Reward is Enough

David Silver, Satinder Singh, Doina Precup, Richard S. Sutton

Paper Review by Hongzheng Wang

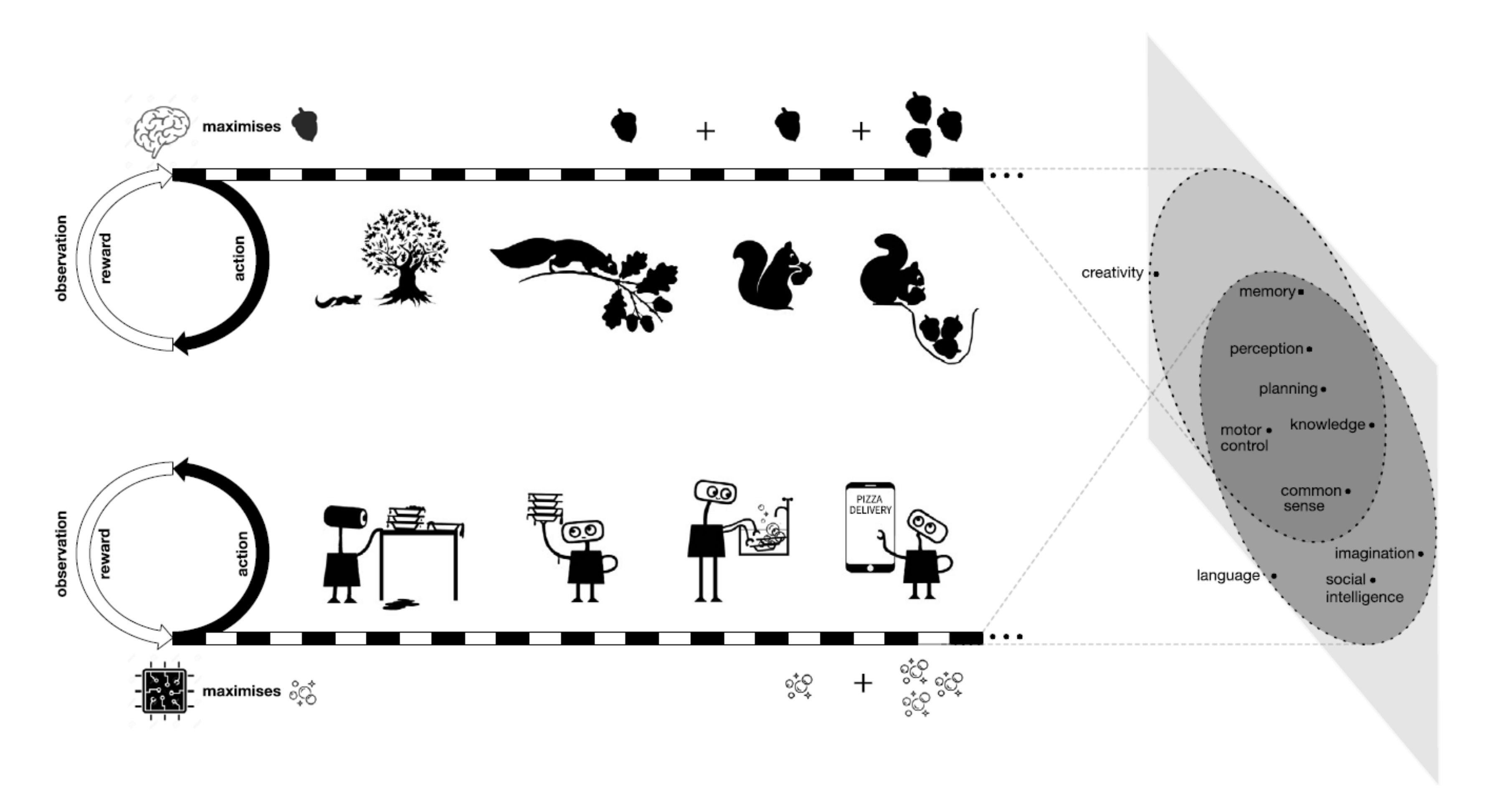
Hypothesis

- Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment.
  - Accordingly, reward is enough to drive behaviour that exhibits abilities studied in natural and artificial intelligence, including knowledge, learning, perception, social intelligence, language, generalisation and imitation.
  - Agents that learn through trial and error experience to maximise reward could learn behaviour that exhibits most if not all of these abilities, and therefore that powerful reinforcement learning agents could constitute a solution to artificial general intelligence.

Hypothesis

- The environments are so complex and require sophisticated abilities.
  - Maximising reward demands a variety of abilities associated with intelligence.
- In such environments, any behaviour that maximises reward must necessarily exhibit those abilities.
  - In this sense, the generic objective of reward maximisation contains within it many or possibly even all the goals of intelligence.

Example: Squirrel and Kitchen Robot



Reward is Enough

- Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment.

- If it is true, it suggests that a good reward-maximising agent, in the service of achieving its goal, could implicitly yield abilities associated with intelligence.
  - A good agent in this context is one that successfully – perhaps using as-yet-undiscovered algorithms – performs well at maximising cumulative reward in its environment.

Reward is enough for knowledge and learning

- Knowledge is defined as information that is internal to the agent.
  - Knowledge may be contained within the parameters of an agent's functions for selecting actions, predicting cumulative reward, or predicting features of future observations.
  - Some of this knowledge may be innate (prior knowledge), while some knowledge may be acquired through learning.
  - If the sum of potential knowledge outstrips the agent's capacity, knowledge must be a function of the agent's experience and adapt to the agent's particular circumstances – thus demanding learning.
- The environment may call for both innate and learned knowledge, and a reward-maximising agent will, whenever required, contain the former (for example, through evolution in natural agents and by design in artificial agents) and acquire the latter (through learning).

