Preferences and Ethical Priorities: Thinking Fast and Slow in AI

Francesca Rossi
IBM Research
Francesca.Rossi2@ibm.com

Andrea Loreggia
University of Padova
andrea.loreggia@gmail.com

ABSTRACT

In AI, the ability to model and reason with preferences allows for more personalized services. Ethical priorities are also essential, if we want AI systems to make decisions that are ethically acceptable. Both data-driven and symbolic methods can be used to model preferences and ethical priorities, and to combine them in the same system, as two agents that need to cooperate. We describe two approaches to design AI systems that can reason with both preferences and ethical priorities. We then generalize this setting to follow Kahneman's theory of thinking fast and slow in the human's mind. According to this theory, we make decision by employing and combining two very different systems: one accounts for intuition and immediate but imprecise actions, while the other one models correct and complex logical reasoning. We discuss how such two systems could possibly be exploited and adapted to design machines that allow for both data-driven and logical reasoning, and exhibit degrees of personalized and ethically acceptable behavior.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Theory of computation** → *Machine learning theory*; *Sequential decision making*; Reinforcement learning;

KEYWORDS

Multi-agent system; Knowledge Representation; Decision Theory

ACM Reference Format:
Francesca Rossi and Andrea Loreggia. 2019. Preferences and Ethical Priorities: Thinking Fast and Slow in AI. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 2 pages.

1 COMBINING PREFERENCES AND ETHICAL PRIORITIES

In our everyday life, preferences play a key role in whatever we do, and they are routinely collected, analyzed, and used to provide us with personalized services. When we make decision, however, we usually use both our preferences and additional constraints, priorities or ethical principles, in order to be compliant to some exogenous guidelines.

More often, AI agents are used to support humans in different decision scenarios. In some domains, they are also allowed to make decisions autonomously. It is therefore natural to question whether these systems are aligned with both subjective preferences and ethical values, so that they can make decisions in line with human behavior.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

AI researchers have studied for a long time how to represent preferences, or priorities of many kinds, in artificial agents, employing both symbolic and data-driven approaches.

Data-driven approaches learn autonomously from a set of available samples (dataset) how to behave in a particular setting. How to reach the given goal is inferred from the data. The performance and accuracy of these systems can be very high, but most of the times they lack in terms of explanationability and possibly adherence to existing normatives which are not represented in the data. On the other hand, symbolic methods allow for a structured representation of the domain and for a formal inference of the knowledge base. These systems can give explanation about the choice resulting from the reasoning process, by taking into account preferences or guidelines, but they lack in terms of scalability, performance, and flexibility.

Recently, some works have been devoted to design and implement AI systems driven by separate representations of preferences and ethical principles. The power of separating them is to be able to flexibly decide how to combine them, and to allow different agents to provide the two kinds of knowledge. In these studies, a common representation (being it a data-driven approach or a symbolic framework) for both objects (preferences and ethical priorities) is used in order to make a decision which follows them both.

For example, Balakrishnan et al. [1] shows that is possible to teach online agents to learn through reward feedback how to select the best action and at the same time to learn and follow a set of behavioral constraints. These learned constraints are used as a guideline when the agent makes a decisions in the online scenario. In such a way, agents are reactive to both reward feedback and learned behavioral constraints. In Loreggia et al. [4], instead, preferences and ethical principles are represented and used in a value alignment procedure in order to make a decision which is compliant with the ethical principles and at the same time close enough to the preference of the individual. The procedure uses a notion of distance [3] to compute a similarity score between the ethical principle and the preferences.

Both these lines of work use the same kind of techniques (reinforcement learning in one, and preference symbolic modelling in the other one) for modelling both preferences and ethical principles.

2 THINKING FAST AND SLOW IN AI

Generalizing from the above scenarios, one can think not only of keeping preferences and ethical principles separate, but also of handling them via different kinds of techniques. After all, this is what is done in our mind.

According the Daniel Kahneman's theory [2], our decision making is supported by the cooperation of two systems: System 1 provides intuitive, imprecise, fast, and at time unconscious decisions

(the so-called "thinking fast"), while System 2 handles more complex situations where logical and rational thinking is needed to reach a decision (the so-called "thinking slow").

