Report of "Think Fast and Slow in AI", by Francesca ROSSI

Joris LIMONIER

November 17, 2021

Francesca ROSSI is an Italian IBM researcher focusing on ethics in Artificial Intelligence (AI). Her talk "Think Fast and Slow in AI" is named after the book *Thinking, Fast and Slow*, by Daniel KAHNEMAN. This book, as well as *Sapiens* (HARARI), *Society of Mind* (MINSKY) and *Rethinking Consciousness* (GRAZIANO) are four pillars of the theory about Human minds. The framework used in this field dates back to Descartes who separated, in the 17th century, intuitive and conscious reasoning. However, it was brought back to life by Posner and Synder in 1975 and was reformulated to a more recent description by Kahneman, who introduced the so-called System 1 ($\mathcal{S}_1$)and System 2 ($\mathcal{S}_2$).

$\mathcal{S}_1$ is fast, unconscious, automatic, error prone and is responsible for everyday decisions. $\mathcal{S}_2$ on the other hand is slow, conscious, effortful, reliable and responsible for complex decisions. It may appear that these systems work in opposite manners, but they are actually complementary, as we shall show below. $\mathcal{S}_1$ quickly reacts in a local and parrallel fashion, almost like a reflex, whereas $\mathcal{S}_2$ takes time to analyze and consider its environment before performing a more accurate response. It is important to understand that the border between the two systems is blurred. Some actions may initially be managed by $\mathcal{S}_1$, before moving onto $\mathcal{S}_2$ once they become more natural, more intuitive. Complex arithmetic operations for example require a lot of cognitive intensity at first, but may become routine given enough practice.

The link with AI is now straightforward, one may easily imagine a multitude of problems where a quick decision is needed, even suboptimal, before performing a more accurate decision. This however leads to several questions: Do we always need a more accurate decision from $\mathcal{S}_2$? Who should decide whether such a decision is required? Is the use of $\mathcal{S}_2$ worth it given the associated costs? The key to those questions lays in meta-cognitive agents. They take as input the world (problems, actions, environment), the two systems (past decisions, rewards, cost) and other considerations such as knowledge and beliefs about other agents. Then they should assess available resources, the expected cost of using $\mathcal{S}_2$ and the expected reward associated with getting a correct solution. Their (broad) role is to determine how to best use $\mathcal{S}_1$ and $\mathcal{S}_2$. To perform this task, meta-cognitive agents take a two-phase approach. In phase one, the meta-cognitive agent ($\mathcal{MC}_1$) asserts whether there are enough resources to perform both $\mathcal{S}_1$ and $\mathcal{S}_2$. If not, $\mathcal{S}_1$ is chosen. Then $\mathcal{MC}_1$ determines the confidence of $\mathcal{S}_1$. If it is higher than the expected reward, then $\mathcal{S}_1$ is chosen. The second meta-cognitive agent ($\mathcal{MC}_2$) then comes into play. It computes the gain of using $\mathcal{S}_2$ (reward minus cost) and compares it to the expected reward of $\mathcal{S}_1$. If the former is greater than the latter, $\mathcal{S}_2$ should be used, otherwise $\mathcal{S}_1$ suffices.

Meta-cognitive agents are also responsible for a major topic in AI: trustworthiness. They can take into consideration fairness, robustness, explainability before choosing whether to apply $\mathcal{S}_2$. In summary, they must be logic-based so that a reasoning can be extracted from them. The clearer this reasoning, the better.

What we presented is actually a simplified model and more sophisticated ones exist, *e.g.* where there are multiple $\mathcal{S}_1$'s and $\mathcal{S}_2$'s. Some more sophisticated models are part of Multi-alternative Decision Field Theory ($\mathcal{MDFT}$). As the name hints, $\mathcal{MDFT}$ considers multiple options as an input and outputs one of these options, ideally the best one. This model is weighted over attributes, where the weights are updated according to the similarity between multiple options. Eventually, either when the maximum number of iterations is reached, or when a given value of a quality metric is obtained, the algorithm should stop.

The $\mathcal{S}_1$ - $\mathcal{S}_2$ framework is an interesting framework, not only because it works in practice, but also because it models the way of thinking of the human mind. Both this framework and its more sophisticated counterpart, the $\mathcal{MDFT}$, arise from natural concepts in cognitive theory. Another field inspired by cognitive and neural theory is deep learning. Merging the $\mathcal{S}_1$ - $\mathcal{S}_2$ framework with deep learning seems to be an interesting idea. Maybe it could lead to more explainability in the deep learning so-called "black-box", and therefore to improvements with respect to fairness in AI?