Concept learning for safe autonomous AI

Kaj Sotala
Machine Intelligence Research Institute

The problem

How do we instill ethical concepts lacking a precise definition into an AI system?

Sophisticated AI systems may need to make decisions without human guidance, based on guidelines that have been provided beforehand but which need to be applied in novel or unanticipated situations. The guidelines may involve constraints that are difficult to define rigorously, such as:

- well-being
- rights
- due diligence
- reasonable doubt
- proportionate force

Non-moral concepts may also be important to get right, for instructing an AI to e.g. not affect an area beyond its physical location (Armstrong, Sandberg & Bostrom 2012).

A possible solution

Research has identified some mechanisms behind human concept generation. Further research in this field may allow us to build an AI that uses similar mechanisms to learn human-like concepts.

- Languages cannot be too complex, but also need to precisely communicate different concepts. Concepts in human languages tend to achieve near-optimal tradeoffs between these constraints (Regier, Kemp & Key in press).
- Kemp & Tenenbaum's (2008) structure learning approach produces similar classifications as humans do in physical, biological, and social domains.
- Probabilistic concept learning may explain how children easily learn that "horse" refers to all horses as opposed to e.g. "all animals" or "all horses except Clydesdales" (Tenenbaum 2011)

Challenges

Unfortunately, human concepts may not be straightforward to derive from logical principles.

Embodied cognition. Many of our concepts are rooted in bodily metaphors, and concepts may be formed by combining modality-specific features of different categories (Niedenthal et al. 2005), requiring an AI to have human-like sensory modalities to learn similar categories.

Evolutionary vestiges. Some concepts may exist for specific evolutionarily purposes: e.g. human reasoning may employ specialized modules evolved for tasks such as detecting cheaters (Cosmiders and Tooby 1992).

Social learning. Humans do not learn concepts in isolation, but rather in a social environment with rich feedback on which concepts are worth focusing on.

Comparing concepts

There needs to be some way of verifying that an AI's conceptual representation matches the human one. One way would be a common format into which human and AI concepts could be mapped.

Gärdenfors (2000) proposes a general theory of representation, Honkela et al. (2010) discuss comparing differences between human individuals, brain studies explore human concept geometry (Kriegeskorte & Kievit 2013).

Armstrong, S.; Sandberg, A.; and Bostrom, N. 2012. Thinking Inside the Box: Controlling and Using an Oracle AI.
Cosmides L and Tooby, J. 1992. Cognitive adaptations for social exchange.
Gärdenfors, P. 2000 Conceptual Spaces: The Geometry of Thought.
Honkela, T.; Janasik, N.; Lagus, K.; Lindh-Knuutila, T.; Pantzar, M.; Raitio, J. 2010. GICA: Grounded Intersubjective Concept Analysis.
Kemp, C. and Tenenbaum, J. B. 2008. The discovery of structural form.
Kriegeskorte, N. and Kievit, R.A. 2013. Representational geometry: integrating cognition, computation, and the brain.
Niedenthal, P.M.; Barsalou, L.W.; Winkielman, P.; Krauth-Gruber, S.; and Ric, F. 2005. Embodiment in Attitudes, Social Perception, and Emotion.
Regier, T.; Kemp, C.; and Kay, P. In press. Word meanings across languages support efficient communication
Tenenbaum, J.B.; Kemp, C.; Griffiths, T.L.; and Goodman, N.D. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction.