Nemmeno le persone reali capiscono le altre figuriamoci sta roba. Potrebbero capire e prevedere un comportamento ma non sempre e soprattutto troppe variabili da tenere conto.

Neanche noi ci conosciamo

Ma poi una macchina che generalizza come può adeguarsi alla variabilità degli esseri umani? Ci sono assassini e persone tranquillissime, hanno optmal policies diversissime.

Si parla di un box in più occasioni e poi lo si mette 4 pagine dopo.

Inglese pessimo: If a cooperative AI is to explain human behavior, then it must do so in terms of a model of these components of the problem and a derived optimal policy.

Come si fa a capire in base a cosa sono misinterpretati? Magari ha ragione Kahneman, manco lo spiega perchè: In cognitive science,it has been shown that explanations for some puzzling aspects of human behavior have some times been misattributed amongs these three components of the bounded optimization problem. Kahneman for example, attributes a number of human behaviors to distorted utility functions when these should, instead, be attributed to the structure of the task environment.

Punti

Computationally Rational model: explains human behavior by establishing a causal connection between subjective utility, capacities, and experience on the one-hand and behavior on the other. In other words, cooperative AI needs a computational model of the human that is defined in terms of an objective function representing human subjective utility, a model of human capacities (memory, perception, reasoning, etc.), and a model of how humans experience the task environment over a lifetime. In addition, to be utilized in cooperative AI, the general human model needs to be “fit” to each individual collaborator. People vary widely along all three of the key dimensions. For example, in subjective utility, while most people are risk-averse, there is considerable variation (Mata et al. 2018), with some individuals even being risk seeking in the domain of gains. Un modello deve sapere ste preferenze, sta per avere a che fare con un sottone o un giga-chad?

While admitting to the possibility of variation in subjective weightings, computational rationality commits to a normative model of subjective utility, which is a requirement of the rationality assumption. Rather than appealing to “biases” in choice, a computationally rational account explains behavior as an emergent consequence of resource limits rather than as a consequence of “irrational” policies

In many, though not all, computationally rational models, the optimal policy is generated by defining the bounded optimality problem as a Markovian decision problem (MDP) and then solving it using classical planning algorithms. All of these problems share the fact that they can be used to precisely capture a sequential bounded optimalisation problem in which goals are expressed as cumulative reward maximization

Standard use di Markovian: agente interagisce con un ambiente esterno.

Computally rational model alternativi: agente interagisce con un ambiente interno con osservazioni e azioni con uno yoked(aggiogato) external environment via stimuli e responses.

In addition, people can make strategic choices about which task to prioritize and whether to pri oritize speed or accuracy. The fact that task performanceis influenced bybothcapacity limits and strategic concerns makes it difficult for an observer to tease out the contribution of each

La gente crede che gli altri siano razionali. Considerazione poco sensata.

Humans also vary in capabilities and experience of the task environment.

Simon pointed out that it is impossible to understand human behavior without understanding that it is adapted to both the structure of the environment and cognitive limitations.

It’s impossible to understand human behavior without taking utility, capabilities, and experienced environment into account simultaneously.

The model is fit to individuals via parameter inference, which is carried out in Equation (2) and results in a set of parameters 𝜃∗ that govern the functioning of the model.

Computationally rational models achieve this by finding a policy that maximizes utility under bounds.

What computational rationality adds is that predictions are made by calculating the bounded optimal policy in a given environment.

Computationally rational models can explain why behavior differs among individuals and conditions, as opposed to just describing them.

Others researchers have gone as far as arguing that progress in AI will curtail without better use of models

Arguably, the latent states and processes that are characteristic of human cog nition are not captured easily by model-free methods that learn from patterns in observed behavior.

One core insight has been that the adaptive problems faced by people can be defined as reinforcement learning (RL) problems and solved accordingly.

The assumption that human behavior is boundedly optimal is controversial in cognitive science. However, there is evidence that people appear optimal when the bounds imposed by computation or environment are taken into account

For the purposes of cooperative AI, the question is not whether people are able to optimize behavior,but whether behavior can be predicted by an optimization algorithm.

Computationally, rational models view cognition as a bounded optimization problem

Bounded optimality is distinguished from bounded rationality, where, in the latter, the role of optimization in understanding behavior is rejected.

An agent is bounded optimal if its program is a solution to the optimization problem presented by its utility function (objective), capacity limits (architecture), and the task environment.

When a model of a human and the environment defines a bounded optimality problem, the solution is a program 𝑃∗ , a policy 𝜋∗.The program simulates the user’s strategy, which determines the choice of actions. 𝑃∗ is adapted through optimization to the bounds imposed by the tuple, possibly using machine learning. This assumption is highlighted by stating that the optimal policy 𝜋∗ is a function of the agent’s subjective utility, capacity limits, and experience of the environment (its history).

Computational rationality is a departure from the earlier idea that human behavior can be predicted by considering only the environment, and, “not the assumed structure of the mind”. It is also a departure from the idea that behavior is shaped to external rewards. Internal bounds matter.