Lab42

Essay Challenge

On the Principles of Intelligence:

Which fundamental principles of intelligence must be considered in the successful design of artificial intelligence?

Building human-like intelligence: an evolutionary perspective

Simon Ouellette

31. December 2022

1 Introduction

Human intelligence is not general, and the human mind does not start as a blank slate: it comes with a variety of domain-specific innate mechanisms (or biases) encoded in the genome, which manifest as rules for wiring up the brain [40, 25, 8, 32, 20]. So far, geneticists have identified almost 1000 genes that relate to human intelligence [13, 28, 29].

As an example of such innate biases, human infants can discriminate faces soon after birth. Primates are equipped with a cortical area known as the *fusiform face area* that, in some sense, hard-wires facial image processing. The specific content of distinct faces to recognize is of course learned, but the inductive biases that allow extraordinarily fast learning of facial recognition (compared to recognition of other objects) reside in that brain region [22, 15, 23, 18]. Intuitively, because facial recognition is so crucial to survival, evolution has put much selective pressure on developing an especially efficient mechanism for it.

Mammals, including humans, possess another specific cognitive mechanism that is located in the hippocampus and in the entorhinal cortex. These brain regions contain neurons known as "place cells" and "grid cells" respectively. The "grid cells" specifically solve the distance measurement problem, while the "place cells" act more like reference points in space. Together, they help form a cognitive map that allows the animal to move intentionally in its environment [17, 24, 14, 9]. Here as well, the scaffolding seems mostly hard-wired, while the specific contents (different spatial environments) are learned.

Even more abstract concepts such as cardinality and ordinality, the roots of mathematics, can be traced back to innate mechanisms present in other species. Indeed, mice, rats, pigeons, lions and honeybees can count, though in a more rudimentary sense than humans [5, 6, 21].

While the aforementioned examples can be seen as anecdotal, the "no free lunch theorem" demonstrates mathematically that no optimization, no learning, and thus no intelligence can be truly general and universal [38]. For every learning model there is a data distribution on which it will underperform relative to another algorithm. In

other words, any optimization or learning algorithm must either implicitly or explicitly restrain its search space.

Practically speaking, there is a known tradeoff between imposing a strong hypothesis structure on the data (in which case learning can be done more efficiently) and a weak hypothesis (in which case learning will require more data). Deep learning tends to fall in the latter camp. Meta-learning seeks to learn that hypothesis structure, as will be discussed in section 2.

Although not a crucial point to the argument being made here, it can be noted that there is an equivalent tradeoff in the realm of biology. While the advantages of adaptability are obvious, they can come at the cost of an excessively slow lifetime skill acquisition. Neuroscientist Anthony M. Zador summarizes this point in his "critique of pure learning" [40]: "There is, thus, pressure to evolve an appropriate tradeoff between innate and learned behavioral strategies, reminiscent of the bias-variance tradeoff in supervised learning".

Shortcut learning [12] is a well-known problem of deep learning that illustrates the need for useful biases. It refers to the observed tendency in deep neural networks to learn the "easy" solutions to the training set, which usually means the superficial ones. Instead of learning deep, meaningful representations in the way humans do, deep neural networks will latch onto any superficial correlations it can find. This is especially obvious in the field of computer vision.

As an example, convolutional neural networks (CNNs) trained to identify a cow will overfit the backgrounds that occur in the training set. It will be unable to recognize the same cow in unfamiliar environments. It probably relies on patterns of color and frequency in the image as a whole, rather than learning the actual shape of the intended object [12]. Similarly, experiments have shown that CNNs prefer to learn texture over global shape, as it identifies a golf-ball textured teapot as a golf ball, rather than a teapot [1]. In the absence of meaningful inductive biases, it is easier to just latch onto texture than to learn a more meaningful representation based on a combination of factors, such as shape, color and texture. After all, the CNN only sees matrices of pixels.

In contrast, it is practically impossible for humans to look at the world as a matrix of pixels: the unconscious mind will automatically group things into objects before

they can even be consciously processed. Depth perception is instinctive: if a human sees part of an object protruding from behind a wall, immediately they will know that the rest of that object (probably) exists behind that wall. Object representation notions such as spatio-temporal cohesion (objects as connected, bounded wholes) and object permanence (fully or partially obstructed objects did not suddenly disappear) are core inductive biases for humans [32], but not for the CNN.

