Concept learning for safe autonomous AI

Kaj Sotala
Machine Intelligence Research Institute
kaj.sotala@intelligence.org

Abstract

Sophisticated autonomous AI may need to base its behavior on fuzzy concepts such as well-being or rights. These concepts cannot be given an explicit formal definition, but obtaining desired behavior still requires a way to instill the concepts in an AI system. To solve the problem, we review evidence suggesting that the human brain generates its concepts using a relatively limited set of rules and mechanisms. This suggests that it might be feasible to build AI systems that use similar criteria for generating their own concepts, and could thus learn similar concepts as humans do. Major challenges to this approach include the embodied nature of human thought, evolutionary vestiges in cognition, the social nature of concepts, and the need to compare conceptual representations between humans and AI systems.

Introduction

Autonomous AI systems need to make decisions without immediate human guidance, according to instructions, goals and guidelines that have been provided beforehand but which may need to be applied in novel and unanticipated situations. Future systems may be intended to deal with complicated situations involving constraints that cannot be specified rigorously, but instead involve inherently fuzzy human concepts such as well-being, rights, due diligence, reasonable doubt, or, in military or law enforcement contexts, proportionate force.

For the AI system to behave correctly in such situations, its understanding of the concepts involved has to closely resemble that of the humans that gave it its instructions. An AI that is applying its instructions to a novel situation will only display desired behavior if the AI has an accurate understanding of the motivations behind the instructions.

Similarly, if the system is to qualify as an Artificial Moral Agent (Allen, Varner, and Zinser 2000), capable of determining the morally correct behavior in a given situation, it needs to have some level of understanding of morally relevant concepts, such as the previously mentioned concepts of rights and well-being.

The need for the AI's concepts to match those of its human designers increases the more autonomous and powerful the AI is. The most extreme case is that of a superintelligent AI, which may become powerful enough to be immune to human attempts to shut it down (Yudkowsky 2008, Chalmers 2010, Bostrom 2014).

This need has been highlighted in some previous work. For example, Armstrong, Sandberg, and Bostrom (2012) discuss the possibility of limiting an AI's behavior via rule-based motivational control, such as forbidding it from leaving a boxed area. They argue that this is a harder task than it might initially seem, for it requires rigorously defining intuitive concepts such as "this lead box here" and "the AI's physical location".

Armstrong, Sandberg, and Bostrom (2012) mention the possibility of an AI internalizing complex concepts through feedback, learning them in a way that resembles that of human children. Similar suggestions have been made by a number of other authors (Allen, Varner, and Zinser 2000, Guarini 2006, Goertzel & Bugaj 2008, Wallach and Allen 2009, Goertzel and Pitt 2012; for a review, see Sotala and Yampolskiy 2013). For example, Goertzel and Pitt (2012) and Goertzel and Bugaj (2008) argue that human ethical judgement relies on the combination of a number of human faculties, and that an AI system could be made to learn ethical behavior if the correct set of analogous faculties was implemented in an AI.

However, Armstrong, Sandberg, and Bostrom (2012) are skeptical of teaching AI complex concepts through feedback, as humans are biologically similar and thus predisposed to learn similar concepts, whereas an AI system with a different cognitive architecture might learn very different concepts.

This paper proposes and examines a strategy for AI engineering which builds on a hypothesis about human cognition. The hypothesis is that the human brain generates

its concepts using a relatively limited set of rules and mechanisms, which we are making good progress on reverse-engineering. If this were the case, it could be feasible to design AI systems that used similar criteria and mechanisms for generating their own concepts, and would thus learn similar concepts as humans did.

We begin by reviewing some of the research on human concept formation.

Research on human concept generation

One school of research views human concepts as being created by a process of optimizing for a specific set of constraints. Prince and Smolensky (1997) suggest that a sentence in a human language is grammatical if it optimally satisfies a conflicting set of constraints, and that differences between grammars can be traced to differences in how the constraints are ranked. For example, the English sentence "it rains" is "piove" (literally, "rains") in Italian. In English, the constraint forbidding subjectless sentences outranks the constraint forbidding meaningless words, and vice versa in Italian. The authors suggest that the set of constraints is universal for every language, and that the constraints emerge from an underlying neural optimization dynamic. A grammar is a set of relative strengths, and learning a grammar involves adjusting those strengths.

