TFTS

Joris LIMONIER

November 17, 2021

- AI lacks capabilities that humans have. See how humans can help and bring them in machine world
- Focus on human capabilities that machines do not have, exploit cognitive theories of human decision making, identify causes enabling such capabilities, [fourth point]
- Four books:

  TFTS

  Sapiens: historian point of view

  Society of mind: how we interact collectively

  theory of consciousness: how humans model the world

- System 1, 2 = TF, TS (respectively)

  TF: automatic, fast thinking, error prone, unconscious bias, used when problem solved is easy, reactive mode, make quick decisions

  TS: slow, deliberate decision making, effortful, requires all attention, not many parrallel decisions, used when problem is cognitively difficult or we really care to solve it correctly (because many things depend on the problem). Sometimes override TF.

  We usually go from system 2 (TS) to system 1 (TF). Example: with child, start with TS, then go to TF when they can read. But not always: arithmetic operations.

- TFTS inspires AI. AAAI 2020 panel, CERN SPARK podcast of Francesca.
- Different approach

  Multi-agent architecture:

  TF solvers: rely on past experience, don't look at caracteristics of the problem / similar problems and what we know about the environment. Very fast, react (activate automatically) when new problems arise if their skills are relevant for the problem to be solved. Propose solution and assert confidence.

TS solvers: Don't rely too much on past experience, even though they have access to it. Computational complexity can be slow depending on size of input. Activate by meta-cognition.

Model solvers / updaters: Act in the backgorund to update model used by agents to do their job.

- Matacognition: cognition about cognition. Improves the quality of the system's decision. Choice of author is to use a centralized meta-cognitive agent. Assert TF vs TS.

- S1 solvers do not wait to be asked to start solving a problem + give confidence. Model solvers assert quality of S1 agent and decide if activation of S2 is required. Model solvers assesses available resources, expected cost of using S2, expected reward for correct solution and solvers. 2 phases: quick assesment, then more reliable (longer one) if needed.

- Two metacognitive phases:

  Goal: avoid using S2 when unnecessary (i.e. expected reward - expected code > what S1 can provide)

  2 phases: 1) check if enough resources for S1 and S2, if not, choose TF

- Design choices:

  S1 by default

  S2 may not be better than S1

  In more complex scenarios: there are several S1 and S2. MC need to choose among them

  AI trustworthiness: take into consideration fairness, robustness, explainability, ...etc and MC must be explainable and logic-based.

- SOFAI vs neuro-symbolic

  We do not assume that S1 are data-driven and S2 are logic-based.

- Human reasoning

  Human deliberation (Multi dimensional theory: MDFT)

  Results: learn the model of the world fril human deomonstrations, comparisaon of MDFT and RL

- MDFT: model how people make decisions about a set of options by choosing an option based on their preferrence and accumulation depending on discount factors and option similarity. Then stop criteria (number of iterations (ie time) or satisfaction with solution). Stopping times are used (with time and upper bound)

- MDFT vs RL: Machine is given a state and has to build a path towards a goal. There are penalties for violating constraints, for going for state where the agent is not supposed to go.
- S1 uses probability distribution based on history. S2 uses MDFT. MC decides between move proposed by S1 with confidence level and activating and MDFT (S2).
- Initially, system has no S1 and only S2, then at some point there is an option to use S1.