It could be argued that part of what is generally known as "common sense" is a collection of such pre-existing inductive biases and priors (both innate and acquired) that restrict the search space of solutions to "sensible" ones. Another aspect of common sense is the integrated use of a vast breadth of knowledge about the world that humans have access to, but that our typically narrow AI models simply do not.

There is a cognitive science theory that calls these built-in inductive biases, these fundamental building blocks of learning, Core Knowledge [32]. It has been proposed [40, 16] that identifying and implementing these Core Knowledge concepts is central to building human-like intelligence. The question is: how do we achieve this?

2 The Case for Meta-Learning

It would be unreasonable to expect to manually discover and implement every single one of these Core Knowledge principles. This is for the same reason that it would be unreasonable to try to manually craft a set of rules or hard-coded algorithms that can categorize animal species from images as well as a deep neural network. Indeed, manually identifying and implementing the result of a billion years of evolutionary experience about the world seems too ambitious. In the spirit of machine learning, these things should be meta-learned.

In a sense, life can be seen as a meta-learning algorithm. Evolution is the outer optimization loop. It selects those hard-wired inductive biases that make learning optimal for individuals [4]. The inner optimization loop, then, corresponds to the learning that occurs within a lifetime. These processes will be referred to as *inter-life* learning and *intra-life* learning respectively.

It should be noted that even *intra-life* learning can be said to contain an element

of meta-learning, in the form of knowledge reuse from past experience. However, this is not the type of meta-learning that will be referred to in this essay, as it is assumed that such *intra-life* meta-learning is a process that would emerge automatically from optimal *inter-life* learning. That is, because *intra-life* knowledge reuse is a useful skill, a well-functioning *inter-life* learning process would mold the *intra-life* model in such a way that it will be able to develop it. As a consequence of that, it can be seen as a notion that is secondary to what is being discussed in this essay.

There is a further sense in which evolution and meta-learning can be seen as analogous, which relates to sample efficiency. In meta-learning, the inner model is efficient at learning new tasks from very few samples, but a very large number of samples are still required to learn the meta-model. Similarly, in the realm of biology, humans are efficient at learning *intra-life* tasks because evolution, an extremely sample inefficient process, optimized our brains for this meta-task distribution which is human life.

This *inter-life* learning process includes what we typically call meta-learning in machine learning. We know for example that *Model-Agnostic Meta-Learning* (MAML) operates not so much by learning to adapt quickly, but by learning initial shared features that can be reused across different tasks of the meta-task distribution [26]. In other words, it learns Core Knowledge for that meta-task distribution. But MAML is a very strict, limited form of meta-learning compared to what is really needed: it only learns initial weights.

Some neural architectures are poorly suited to some types of problems, regardless of the values of the weights, which is why Neural Architecture Search [11] needs to play a role as well. For example, the presence (or absence) of external memory modules, attention mechanisms, and components such as recurrence and convolutions should all be part of that *inter-life* learning process' search space.

Even so, meta-learning initial weights and neural architecture are not sufficient. There are questions to be answered even with regards to principles as fundamental as the optimization algorithm itself. This is an open question, which will be explored further in the last section.

To summarize this section, these Core Knowledge principles that allow sample efficient, yet flexible, *intra-life* learning of tasks from the "human meta-task distribution"

should be meta-learned. This *inter-life* learning process implies more than just learning neural network weights, which is why more direct, conventional approaches would fail (unless all of the required indicative biases were known). Meta-learning implies, however, a meta-task distribution to learn from. Since we are trying to meta-learn the inductive biases that allow humans to gain an efficient grasp of the world that surrounds them, it is almost as if a simulation of life itself needs to be created. Is this even feasible?

3 Meta-learning the world?

Learning human-like intelligence in the real world, instead of a simulation, would almost necessarily take the same time it took for evolution to build it (perhaps slightly less time due to more efficient optimization algorithms). One cannot avoid the necessity of learning it in a simulation or, at least, of meta-learning Core Knowledge in a simulation, such that the *intra-life* learning in the physical world is relatively efficient. This is another reason why a kind of meta-learning is more appropriate than just directly learning to solve the simulations: some sort of efficient generalization to new environments is needed.

Obviously, one cannot build a perfect simulation of life. Instead, the solution might start with a successful series of educated guesses about what those Core Knowledge principles should be. Cognitive science already provides us with a lot of plausible starting points. A few examples would be: object permanence, cardinality, inductive biases for grounded language learning, inductive biases for coordination with other agents, logical reasoning, etc. [32, 31, 30, 34]

As a next step, one should implement sufficiently domain-randomized [33] batteries of tasks that test for each of these Core Knowledge elements. The key is to build those simulations in a such a way that they will generalize well to similar real-life tasks, without requiring them to be imitations of real life. This is why the field of research sometimes referred to as *Sim-to-real transfer* [27, 41] is crucial.

Taking the example of object permanence as a Core Knowledge concept, a variety of tasks that require this fundamental notion can be built. One could show mov-

ing objects, sometimes in ways that are visibly obstructed, and have the model count how many distinct objects exist in the video sequence. Temporarily obstructed objects should not be double-counted, otherwise object permanence has not been learned. The domain randomization can take many forms: a battery of such tasks could have realistic textures and objects, another could be based on abstract shapes with simplistic textures. Various different backgrounds are possible. An alternance of three-dimensional and two-dimensional representations could be used.

These task simulations should not necessarily be thought of as separate, independent universes. Ideally, they would all be integrated into a common environment in which evolve a multitude of reinforcement learning-based agents. In fact, using exactly such an approach, researchers [37] were able to teach a general notion of object permanence to a virtual agent. In this case, the simulation was a kind of *hide-and-seek* game using objects, called *cache*. In order to solve this task, reinforcement learning agents needed to implicitly learn concepts such as object permanence, occlusion and depth.

Because a reinforcement learning context seems almost necessary to build an intelligent agent, it should be noted that a well constructed *inter-life* learning process should also be able to develop and optimize intermediate rewards. One of the very first skills that a successful agent should develop is the ability to actively learn from its environment through careful experimentation. This almost necessarily implies the development of intermediate rewards such as curiosity.

To summarize, it should be possible to create simulations for which the goals can only be attained by learning the necessary Core Knowledge skills that have been identified as desirable for human-like intelligence. It is, in general, much easier to define a problem than to find its solution. That is the underlying assumption behind the proposed approach. By letting reinforcement learning agents play in such a simulation, researchers have shown that it was able to learn certain object representation principles such as permanence. The suggestion here is: why not extend this principle to all other Core Knowledge skills?

4 The Heterogeneous Mind

The paradigm presented in the previous sections is a central element of building human-like intelligence, but it does not cover all of it. The current mainstream paradigm of backpropagation should be questioned, since backpropagation is limited to learning differentiable functions (while, at the same time, being biologically implausible). A great example of this is the Abstraction & Reasoning Corpus [7]. It consists of a battery of very diverse visual reasoning problems based on grids. These tasks are fairly easy for humans to solve. End-to-end differentiable, gradient descent-driven algorithms such as backpropagation have a poor track record so far on this dataset. Instead, the state of the art is set by discrete search algorithms [39].

It has been suggested by many that neurally-guided discrete search (NGS) solutions, in particular those that belong to the field of inductive program synthesis, could be the most promising approach to solving the Abstraction & Reasoning Corpus. Examples of such algorithms include DreamCoder [10] and DeepSynth [19]. These approaches work by using a neural network to propose probabilities on a context-free grammar after having observed a number of input/output examples for a task.

This probabilistic context-free grammar is essentially a library of function primitives and a set of construction rules to which probabilities are assigned. As such, it defines a search space over the possible solutions for the input problem. A combinatorial search algorithm is then used to search this space and return the most probable program solution. Because the neural network can learn to suggest different search probabilities for different tasks over a problem domain (i.e. meta-task distribution), it can be considered a meta-learning approach.

Neurally-guided search has an interesting analogy in cognitive science: procedural versus descriptive knowledge [10]. In the same way that humans can be said to combine intuitive (fast, procedural) thinking and symbolic (slow, descriptive) reasoning, these algorithms combine a neural network that learns this procedural know-how with a symbolic search that learns the descriptive knowledge.

