"I'll have an H_2O." The second says, "I'll have an H_2O too." The second guy dies.) He should be able both to age and to die.

10. He should be able to know the differences between Scotch and Bourbon, and to develop a preference for one or the other, and enjoy it occasionally. Same for wine.

I'm human, and I can do, or have done, all those things (except die), which is precisely why I think this is a fool's errand. I think it is a terrible idea to develop robots that are like humans. There are 7 billion humans on earth already. Why do we need fake humans when we have so many real ones? The robots we have now are (primarily) extremely useful single-function machines that can weld a car together in minutes, 300 a day, and never feel like, well, a robot, or a rivethead (Hamper 2008).

Even this sort of robot can cause lots of problems, as substantial unemployment in industry can be attributed to them. They tend to increase productivity and reduce the need for workers (Baily & Bosworth 2014). If that's what single-purpose (welding) robots can do, imagine what a HLM could do. If you think it might not be a serious problem, read Philip K. Dick's story, *Do Androids Dream Electric Sheep* (Dick 1968), or better yet, watch Ridley Scott's film *Blade Runner* (Scott 2007) based on Dick's story. The key issue in this film is that HLMs are indistinguishable from ordinary humans and are allowed legally to exist only as slaves. They don't like it. Big trouble ensues. (Re number 6, above, our HLM should probably *not* enjoy Philip Dick or *Blade Runner*.)

What kinds of things should machines be able to do? Jobs inimical to the human condition. Imagine an assistant fireman which could run into a burning building and save the 4-year-old reading Dr. Seuss. There is work going on to develop robotic devices – referred to as exoskeletons – that can help people with profound spinal cord injuries to walk again (Brenner 2016). But this is only reasonable if the device helps the patient go where he wants to go, not where the robot wants to go. There is also work going on to develop robotic birds, or orniothopters, among them the "Nano Hummingbird" and the "SmartBird." Both fly with flapping wings (Mackenzie 2012). The utility of these creatures is arguable; most of what they can do could probably be done with a $100 quad-copter drone. (Our HLM should be able to fly a quad-copter drone. I can.)

Google recently reported significant improvements in language translation as a result of the adoption of a neural-network approach (Lewis-Kraus 2016; Turovsky 2016). Many users report dramatic improvements in translations. (My own experience has been less positive.) This is a classic single-purpose "robot" that can help translators, but no one ought to rely on it alone.

In summary, it seems that even with the development of large neural-network style models, we are far from anything in *Blade Runner*. It will be a long time before we can have an HLM that can both display a patellar reflex and move the pieces in a chess game. And that, I think, is a very good thing.

Autonomous development and learning in artificial intelligence and robotics: Scaling up deep learning to human-like learning

doi:10.1017/S0140525X17000243, e275

Pierre-Yves Oudeyer

Inria and Ensta Paris-Tech, 33405 Talence, France.

pierre-yves.oudeyer@inria.fr http://www.pyoudeyer.com

Abstract: Autonomous lifelong development and learning are fundamental capabilities of humans, differentiating them from current deep learning systems. However, other branches of artificial intelligence have designed crucial ingredients towards autonomous learning: curiosity and intrinsic motivation, social learning and natural interaction with peers, and embodiment. These mechanisms guide exploration and autonomous choice of goals, and integrating them with deep learning opens stimulating perspectives.

Deep learning (DL) approaches made great advances in artificial intelligence, but are still far from human learning. As argued convincingly by Lake et al., differences include human capabilities to learn causal models of the world from very few data, leveraging compositional representations and priors like intuitive physics and psychology. However, there are other fundamental differences between current DL systems and human learning, as well as technical ingredients to fill this gap that are either superficially, or not adequately, discussed by Lake et al.

These fundamental mechanisms relate to *autonomous development and learning*. They are bound to play a central role in artificial intelligence in the future. Current DL systems require engineers to specify manually a task-specific objective function for every new task, and learn through offline processing of large training databases. On the contrary, humans learn autonomously open-ended repertoires of skills, deciding for themselves which goals to pursue or value and which skills to explore, driven by intrinsic motivation/curiosity and social learning through natural interaction with peers. Such learning processes are incremental, online, and progressive. Human child development involves a progressive increase of complexity in a curriculum of learning where skills are explored, acquired, and built on each other, through particular ordering and timing. Finally, human learning happens in the physical world, and through bodily and physical experimentation, under severe constraints on energy, time, and computational resources.

