Generality, Capability and Control

John Burden and José Hernández-Orallo

2020
Artificial General Intelligence

- AI Systems that are very capable over a broad range of tasks.
- "Superintelligence" — Many proposed risks and associated challenges [1]
Capability

- Capability is how good an AI system is at the tasks for which it was designed.
- Easy to evaluate if the task has a performance metric.
- Difficult to compare capabilities across different tasks.
Generality

- Generality is a measure of the range and types of domain an AI system can perform well in.
- Difficult to evaluate — how are different types of tasks compared? How well does an agent need to perform?
Agent Characteristic Curves

- IRT incorporates the item difficulty into the score.
- Agent Characteristic Curves map item difficulty to a probability of success.

Figure: An ACC showing performance $\Psi$ of a Q-learning agent over environments of increasing difficulty ($h$, on the x-axis). Data from [3].
Agent Characteristic Curves (2)

- The area under the ACC can correspond to the agent’s capability.
- The gradient of the ACC can correspond to the agent’s generality.

Figure: An ACC showing performance $\Psi$ of a Q-learning agent over environments of increasing difficulty ($h$, on the $x$-axis). Data from [3].
Control

- Agent control is the reliability and deliberate intent of an agent’s actions and decisions.
- Often control is required with respect to a task property (such as goal completion or avoiding unsafe behaviour)
- Control is opposite to the expected entropy of visited states conditioned over the behaviour to control.
- Control is also related to the dispersion of data-points along an ACC.
Risk

- Risk can be thought of as a measure of the likeliness and severity of “bad” events occurring.
- Incompetent AI systems (lacking in generality, capability or control) can be very risky.
- However extremely competent AI systems can also be very risky.
Entanglement

- Generality, Capability and Control can all independently affect the level of risk posed by an AI system.
- However they can also affect each other.
- How we can untangle these factors is currently unclear — but they can certainly be measured.
- These measurements could give insight into a system’s behaviours and help to mitigate risk.
Consider a class of environments where an agent must identify and move to a goal avoiding traps along the way.

In the paper we demonstrate how we could calculate the agent’s capability, generality, control and risk.

We show that our notion of control (when aligned with safety) does reduce the expected risk.
More work is needed to characterise and disseminate the dimensions of agents’ “intelligence”.

We want to extend these types of analysis to more complex agents and domains.

Verify these types of analysis with empirical results.
References

*Superintelligence: Paths, dangers, strategies.*  
Oxford University Press, 2014.

*Item response theory for psychologists.*  

*Comparing humans and AI agents.*  