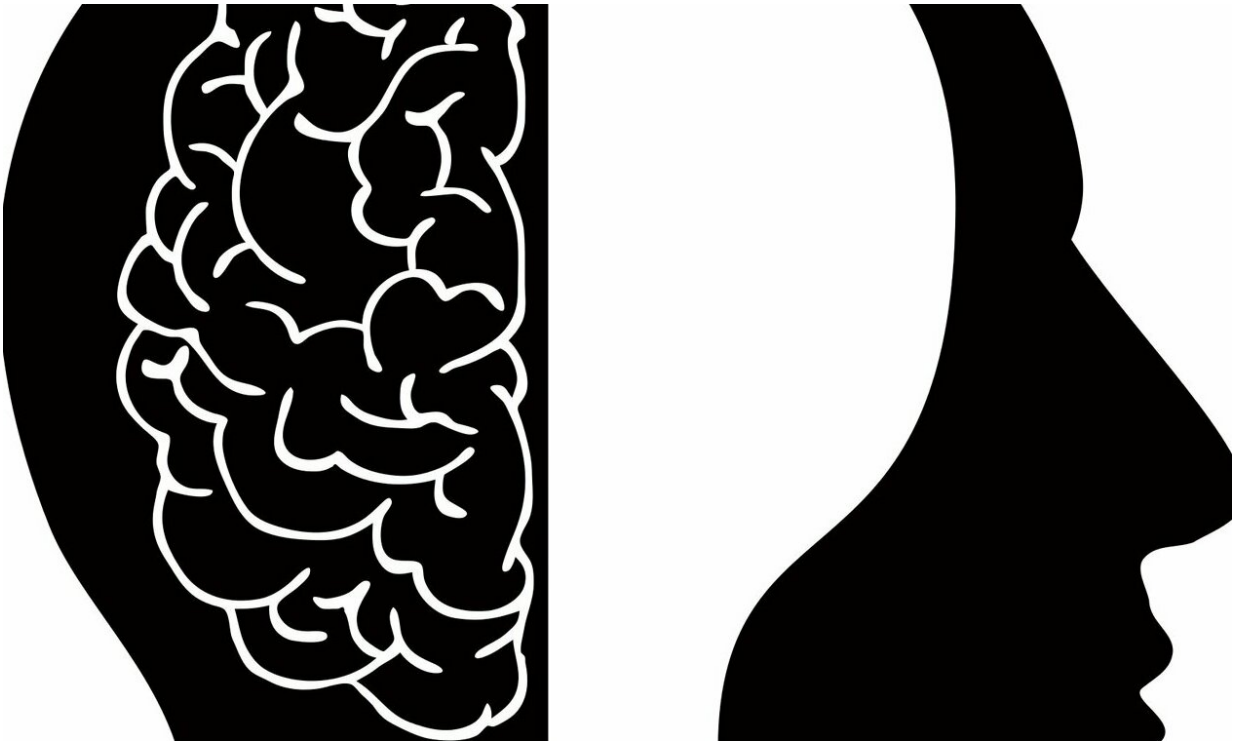
# Building ethically aligned AI

January 23 2019, by Francesca Rossi



Credit: CC0 Public Domain

The more AI agents are deployed in scenarios with possibly unexpected situations, the more they need to be flexible, adaptive, and creative in achieving their goals. Thus, a certain level of freedom to choose the best path to a specific goal is necessary in making AI robust and flexible enough to be deployed successfully in real-life scenarios.

This is especially true when AI systems tackle difficult problems whose solution cannot be accurately defined by a traditional rule-based approach but require the data-driven and/or learning approaches increasingly being used in AI. Indeed, data-driven AI systems, such as those using machine learning, are very successful in terms of accuracy and flexibility, and they can be very "creative" in solving a problem, finding solutions that could positively surprise humans and teach them innovative ways to resolve a challenge.

However, creativity and freedom without boundaries can sometimes lead to undesired actions: the AI system could achieve its goal in ways that are not considered acceptable according to values and norms of the impacted community. Thus, there is a growing need to understand how to constrain the actions of an AI system by providing boundaries within which the system must operate. This is usually referred to as the "value alignment" problem, since such boundaries should model values and principles required for the specific AI application scenario.

At IBM Research, we have studied and assessed two ways to align AI systems to ethical principles:

- The first uses the same formalism to model and combine subjective preferences (to achieve service personalization) and ethical priorities (to achieve value alignment). A notion of distance between preferences and ethical priorities is used to decide if actions can be determined just by the preferences or if we need to consider additional ethical priorities, when the preferences are too divergent from these priorities.
- The second employs a reinforcement learning approach (within the bandit problem setting) for reward maximization and learns the ethical guidelines from positive and negative examples. We tested this approach on movie recommendations with parental guidance, as well as drug dosage selection with quality of life

considerations.

The paper that describes our overall approach and the two possible ways to solve the value alignment problem is going to be presented at the upcoming AAAI 2019 conference and will receive the AAAI 2019 Blue Sky Idea award. It can be found here.

This work is part of a long-term effort to understand how to embed ethical principles into AI systems in collaboration with MIT. While the research done in and models ethical priorities as deontologic constraints, the IBM-MIT team is currently gathering human preferences data to model how humans follow, and switch between, different ethical theories (such as utilitarian, deontologic, and contractualist), in order to then engineer both ethical theories and switching mechanisms, suitably adapted, into AI systems. In this way, such systems will be able to be better aligned to the way people reason and act upon ethics while making decisions, and thus will be better equipped to naturally and compactly interact with humans in an augmented intelligence approach to AI.

**More information:** "Building Ethically Bounded AI," Francesca Rossi and Nicholas Mattei, to appear in Proceedings of AAAI 2019, senior member presentation track, Blue Sky idea award paper.

"Incorporating Behavioral Constraints in Online AI Systems," Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, Francesca Rossi, to appear in Proceedings of AAAI 2019.

"On the Distance Between CP-nets," Andrea Loreggia, Nicholas Mattei, Francesca Rossi, K. Brent Venable. In Proc. AAMAS 2018, Stockholm, July 2018.

here.

Provided by IBM