In the two last decades, the field of Developmental and Cognitive Robotics (Asada et al. 2009; Cangelosi and Schlesinger 2015), in strong interaction with developmental psychology and neuroscience, has achieved significant advances in computational modeling of mechanisms of autonomous development and learning in human infants, and applied them to solve difficult artificial intelligence (AI) problems. These mechanisms include the interaction between several systems that guide active exploration in large and open environments: curiosity, intrinsically motivated reinforcement learning (Barto 2013; Oudeyer et al. 2007; Schmidhuber 1991) and goal exploration (Baranes and Oudeyer 2013), social learning and natural interaction (Chernova and Thomaz 2014; Vollmer et al. 2014), maturation (Oudeyer et al. 2013), and embodiment (Pfeifer et al. 2007). These mechanisms crucially complement processes of incremental online model building (Nguyen and Peters 2011), as well as inference and representation learning approaches discussed in the target article.

Intrinsic motivation, curiosity and free play. For example, models of how motivational systems allow children to choose which goals to pursue, or which objects or skills to practice in contexts of free play, and how this can affect the formation of developmental structures in lifelong learning have flourished in the last decade (Baldassarre and Mirolli 2013; Gottlieb et al. 2013). In-depth models of intrinsically motivated exploration, and their links with curiosity, information seeking, and the "child-as-a-scientist" hypothesis (see Gottlieb et al. [2013] for a review), have generated new formal frameworks and hypotheses to understand their structure and function. For example, it was shown that intrinsically motivated exploration, driven by maximization of learning progress (i.e., maximal improvement of predictive or control models of the world; see Oudeyer et al. [2007] and Schmidhuber [1991]) can self-organize long-term developmental structures, where skills are acquired in an order and with timing that share fundamental properties with human development (Oudeyer and Smith 2016). For example, the structure of early infant vocal development self-organizes spontaneously from such intrinsically motivated exploration, in interaction with the physical properties of the vocal systems (Moulin-Frier et al. 2014). New experimental paradigms in psychology and neuroscience were recently developed and support these hypotheses (Baranes et al. 2014; Kidd 2012).

These algorithms of intrinsic motivation are also highly efficient for multitask learning in high-dimensional spaces. In robotics, they allow efficient stochastic selection of parameterized experiments and goals, enabling incremental collection of data and learning of skill models, through automatic and online curriculum learning. Such active control of the growth of complexity enables robots with high-dimensional continuous action spaces to learn omnidirectional locomotion on slippery surfaces and versatile manipulation of soft objects (Baranes and Oudeyer 2013) or hierarchical control of objects through tool use (Forestier and Oudeyer 2016). Recent work in deep reinforcement learning has included some of these mechanisms to solve difficult reinforcement learning problems, with rare or deceptive rewards (Bellemare et al. 2016; Kulkarni et al. 2016), as learning multiple (auxiliary) tasks in addition to the target task simplifies the problem (Jaderberg et al. 2016). However, there are many unstudied synergies between models of intrinsic motivation in developmental robotics and deep reinforcement learning systems; for example, curiosity-driven selection of parameterized problems/goals (Baranes and Oudeyer 2013) and learning strategies (Lopes and Oudeyer 2012) and combinations between intrinsic motivation and social learning, for example, imitation learning (Nguyen and Oudeyer 2013), have not yet been integrated with deep learning.

Embodied self-organization. The key role of physical embodiment in human learning has also been extensively studied in robotics, and yet it is out of the picture in current deep learning research. The physics of bodies and their interaction with their environment can spontaneously generate structure guiding learning and exploration (Pfeifer and Bongard 2007). For example, mechanical legs reproducing essential properties of human leg morphology generate human-like gaits on mild slopes without any computation (Collins et al. 2005), showing the guiding role of morphology in infant learning of locomotion (Oudeyer 2016). Yamada et al. (2010) developed a series of models showing that hand-face touch behaviours in the foetus and hand looking in the infant self-organize through interaction of a non-uniform physical distribution of proprioceptive sensors across the body with basic neural plasticity loops. Work on low-level muscle synergies also showed how low-level sensorimotor constraints could simplify learning (Flash and Hochner 2005).

