AI ETHICS

# Moral Agency in Artificial Intelligence: A Model for an Ongoing Discussion

*By Javier Camacho Ibáñez, Director of Ethical Sustainability, ICADE, Comillas Pontifical University and José-Luis Fernández-Fernández, Professor, ICADE, Comillas Pontifical University*

*How is artificial intelligence influencing the concept of moral agency? How is it potentially modifying the concept of moral responsibility? Who's responsible for what? In this module, Javier and José-Luis tackle these questions with the help of a simple moral agency model. Through a series of thought exercises, they argue that responsibility ought to rest solely with autonomous human actors, rather than with automatic machines. They make the case for responsibility as the core of ethics, and outline key questions that they feel will shape future ethical AI discussions.*

**Lecture Transcript**

**Ibáñez 0:12** In this module, we are going to try to answer the question related to moral agency of artificial intelligence. How is artificial intelligence influencing the moral agency, the concept of moral agency? How is it modifying, potentially modifying, the concept of moral responsibility? Who's responsible for what? We will try to present a very simplified model in order to illustrate these moral agency issues with some examples. At the end, we will present some ongoing questions for this ongoing discussion. The way the module has been structured is the following. We will have a short introduction to the topic, we will introduce the moral agency model. We will present you with some of the characteristics of the model and some examples. And finally, we will wrap up with some conclusions and ongoing questions.

**Fernández-Fernández 1:30** Philosophy always has been a task, a human task, very, very demanding. The origin of the philosophy is just the admiration. Wilkes tends to say, "thavmázo". "Thavmázo" is admiration. Kant in the 18th century told us two things: admire and respect the sky and the law, moral inside of me. What is the "thavmázo" of today? What is the admiration of today? There are a lot of things but one is very important - we are now in a context where technology can help us to change, to transform, to amplify fantastically, the possibilities of human beings. And this is not an easy task to understand. And above all, it is not an easy task to manage properly. So we need to think about what are the possibilities that we have with us with the help of technology, intelligence, artificial intelligence, algorithms, analytics. Could have for a long period of history, some learnings, some apprenticeships. We now know that not everything that is technically possible, is at the same time, equally, morally desirable. Not everything that we can do, we should do. This is an axiom. I can't prove this as a demonstration in mathematics. But this is axiomatic result of a long history and reflection about the history of human beings on earth. Not everything that is possible technologically is equally well, ergo morally. This is the question. What will we do with these possibilities that we have now with technology and technocratic war? And this is a question of moral philosophy. It is a question of ethics. It is a question of reflection. We have to stop and to think systematically, because we are facing a new reality, and it deserves us to try to conduct this for good.



Transcribed by https://otter.ai

Putting people in the center of the process and making a new situation where humanity can flourish, humanity. There are people who are dreaming of post humanism, no - why? Humanity is an ever better thing to achieve. And this is a question of ethics in the technology field. And this is a question of ethics in the artificial intelligence and in the algorithmic intelligence, and if you want, if this is even the ethics for machines. But this is of course, an analogical way of speaking, because are really, are truly, the machines moral subjects? This is the question. This is the philosophical question that we have to, if not answer, certainly at least to formulate and to think about.