We can consider the process of learning the concept of "a valid English sentence" to be equivalent to learning a grammar, thus making the process of grammar learning an instance of concept learning. Many procedural skills seem to similarly involve mutually conflicting rules that all have to be satisfied: consider for instance a social skill such as "being funny without offending anyone present". An AI might need to similarly learn various concepts that were implicit in procedural skills, and which might be learned by a process of optimizing for the mutual satisfaction of a set of constraints.

More explicit concepts may also be the results of an optimization process. Regier, Kemp, and Key (in press) argue that although different languages have somewhat different concepts, the variation is constrained by simplicity on one hand, and a need for precisely communicating different concepts on the other. Drawing on work in the domains of color, kinship, and binary feature vectors, they propose that human languages tend to develop concepts that achieve a near-optimal tradeoff between simplicity and efficient communication. Khetarpal et al. (2013) find similar results for the domain of spatial terms.

Other modeling approaches have also found rules that produce human-like concepts. Kemp and Tenenbaum (2008) developed a structure learning approach that considered a number of different kinds of structures including trees, linear orders, multidimensional spaces, rings, cliques, and others. When applied to different physical, biological, and social domains, it produced similar classifications as humans would.

Tenenbaum (2011) reviews work on probabilistic concept learning and mentions the example of a child seeing three examples of different kinds of horses and correctly learning to generalize the word "horse" based on this information. Why does the child learn this hypothesis and not some other, such as "all animals" or "all horses except Clydesdales"? Tenenbaum argues that Bayesian likelihood favors the smaller sets, "horses" and "all horses except Clydesdales", since it would be less likely for three random samples to fall within the smaller sets if they were actually drawn from the larger set of "all animals". At the same time, there's a reasonable prior belief distribution that favors "all animals" and "all horses" for being more coherent and distinctive categories; "all horses" is then the only hypothesis favored by both the prior and the likelihood.

Other approaches also create similar classifications as humans tend to do, even though they have not been explicitly designed with the intention of mimicking human thought. For example, Shamir and Tarakhovsky (2012) used a classification scheme originally developed for biomedical image analysis, and found that it could classify artists by their artistic movements in a manner that agreed with the analysis of art historians.

Engineering human concepts in AI systems

In the previous section, we reviewed some work that has studied human concept learning, and which has either identified potential rules or constraints that humans follow while learning different concepts, or which has come up with explicit algorithms that produce similar concepts and classifications as humans do.

Our proposal is for a research and engineering program that would 1) map the rules behind human concept learning in more detail and 2) use the results of that learning to build AI systems that follow the same rules to produce the same concepts.

It should be noted that it may be very important to get the learning rules exactly right: a minor-seeming difference in a concept may turn out to be crucial (Yudkowsky 2011). Among humans, minor differences between concepts already lead to strongly differing moral judgments. For example, the question of whether unborn fetuses or brain-dead patients on life support count as humans worthy of protection has caused considerable controversy. In humans this is arguably caused by differences in the concept of "human"; similarly, a relatively minor difference between an AI's and a human's understanding of a concept could

lead the AI to take actions that humans considered clearly wrong.

Although the preceding sections have suggested that human concept learning may be reducible to relatively straightforward mathematical problems, it needs to be noted that the full picture is likely more complex. One complicating factor is the question of embodiment: human concepts and thought may be deeply rooted in our physical body.

A particularly relevant form of embodiment is that many of our moral judgments may be based on physical feelings of disgust and intuitions that were originally related to physical purity (Schnall et al. 2008). Correctly incorporating these intuitions might in principle require a simulation of the body's disgust responses.

Other thought processes also seem to have fewer purely abstract components than a mathematical analysis might suggest. For example, imagining an action involves a partial activation of the same neural systems that are involved in actually performing that action. Niedenthal et al. (2005) argue that the brain contains few to no abstract, amodal representations, and that concepts are formed by combining modality-specific features of different categories. This suggests that an AI might not be capable of learning human-like concepts unless it had similar sensory modalities as humans did. Lakoff and Núñez (2000) argue that even abstract mathematics is grounded in specific features of the human body, such as some of the axioms of set theory being derived from the structure of the human visual system.