Human learning as a complex dynamical system. Deep learning architectures often focus on inference and optimization. Although these are essential, developmental sciences suggested many times that learning occurs through complex dynamical interaction among systems of inference, memory, attention, motivation, low-level sensorimotor loops, embodiment, and social interaction. Although some of these ingredients are part of current DL research, (e.g., attention and memory), the integration of other key ingredients of autonomous learning and development opens stimulating perspectives for scaling up to human learning.

Human-like machines: Transparency and comprehensibility

doi:10.1017/S0140525X17000255, e276

Piotr M. Patrzyk, Daniela Link, and Julian N. Marewski

Faculty of Business and Economics, University of Lausanne, Quartier UNIL-Dorigny, Internef, CH-1015 Lausanne, Switzerland

piotr.patrzyk@unil.ch **daniela.link@unil.ch**
julian.marewski@unil.ch

Abstract: Artificial intelligence algorithms seek inspiration from human cognitive systems in areas where humans outperform machines. But on what level should algorithms try to approximate human cognition? We argue that human-like machines should be designed to make decisions in transparent and comprehensible ways, which can be achieved by accurately mirroring human cognitive processes.

How to build human-like machines? We agree with the authors' assertion that "reverse engineering human intelligence can usefully inform artificial intelligence and machine learning" (sect. 1.1, para. 3), and in this commentary we offer some suggestions concerning the direction of future developments. Specifically, we posit that human-like machines should not only be built to match humans in performance, but also to be able to make decisions that are both *transparent* and *comprehensible* to humans.

First, we argue that human-like machines need to decide and act in transparent ways, such that humans can readily understand how their decisions are made (see Arnold & Scheutz 2016; Indurkhya & Misztal-Radecka 2016; Mittelstadt et al. 2016). Behavior of artificial agents should be predictable, and people interacting with them ought to be in a position that allows them to intuitively grasp how those machines decide and act the way they do (Malle & Scheutz 2014). This poses a unique challenge for designing algorithms.

In current neural networks, there is typically no intuitive explanation for *why* a network reached a particular decision given received inputs (Burrell 2016). Such networks represent statistical pattern recognition approaches that lack the ability to capture agent-specific information. Lake et al. acknowledge this problem and call for structured cognitive representations, which are required for classifying social situations. Specifically, the authors' proposal of an "intuitive psychology" is grounded in the *naïve utility calculus* framework (Jara-Ettinger et al. 2016). According to this argument, algorithms should attempt to build a causal understanding of observed situations by creating representations of agents who seek rewards and avoid costs in a rational way.

Putting aside extreme examples (e.g., killer robots and autonomous vehicles), let us look at the more ordinary artificial intelligence task of scene understanding. Cost-benefit–based inferences about situations such as the one depicted in the leftmost picture in Figure 6 of Lake et al. will likely conclude that one agent has a desire to kill the other, and that he or she values higher the state of the other being dead than alive. Although we do not argue this is incorrect, a human-like classification of such a scene would rather reach the conclusion that the scene depicts either a legal execution or a murder. The returned alternative depends on the viewer's inferences about agent-specific characteristics. Making such inferences requires going beyond the attribution of simple goals – one needs to make assumptions about the roles and obligations of different agents. In the discussed example, although both a sheriff and a contract killer would have the same goal to end another person's life, the difference in their identity would change the human interpretation in a significant way.

We welcome the applicability of naïve utility calculus for inferring simple information concerning agent-specific variables, such as goals and competence level. At the same time, however, we point out some caveats inherent to this approach. Humans interacting with the system will likely expect a justification of why it has picked one interpretation rather than another, and algorithm designers might want to take this into consideration.

This leads us to our second point. Models of cognition can come in at least two flavors: (1) *As-if models*, which only aspire to achieve human-like performance on a specific task (e.g., classifying images), and (2) *process models*, which seek both to achieve human-like performance and to accurately reproduce the cognitive operations humans actually perform (classifying images by combining pieces of information in a way humans do). We believe that the task of creating human-like machines ought to be grounded in existing process models of cognition. Indeed, investigating human information processing is helpful for ensuring that generated decisions are comprehensible (i.e., that they follow human reasoning patterns).