**Fernández-Fernández 5:58** The question, the key question is, what compacts a subject and a moral subject? What are the traits? What are the characteristics of a subject that can be understood as a moral subject? Philosophy has taught us that human people are moral subjects, because they are people who have conscience. They know what are they doing and they can anticipate the results of their actions. So, they have conscience. They are people who is free, not absolutely free, but has freedom. Freedom to choose and the question is how to choose well. And the subject, moral subject, has a third characteristic that is the will. I am conscious that I can choose E or N, and I decide, my will, imposes me to decide for X. I do X, I am responsible for the consequences that can follow my election, my choice. This is the subject as a moral subject. And this is a question, very, very human. The question now is what is the moral agency of an artifact, a machine, an algorithm, a system? Are they responsible or not? Because they have consequences, their actions have positive and negative consequences. When the consequences are negative, are they responsible? Even a tsunami has bad consequences, but is the weather responsible for the bad consequences? Now, is responsible, is a moral subject, an artifact? The question is not so easy to answer. In fact, there are a lot of people who are writing and thinking and researching about machine morals, moral machines, ethics for machines. And so, we are now touching the point – this is a point the deserves to think about. So, in doing that we are now proposing a simple moral model. It is like in economics or whatever, it's a way of depicting a complex reality in order to be able to capture, to get the most clear thing and to try to understand it. Our model has, this is very simple, has two agents. We have an Agent A and Agent B. Agent A is a complete person, Agent B is a company, and A and B agents are interacting. But suppose that agent B is able to build an Agent Beta (ß). Beta is a construct built by the Agent B, and interacts. But the question is Agent Beta is able to learn faster than everyone in the world, is able to self direct, is able to be much, much more intelligent than even the persons behind the Agent B and the company that is building the artifact data and they are interacting with Agent A. In this relationship between Agent A (human), Agent B (company) and Agent Beta (this is an artifact) mediated by the technology and the artificial intelligence and by the algorithms, we have a new field of reflection. A new field to think about. This is a connection between the old questions of moral philosophy in the current time.

**Ibáñez 11:20** So as we have seen the simple model, there are some characteristics on the different agents and even in the Agent Beta. Agent A can be also defined as the trustor – they have to trust on Agent B and Agent Beta, on the organizations and the algorithm, the artificial intelligence system that they will be providing him or her with. So, this relationship of trust, of confidence, is at the center of many of the efforts for the last years in terms of building trustable artificial intelligence systems. It is also true that there is an interaction between Agent A and the Artifact Beta and Organization B. There's a dialogue, there's an interaction, and maybe the goals are not the same, maybe those goals are not completely aligned. Probably Organization B or Agent B is thinking about efficiency. It is thinking about goals. It is thinking about the specific target. It is thinking about maximizing resources, maybe profits. And Agent A is definitely trying to use the services to fulfill some needs, to match some expectations that they have created, and at the end is taking a certain risk. It is making, has got some expectations,


Transcribed by https://otter.ai

some assumption, assuming some width. As we have explained, Agent A and Agent B are autonomous agents, they are fully autonomous agents. However artifact, the artifact or our Agent Beta, is not autonomous. Because, well the plane is an automatic or auto-motivated, however since it doesn't gather all the three needed characteristics: freedom, consciousness and will. However, the fact is that within this interaction, there is some sort of mediation, some sort of transformation of the overall moral agency of Agents A and Agents B. We can see this in a number of examples. Some matter of fact, artificial intelligence has brought so much attention because it can amplify in a positive way or in a negative way, the traditional issues we have found in marketing, in finance, in human resources or in many other business related disciplines. Our suggestion is that we should not consider Agent Beta as a full moral agent, but as an analogic moral agent. What we have to take into consideration is the effect, the transformational effect, the amplification effect if you will, that this Agent Beta has on both Agent B and Agent A. This amplification doesn't necessarily have to be negative. We can see also some, some positive sides, some positive consequences of this. Let's take a look into some of them. Even though Agent Beta is not sympathetic, is not compassionate, you know, it cannot be considered a fully moral agent. It has a one clear positive impact on Agent A, and also should be on Agent B.