Human cognition may also include other properties that are at their core evolutionary vestiges and which cannot be naturally derived from mathematical principles. Human writing systems may be constrained by the properties of the evolutionarily older circuits in the visual system that have been recruited for the task deciphering written text (Dehaene and Cohen 2007), and human reasoning may employ specialized modules (Barrett and Kurzban 2006) evolved for specific evolutionarily useful tasks such as detecting cheaters (Cosmides and Tooby 1992).

Another factor which may complicate the task of concept learning is the social nature of concepts. Humans do not learn concepts in isolation, but rather in a social environment where they gain rich feedback on both their communication and general behavior. More detailed concepts are learned if they are useful for some specific task, and human attention is guided towards noticing subtle differences in the domains that we are interested in and encouraged to attend to.

Goertzel and Pitt (2012) recommend teaching AI systems morality by "[p]rovid[ing] rich ethical interaction and instruction, respecting developmental stages", and this may indeed be necessary. In addition, it may also be necessary to understand the reward system that guides human attention and learning, and build the AI in such a way that the reward system is sufficiently similar.

Verifying conceptual equivalence

Building an AI which does actually have human-like concepts requires an ability to inspect and verify its internal concepts and compare them to human ones. Although testing the AI in different situations to see whether it behaves as expected given the desired concepts may be of some use, the AI also needs to behave correctly in entirely unanticipated situations. Particularly an AI that becomes more powerful than its programmers may end up in completely novel situations, and manifest unintended behavior outside its training and testing environment (Yudkowsky 2008), but the danger applies to less powerful AI systems as well. Furthermore, if the AI's concepts cannot be directly examined, there exist the possibility of a "treacherous turn" (Bostrom 2014), with the AI intentionally acting in ways that misguide its examiners about its intentions.

This may require not only understanding the rules that generate human concepts, but also mapping out the actual concepts that humans have so that they can be compared with those of the AI. For this purpose, there needs to be a format that both human and AI concepts can be mapped into, so that they could be compared with each other.

Gärdenfors (2000) proposes a general theory of representation, representing concepts as geometrical structures within a multidimensional space. Some work has built on this foundation and discussed communication between agents that have differing conceptual spaces (Honkela et al. 2008), as well as comparing the conceptual differences between individuals (Honkela et al. 2010, Honkela et al. 2012).

A number of brain imaging studies have also tried to understand how the brain represents information and to uncover the representational geometry of different concepts (Davis and Poldrack 2013, Kriegeskorte and Kievit 2013): possibly some of this research could eventually be leveraged for creating a way to compare the concepts of a particular group of humans with those stored in an AI's reasoning systems.

Discussion

Above, we have reviewed work that is aimed at uncovering the way the human brain generates concepts, and suggested that an AI system might come to have human-like concepts if it implemented similar mechanisms.

The project of designing such an AI involves at least the following components: (1) discovering the logical and mathematical criteria used in concept learning (2)

incorporating the effects from the embodied nature of human thought (3) incorporating any evolutionary vestiges that influence our concept formation and which cannot be naturally derived from mathematical principles (4) incorporating the social learning mechanisms which guide concept formation in humans (5) creating the means to directly compare conceptual representations between humans and AI systems.

The effort involved with these different steps is unclear. In particular, the effort involved with each of the stages of 2-4 might range between "trivial" and "insurmountable".

There is also the possibility that an AI that were to reason about concepts in a human-like way would require a design that was itself almost human, and an AI design that was very close to human might be easily outcompeted by AIs that were not so constrained (Sotala and Yampolskiy 2013).

On the other hand, the results from existing concept learning research suggests that there might be room to be optimistic, since relatively simple principles seem sufficient for learning many human concepts correctly. Whether or not this remains true once the field moves past toy models remains to be seen.

Acknowledgments

Thanks to Steve Rayhawk for invaluable assistance on this paper. This work was supported by Meru Foundation grant BAARI-001.

References

Allen, C.; Varner, G.; and Zinser, J. 2000. Prolegomena to any future artificial agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12(3):251-261.

Armstrong, S.; Sandberg, A.; and Bostrom, N. 2012. Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines* 22(4):299–324.

Barrett, H.C. and Kurzban, R. 2006 Modularity in Cognition: Framing the Debate. *Psychological Review* 113(3):628-647.