It is noteworthy that discrete search solutions struggle with continuous parameters: they need to integrate a continuous component such as gradient descent for that

purpose [10, 3]. Furthermore, discrete search solutions are inappropriate in fields such a computer vision, where end-to-end differentiable neural networks currently dominate. We can gather from this that at least two fundamentally different types of optimization would be helpful: a continuous one and a discrete (symbolic) one.

Yet, it is implausible that the human brain implements two distinct mechanisms for learning. Furthermore, because synaptic weights are adjusted gradually, the brain's learning process is much closer to gradient descent and similar continuous optimization methods than to a discrete search, at least at the fundamental level. A neuroscience study [2] offers some insight into this dilemma: researchers have analyzed behavioral and neuroimaging data in stroke patients. Using Raven's Colored Progressive Matrices to evaluate a patient's reasoning ability (in what is generally considered a non-verbal task), they found that language-impaired patients underperformed relative to a control group. With further neuroimaging-based analysis, they concluded that deficits on the relational reasoning problems were associated with lesions in the brain regions that are necessary for language processing. This suggests that language is necessary for higher-level reasoning and problem-solving.

Another empirical result that supports this idea is the relatively recent realization, in the field of natural language processing, that reasoning appears as an emergent property in sufficiently large language models [35, 36].

In other words, much like in neurally-guided search methodologies, the symbolic grammar (or language) is not just a means of communication but also a fundamental building block of reasoning. The proposed hypothesis is that combinatorial search ability is an emergent faculty made possible through the evolution of language. The learning process, at the foundation of it all, is still continuous and gradual rather than discrete and combinatorial. Perhaps somewhere between NGS methods and large language models lies the solution to a system that can truly reason, yet can meta-learn its function primitives from experience.

In conclusion, human-like intelligence is not to be understood as a monolithic, universal, blank-slate algorithm. Instead, it is a heterogeneous, organic collection of submodules and inductive biases that were carefully molded by evolution. These learning biases allowed humans to be efficient at learning the relevant tasks of their existence

on earth, while staying sufficiently adaptable and flexible. An approach that is analogous to evolution could be used to meta-learn similar biases, through the careful use of task simulations guided by cognitive science principles.

Currently, deep learning approaches tend to underperform on reasoning tasks, where discrete algorithms still mostly dominate. However, large language models have recently demonstrated an emergent ability to reason through the use of language. It is hypothesized that once an otherwise continuous learning process has evolved the notion of language, combinatorial search at that higher level of symbols is made possible. This, in turn, potentially solves the "continuous vs discrete" dilemma.

Combining the insights presented here, a final hypothesis is suggested: if one were to ground (in the technical sense of "grounded language learning") these large language models into a simulation based on a sufficiently general meta-task distribution of Core Knowledge problems, what extraordinary capabilities might emerge in these virtual agents?

References

[1] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.

[2] Juliana V Baldo, Silvia A Bunge, Stephen M Wilson, and Nina F Dronkers. Is relational reasoning dependent on language? a voxel-based lesion symptom mapping study. *Brain and language*, 113(2):59–64, 2010.

[3] Matej Balog, Alexander L Gaunt, Marc Brockschmidt, Sebastian Nowozin, and Daniel Tarlow. Deepcoder: Learning to write programs. *arXiv preprint arXiv:1611.01989*, 2016.

[4] Patrick Bateson. Adaptability and evolution. *Interface Focus*, 7(5):20160126, 2017.

[5] Bilgehan Çavdaroğlu and Fuat Balcı. Mice can count and optimize count-based decisions. *Psychonomic bulletin & review*, 23(3):871–876, 2016.

[6] Lars Chittka and Karl Geiger. Can honey bees count landmarks? *Animal Behaviour*, 49(1):159–164, 1995.

[7] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[8] Leda Cosmides and John Tooby. Origins of domain specificity: The evolution of functional organization. *Mapping the mind: Domain specificity in cognition and culture*, 853116, 1994.

[9] Christian F Doeller, Caswell Barry, and Neil Burgess. Evidence for grid cells in a human memory network. *Nature*, 463(7281):657–661, 2010.

[10] Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B Tenenbaum. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd acm sigplan international conference on programming language design and implementation*, pages 835–850, 2021.

[11] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[13] William David Hill, Riccardo E Marioni, Omeed Maghzian, Stuart J Ritchie, Saskia P Hagenaars, AM McIntosh, Catharine R Gale, Gail Davies, and Ian J Deary. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Molecular psychiatry*, 24(2):169–181, 2019.