**Ibáñez 15:57** It helps both moral agents, all the moral agents involved, to expand their moral sensitivity. Let's take for example, the word "to bake" - this historic bias in language model. I'm sure that those biases, or, that bias is based on historical data, and we might have never been able to find it out if it wasn't because of artificial intelligence algorithms. So there is indeed the possibility that this Agent Beta (analogic moral agent) helps us to expand our awareness for the moral dimension of certain issues. It can also help in terms of ethical reasoning, or in help in terms of being able to discern which are consequences even from a totalitarian point of view. So, if we include certain systems in our ethical reasoning framework, you can help us, you know, expand our decision making capability as well. So, what we are trying to say is that moral agency for both Agent A and Agent B is influenced by the salient data. It can also amplify at some time, or it can modify, or it can transform the scope, the radar of the moral agency of both Agents A and B. Another example we can find related with for example, with discrimination or fairness. So, you find that we have a very simple credit loan algorithm in a bank that has been fed by historical data. And based on the data, someone, maybe Organization B, Agent B, or Agent A determined that the way we were given loans, credit loans, was not fair. Because, you know, in the past, there were like more men that got the credit, a more absolute number of men than a number of women. So Agent B with a clearly good intention has kind of a trick, the algorithm has applied certain fairness rules, to try to match that in a way that he's offering Agent A a possibility of giving the same number of credits through loans to men and women. In order to do that, as we as we can see, mainly because the number of, or the profile of women customers, have, for whatever reason, as an average are less income, we have to address different income levels, if that were the only previous criteria to use for men, for the group of men and the group of women. And therefore, as you can see, using this, we give the same number of credits to men and women. But is it fair? Because, probably there is a single person, a single man with higher income that has been denied a credit loan just because of these parity criteria. Of course, this is a very simple example. And there are many more complex ways of approaching it. But the important thing about this example is the influence that Agent B and Artifact Beta can have on the decisions taken by Agent A, in this case for example the bank, and how it might affect other people. And this is a very interesting example to illustrate the following concept. Even though the moral agency of Agents A and B might be altered, modified, amplified, transformed, mediated, the key question is that responsibility, which lies behind that moral agency, is not. So no matter if our moral sensitivity is increased, if our moral decisions are biased byArtifact B, our responsibility remains the same, as well as Agent B's responsibility. And that's a clear


Transcribed by https://otter.ai

distinction that we wanted, we wanted to make. Analogic moral agents might transform the moral agency, but their responsibility untouched. And related to that, there is a problem of, if you will, subcontracting that moral agency. So what if Agent A blindly trusts Agent B and the Artifact Beta in terms of moral agency, saying, “No, no, no, don't worry, I will, I will trust anything you do”. Well, the problem with that, and it's a problem that we are facing these days in certain applications, is that it will be okay to do that. However, they cannot neglect, our Agent A cannot reject, cannot neglect their responsibility on the actions and on their decisions. And there certainly are risks that we have to be aware of. So, that effect of blindly trusting, of trying to move the locus of moral responsibility, even though the moral agency might be transformed, is not is not the right thing to do.

**Fernández-Fernández 22:54** To sum up something of the idea that we have been dealing with in this short video, we have talked about two different kinds of agents: one moral agent - one person that is autonomous, autonomous is the person, the subject, that is able to give to himself or to herself the law auto-nomina, autonomous. And this is just for human beings, and we have automats, automatic agents, that need to receive their way of acting, of doing, from the builder. So, a machine is automatic and expands our possibilities tremendously as we can agree, expand our possibilities. But it is not the last responsible for the consequences. So, the responsibility is always something very human. The responsibility you can avoid. You can do this and you can amplify your possibilities of choice. You can amplify with the use of the artifacts, your ways of doing things in the world, but the responsibility is yours, not of the medium. You are the end, and the medium you use is a tool. We have the question of responsibility. The responsibility is yes, the basic, the core of ethics. It is the essence of your freedom, your ability to choose, you are responsible for what you are doing. And we can prepare for the future by asking in anticipation, some powerful questions.

**Ibáñez 25:19** The questions that we might ask ourselves and we might ask you, for with ongoing discussion, would be as follows: How could a non-human agent with the characteristics of Agent Beta be considered a true moral agent in the mid-term or in the long term? What is missing in order to consider Agent Beta as further more than an analogic moral agent? How is the trust relationship between Agents A and B modified because of the interposition of Agent Beta? How to protect the confidence, how to protect the trust, in the relationship between Agent A and Agent B, when Agent Beta is interposed? We hope you liked this module and we would like to hear your comments. You can find us in our email address or in our LinkedIn profile. Thank you very much.



Transcribed by https://otter.ai