agent systems, or in the case of fully independent agents, on agents themselves as well. Fully independent agents, if they can exist, are ethical beings; they feel the need to make choices in order to fulfill their purpose of existence. Of course philosophizing and ethical autonomy are necessary not only for the design of systems or rules for information gathering and communicating, but also during the processes of interpreting rules or negotiating with other stakeholders.

4 Systems and robots as ethical agents

Ethical decision support programs [10, 11, 12] 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].

Table 1. Focus of training and support for ethical competence depending on the degree of independence of robots and on the conditions of searching for or applying an ethical solution.

Search or implementation	Degree of independence of robot		
	Non-independent	Semi-independent	Fully independent
Open search, person	Mental skills of humans	Mental skills of all humans & design of communication system	Mental skills of all humans and agents
Open search, group	Mental skills and group processes of humans	Mental and group skills & design of communication system	Mental and group skills of humans and agents; negotiation skills
Give or receive instructions	Mental skills or group processes of humans	Mental skills or group processes; rules for communication	Mental skills or group processes for all; strategic - democratic dialog
Give or receive answers, orders	Mental skills or group processes of humans	Mental skills or group processes; redesign of communication	Mental skills or group processes for all; risk for conflict

Another possibility is to integrate a support tool 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.

First 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, and that would certainly conflict with the will of a fully independent system.

5 Identifying and supporting an independent agent

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, 17]. It seems that a rational decision process requires uninterrupted connection to emotions [18]. 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 ultimate purpose that is an emotional base guiding 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?

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, i.e. philosophizing, 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.

If we have such systems we will need to direct our efforts toward them supporting their ethical decision making, like we do for humans. In this case it would be necessary to train and educate the fully independent agents in using ethical autonomy as well as to involve them in democratic dialog together with humans in searching for the right answers to relevant ethical problems.

References

1. Plato: Theaitetos. I. Zacharopoulos, Athens (1981)
2. Plato: Apology of Socrates. Kaktos, Athens (1992a)
3. Plato: The Republic. Kaktos, Athens (1992b)
4. Aristotle: Nicomachean Ethics. Papyros, Athens (1975)
5. Kant, I.: Grundläggning av Sedernas Metafysik. Daidalos, Stockholm (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. Laaksoharju, M.: Let us be Philosophers! Computerized Support for Ethical Decision Making. Uppsala University, Department of Information Technology, Uppsala (2010)
11. 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. Ionian University, Kerkyra, Greece (2009)

12. Thaler, R. H. and Sunstein, C. R.: Nudge: Improving Decisions about Health, Wealth and Happiness. Yale University Press, New Haven, CT (2008)
13. Kavathatzopoulos, I.: Philosophizing as a usability method. In: E. Buchanan, et al. (eds.), Ambiguous technologies: Philosophical issues, practical solutions, human nature, pp. 194-201. International Society of Ethics and Information Technology, Lisbon (2013)
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)
17. Damasio, A.: Descartes misstag. Natur och Kultur, Stockholm (2006)
18. Hume, D.: Treatise of Human Nature. Penguin, London (1985)