[14] Joshua Jacobs, Christoph T Weidemann, Jonathan F Miller, Alec Solway, John F Burke, Xue-Xin Wei, Nanthia Suthana, Michael R Sperling, Ashwini D Sharan,

Itzhak Fried, et al. Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature neuroscience*, 16(9):1188–1190, 2013.

[15] Nancy Kanwisher and Galit Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476):2109–2128, 2006.

[16] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[17] Rosamund F Langston, James A Ainge, Jonathan J Couey, Cathrin B Canto, Tale L Bjerknes, Menno P Witter, Edvard I Moser, and May-Britt Moser. Development of the spatial representation system in the rat. *Science*, 328(5985):1576–1580, 2010.

[18] Klaus Libertus, Rebecca J Landa, and Joshua L Haworth. Development of attention to faces during the first 3 years: Influences of stimulus type. *Frontiers in psychology*, 8:1976, 2017.

[19] Théo Matricon, Nathanaël Fijalkow, Guillaume Lagarde, and Kevin Ellis. Deepsynth: Scaling neural program synthesis with distribution-based search. *Journal of Open Source Software*, 7(78):4151, 2022.

[20] Tetsuro Matsuzawa. Cognitive development in chimpanzees: A trade-off between memory and abstraction. *The making of human concepts*, pages 227–244, 2010.

[21] Karen McComb, Craig Packer, and Anne Pusey. Roaring and numerical assessment in contests between groups of female lions, panthera leo. *Animal Behaviour*, 47(2):379–387, 1994.

[22] Elinor McKone, Kate Crookes, and Nancy Kanwisher. The cognitive and neural development of face recognition in humans. 2009.

[23] CONSPEC Morton. Conlern: A two-process theory of infant face recognition. *Psychological Review*, (63):1743, 2019.

[24] Edvard I Moser, Emilio Kropff, May-Britt Moser, et al. Place cells, grid cells, and the brain's spatial representation system. *Annual review of neuroscience*, 31(1):69–89, 2008.

[25] Steven Pinker. The blank slate: The modern denial of human nature. *New York, NY, Viking. Popper, K.(1974). Unended Quest. Fontana, London*, 2004.

[26] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[27] Erica Salvato, Gianfranco Fenu, Eric Medvet, and Felice Andrea Pellegrino. Crossing the reality gap: a survey on sim-to-real transferability of robot controllers in reinforcement learning. *IEEE Access*, 2021.

[28] Jeanne E Savage, Philip R Jansen, Sven Stringer, Kyoko Watanabe, Julien Bryois, Christiaan A De Leeuw, Mats Nagel, Swapnil Awasthi, Peter B Barr, Jonathan RI Coleman, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics*, 50(7):912–919, 2018.

[29] Suzanne Sniekers, Sven Stringer, Kyoko Watanabe, Philip R Jansen, Jonathan RI Coleman, Eva Krapohl, Erdogan Taskesen, Anke R Hammerschlag, Aysu Okbay, Delilah Zabaneh, et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature genetics*, 49(7):1107–1112, 2017.

[30] Elizabeth Spelke, Sang Ah Lee, and Véronique Izard. Beyond core knowledge: Natural geometry. *Cognitive science*, 34(5):863–884, 2010.

[31] Elizabeth S Spelke. Core knowledge. *American psychologist*, 55(11):1233, 2000.

[32] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

[33] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[34] Giorgio Vallortigara. Core knowledge of object, number, and geometry: A comparative and neural approach. *Cognitive neuropsychology*, 29(1-2):213–236, 2012.

[35] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*, 2022.

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[37] Luca Weihs, Aniruddha Kembhavi, Kiana Ehsani, Sarah M Pratt, Winson Han, Alvaro Herrasti, Eric Kolve, Dustin Schwenk, Roozbeh Mottaghi, and Ali Farhadi. Learning generalizable visual representations via interactive gameplay. *arXiv preprint arXiv:1912.08195*, 2019.

[38] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[39] Yudong Xu, Elias B Khalil, and Scott Sanner. Graphs, constraints, and search for the abstraction and reasoning corpus. *arXiv preprint arXiv:2210.09880*, 2022.

[40] Anthony M. Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7, 2019.

[41] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744. IEEE, 2020.