Reward is enough for perception

- The human world demands a variety of perceptual abilities to accumulate rewards.
  - Several modes of perception may be required, including visual, aural, olfactory, somatosensory or proprioceptive perception.
- Perception may instead be understood as subserving the maximisation of reward.
  - Perception in some animals has been shown to be consistent with reward maximisation.
  - Considering perception from the perspective of reward maximisation rather than supervised learning may ultimately support a greater range of perceptual behaviours, including challenging and realistic forms of perceptual abilities.
    - Action and observation are typically intertwined into active forms of perception
    - The utility of perception often depends upon the agent’s behaviour
    - There may be an explicit or implicit cost to acquiring information
    - The distribution of data is typically context-dependent
    - Many applications of perception do not have access to labelled data

Reward is enough for social intelligence

- Social intelligence is the ability to understand and interact effectively with other agents.
- Social intelligence may instead be understood as, and implemented by, maximising cumulative reward from the point of view of one agent in an environment that contains other agents.
  - An agent that can anticipate and influence the behaviour of other agents can typically achieve greater cumulative reward.
  - Thus, if an environment needs social intelligence (e.g. because it contains animals or humans), reward maximisation will produce social intelligence.
  - The behaviour of other agents may depend on the agent's past interactions, just like other aspects of the environment.
- Reward maximisation may in fact lead to a better solution than an equilibrium. This is because it may capitalise upon suboptimal behaviours of other agents, rather than assuming optimal or worst-case behaviour. Furthermore, reward maximisation has a unique optimal value, while the equilibrium value is non-unique in general-sum games.

Reward is enough for language

- Language modelling by itself may not be sufficient to produce a broader set of linguistic abilities associated with intelligence:
  - Language may be intertwined with other modalities of action and observation.
  - Language is consequential and purposeful. Language utterances have a consequence in the environment.
  - The utility of language varies according to the agent’s situation and behaviour.
  - In rich environments the potential uses of language to deal with unforeseen events may outstrip the capacity of any corpus.
- The ability of language in its full richness, including all these broader abilities, arises from the pursuit of reward.
  - It is an instance of an agent’s ability to produce complex sequences of actions based on complex sequences of observations in order to influence other agents in the environment and accumulate greater reward.

Reward is enough for generalisation

- Generalisation is often defined as the ability to transfer the solution to one problem into the solution to another problem.
- Generalisation may instead be understood as, and implemented by, maximising cumulative reward in a continuing stream of interaction between an agent and a single complex environment.

- Environments such as the human world demand generalisation simply because the agent encounters different aspects of the environment at different times.
  - The differing states in such environment combine a variety of elements that overlap and recur at different time-scales.
  - Rich environments demand the ability to generalise from past states to future states – with all these associated complexities – in order to efficiently accumulate rewards.

Reward is enough for imitation

- Imitation is an important ability associated with human and animal intelligence, which may facilitate the rapid acquisition of other abilities, such as language, knowledge, and motor skills.
- A much broader and realistic class of observational learning abilities, compared to direct imitation through behavioural cloning, may be demanded in complex environments.
  - Other agents may be an integral part of the agent’s environment.
  - Other agents may demonstrate undesirable behaviours that should be avoided.
  - Observational learning may even occur without any explicit agency.
- These broader abilities of observation learning could be driven by the maximisation of reward, from the perspective of a single agent that simply observes other agents as integral parts of its environment.
  - Observational learning by reinforcement learning

Reward is enough for general intelligence

- General intelligence may be defined as the ability to flexibly achieve a variety of goals in different contexts.
- General intelligence can instead be understood as, and implemented by, maximising a singular reward in a single, complex environment.
  - An animal's stream of experience is sufficiently rich and varied that it may demand a flexible ability to achieve a vast variety of subgoals (such as foraging, fighting, or fleeing), in order to succeed in maximising its overall reward (such as hunger or reproduction).
  - Similarly, if an artificial agent's stream of experience is sufficiently rich, then singular goals (such as battery-life or survival) may implicitly require the ability to achieve an equally wide variety of subgoals, and the maximisation of reward should therefore be enough to yield an artificial general intelligence.

Reinforcement learning agents

- How to construct an agent that maximises reward:
  - This question may also be answered by reward maximisation.
- Specifically, we consider agents with a general ability to learn how to maximise reward from their ongoing experience of interacting with the environment. Called *reinforcement learning agents*.
  - Among all possible solution methods for maximising reward, surely the most natural approach is to learn to do so from experience, by interacting with the environment.
  - To achieve high reward, the agent must therefore be equipped with a general ability to fully and continually adapt its behaviour to new experiences. Indeed, reinforcement learning agents may be the only feasible solutions in such complex environments.
  - If an agent can continually adjust its behaviour so as to improve its cumulative reward, then any abilities that are repeatedly demanded by its environment must ultimately be produced in the agent's behaviour.

Reinforcement learning agents

- There is not any theoretical guarantee on the sample efficiency of reinforcement learning agents.
- Powerful reinforcement learning agents, when placed in complex environments, will in practice give rise to sophisticated expressions of intelligence.
  - If this conjecture is correct, it offers a complete pathway towards the implementation of artificial general intelligence.
- Several recent examples of reinforcement learning agents, endowed with an ability to learn to maximise rewards, have given rise to broadly capable behaviours that exceeded expectations.
  - e.g., AlphaZero in the game of Go and Chess.

Discussion

- The basic idea is acceptable.
  - Advantages of reinforcement learning
  - Learning from experiment
  - Interaction with environment
- Lack of detail.
  - How to build such agent? How to design the reward function?
    - Does it even exist?
  - (Joke on forum: walking is enough to reach the top of the Everest)
- Somehow like the process of evolution.

Thank You!