Bostrom, N. 2014 *Superintelligence: Paths, Dangers, Strategies*. Italy: Oxford University Press.

Chalmers, D.J. 2010. The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17(9–10):7–65.

Cosmides L and Tooby, J. 1992. Cognitive adaptations for social exchange. In *The adapted mind: Evolutionary psychology and the generation of culture*, 163–228. Oxford: Oxford University Press.

Davis, T. and Poldrack, R.A. 2013. Measuring neural representations with fMRI: practices and pitfalls. *Annals of the New York Academy of Sciences* 1296:108-34.

Dehaene, S. and Cohen, L. 2007 Cultural Recycling of Cortical Maps. *Neuron* 56(2):384–398.

Gärdenfors, P. 2000 *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Goertzel, B. and Bugaj, S.V. 2008. Stages of Ethical Development in Artificial General Intelligence Systems. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 448–459. Amsterdam: IOS.

Goertzel, B. and Pitt, J. 2012. Nine Ways to Bias Open-Source AGI Toward Friendliness. *Journal of Evolution and Technology* 22(1):116–131.

Guarini, M. 2006. Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems* 21(4):22–28.

Honkela, T; Könönen, V; Lindh-Knuutila, T.; and Paukkeri, M-S. 2008. Simulating Processes of Concept Formation and Communication. *Journal of Economic Methodology* 15(3):245–259.

Honkela, T.; Janasik, N.; Lagus, K.; Lindh-Knuutila, T.; Pantzar, M.; Raitio, J. 2010. GICA: Grounded Intersubjective Concept Analysis. A Method for Enhancing Mutual Understanding and Participation. Technical Report, TKK-ICS-R41. Aalto-ICS, Espoo, Finland.

Honkela, T.; Raitio, J.; Lagus, K.; Nieminen, I.T.; Honkela, N.; Pantzar, M. 2012. Subjects on Objects in Contexts: Using GICA Method to Quantify Epistemological Subjectivity. WCCI 2012 IEEE World Congress on Computational Intelligence.

Khetarpal, N.; Neveu, G.; Majid, A.; Michael, L.; and Regier, T. 2013. Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Kemp, C. and Tenenbaum, J. B. 2008. The discovery of structural form. *Proceedings of the National Academy of Sciences* 105(31), 10687-10692.

Kriegeskorte, N. and Kievit, R.A. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences* 17(8):401-12.

Lakoff, G. & Núñez, R. 2000. *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being.* New York: Basic Books.

Niedenthal, P.M.; Barsalou, L.W.; Winkielman, P.; Krauth-Gruber, S.; and Ric, F. 2005. Embodiment in Attitudes, Social Perception, and Emotion. *Personality and Social Psychology Review* 9(3):184–211.

Prince, A. and Smolensky, P. 1997. Optimality: From Neural Networks to Universal Grammar. *Science* 275(14):1604-1610.

Regier, T.; Kemp, C.; and Kay, P. In press. Word meanings across languages support efficient communication. In *The handbook of language emergence*. Hoboken, NJ: Wiley.

Schnall, S.; Haidt, J.; Clore, G.L; and Jordan, A.H. 2008. Disgust as Embodied Moral Judgment. *Personal and Social Psychology Bulletin*. 34(8):1096–1109.

Shamir, L. and Tarakhovsky, J.A. 2012 Computer Analysis of Art. *Journal on Computing and Cultural Heritage* 5(2).

Sotala, K. and Yampolskiy, R.V. 2013. Responses to Catastrophic AGI Risk: A Survey. Technical report, 2013-2. Berkeley, CA: Machine Intelligence Research Institute.

Tenenbaum, J.B.; Kemp, C.; Griffiths, T.L.; and Goodman, N.D. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* 331(6022):1279-1285.

Wallach, W. and Allen, C. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Yudkowsky, E. 2008. Artificial Intelligence as a Positive and

Negative Factor in Global Risk. In *Global Catastrophic Risks*, 308–345. New York: Oxford University Press.

Yudkowsky, E. 2011. Complex Value Systems are Required to Realize Valuable Futures. In *Artificial General Intelligence: 4th International Conference, Proceedings*, 388–393. Berlin: Springer.