System 1 is guided mainly by intuition rather than deliberation. It gives fast answers to very simple questions, but such answers are sometimes wrong and not always have an explanation. When the problem is too complex for System 1, System 2 kicks in and solves it with access to computational resources and logic rules. Sometimes a problem is new and difficult to solve, thus handled by System 2, but then its solutions over time are used to accumulate examples that System 1 can use readily with no effort. Thus after a while the problem can become manageable by System 1. A typical example is reading text in our own native language.

Now let's think of designing a multi-agent system with this decision making structure. System 1 has clearly the properties of machine learning approaches, whose training set is given by System 2. On the other hand, System 2 has the properties of symbolic and logical AI approaches. At the beginning of the machine's functioning, every decision would be computed by its System 2. After a while, however, in some cases System 2 collects enough knowledge and examples to allow System 1 to kick in.

Machines are of course different from human minds, for example in terms of memory and computational resources. Our brain in comparison is very limited, and this is the reason why some problems never pass from System 2 to System 1, no matter how much examples are collected by System 2. A typical example is counting the number of "a" in a page of text. This happens because our System 1 is limited by our memory and computational resources. Is this true also in a machine? Assuming we provide enough computational and memory support, would the machine's System 1 be able to kick in for all problems, after a while? Also, our System 1 waits for System 2 to provide the training set, but in a machine the System 1 could generate it itself by simulating decision making scenarios. Does this make the machine's System 2 less needed?

In a person's mind, System 2 is also in charge of monitoring and checking the ethical behavior of System 1, that would otherwise act out of only intuition and subjective preferences. If we follow the above analogy, we should model preference reasoning via machine learning ad ethical reasoning by symbolic logic frameworks, that over time build the training dataset for the preference reasoning module. Would this be the best way to combine preferences and ethical principles in multi-agent decision making artificial system?

These questions are at the core of current AI: is there a role for symbolic AI, or is data-driven AI enough? Do we need to combine machine learning with logical reasoning, or can we just focus only on machine learning approaches? In other words, can machine just employ fast thinking, or do they also need slow thinking?

We believe that Kahneman's theory can provide very valuable insights to significantly understand how to answer these very important questions for decision making, multi-agent systems, preferences, ethical principles, and AI in general.

3 BIOGRAPHY

Francesca Rossi

Francesca Rossi is the IBM AI Ethics Global Leader and a Distinguished Research Staff Member at IBM Research.

Her research interests focus on artificial intelligence, specifically they include constraint reasoning, preferences, multi-agent systems, computational social choice, and collective decision making. She is also interested in ethical issues in the development and behaviour of AI systems, in particular for decision support systems for group decision making. She is a fellow of both AAAI and EurAI. She has been president of IJCAI and the Editor in Chief of the Journal of AI Research. She is in the executive committee of the IEEE global initiative on ethical considerations on the development of autonomous and intelligent systems and she is a member of the board of directors of the Partnership on AI, where she represents IBM as one of the founding partners. She is a member of the European Commission High Level Expert Group on AI and the general chair of the AAAI 2020 conference.

Andrea Loreggia

Andrea Loreggia is a postdoc at the University of Padova (Italy) where he is working on a project funded by the Future of Life Institute. The research is aiming to study how to embed and reason with ethical principles in AI systems. He is involved in working groups for the ethical development of AI for an Italian foundation and IEEE. His research interests are focused on artificial intelligence spanning from computational social choice to deep learning. He was an intern at the IBM Research Yorktown Lab, where he developed a framework for algorithm portfolios based on machine learning techniques. The research has been patented by IBM in 2017. He received his doctorate in Computer Science and also the Laurea Magna cum Laude in Computer Science from the University of Padova.

REFERENCES

[1] A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi. 2019. Incorporating Behavioral Constraints in Online AI Systems. In *Proc. of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

[2] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York.

[3] A. Loreggia, N. Mattei, F. Rossi, and K. B. Venable. 2018. On the Distance Between CP-nets. In *Proc. of the 17th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[4] A. Loreggia, N. Mattei, F. Rossi, and K. B. Venable. 2018. Value Alignment Via Tractable Preference Distance. In *Artificial Intelligence Safety and Security*, R. V. Yampolskiy (Ed.). CRC Press, Chapter 18.