Apply rich psychological terms in AI with care

Henry Shevlin and Marta Halina

March 2019

There is much to be gained from interdisciplinary efforts to tackle complex psychological notions such as 'theory of mind' by combining the rich history of study and debates in cognitive science and recent findings from AI research. However, careful and consistent communication is essential when comparing artificial and biological intelligence, say Henry Shevlin and Marta Halina.

Many of the most promising approaches in cognitive science seek to explicate notions like perception, belief, and motivation in information processing terms [1]. A similar move from information processing to psychology is occurring in artificial intelligence (AI) research and machine learning. Currently, few people working in AI would literally attribute beliefs, thoughts, or feelings to machines. However, as new techniques extend the capabilities of artificial systems, it has become increasingly common to use psychological terms to describe processing architectures [2–6].

Such use of psychological terms in AI research may be in many cases justified and unproblematic. Some terms with both narrow computational and psychological meanings, such as "memory" and "reinforcement", have clearly distinct histories in psychology and artificial intelligence and their senses are unlikely to be confused. Others such as "learning" and "behavior" straddle psychology and artificial intelligence, but are broad enough that there is little reason to quibble with their use to describe machine capacities. However, we argue here that there is a third kind of more robustly psychological concepts—including notions like awareness, perception, agency, and theory of mind—that have rich and complex histories in cognitive science. We suggest that these terms – which we call rich psychological concepts—require greater caution when employed to describe the capabilities of machine intelligence.

1 From psychology to AI and back

Our first concern for the use of rich psychological terms in artificial intelligence comes from the fact that for many such concepts (for example, theory of mind, or perception) there is already considerable controversy in cognitive science as to whether and when they can be applied even to biological organisms other than humans. In order to maintain consistency and clear communication across

different branches of cognitive science, it is thus key for artificial intelligence researchers who wish to employ these concepts to do so with an awareness of the standards of evidence and proof applied to them by researchers concerned with biological intelligence. More positively, we would also suggest that many in the machine learning community might gain insights into the development of sophisticated artificial intelligences by paying close attention to these debates.

Consider, for example, the debate concerning theory of mind (ToM) in non-human animals. In 1978, Premack and Woodruff asked whether chimpanzees have a ToM or the ability to attribute mental states to others—such as goals, knowledge, and beliefs [7]. The challenges involved in answering this question quickly became apparent, however, as researchers realised there were multiple available explanations for subjects' successful performance on ToM tasks. A subject could pass a false belief task, for example, by applying a behavioural rule, such as "an agent will choose the container that had the preferred reward placed into it while that agent was present" [8]. In grappling with this problem, psychologists developed alternative paradigms and standards for testing ToM. New experimental approaches, such as experience-projection tasks were designed to dissociate the attribution of mental states from the application of behavioural rules [9, 10]. Forty years and numerous studies later, a consensus is beginning to emerge that chimpanzees have the ability to attribute some mental states to others [11], although not everyone agrees [12]. Crucially for our purposes, however, the above challenges have led to a more nuanced understanding of ToM and how to test for it.

Recently, AI researchers have been exploring the idea of building an artificial theory of mind [4, 13, 14]. The results of these studies are impressive, with machines capable of predicting the behaviour of agents they have never seen before. Researchers working in this area draw heavily on contemporary work in cognitive science in order to guide the development and assessment of machine ToM [4, 14]. Engaging with cognitive science in this way provides AI developers with a rich database of theories, mechanisms, behaviours, and experimental tasks [15].

We think this is the right approach and would like to encourage more engagement of this kind. In the case of ToM research, for example, an artificial system like that developed by Rabinowitz and colleagues 2018 is an excellent example of how well an artificial system can learn to predict social behaviour. Within the context of a simple gridworld, their "Theory of Mind neural network" or ToMnet was able to pass traditional ToM tasks (including a false belief task) after being trained on no more than a population of behaving agents. The fact that psychologists would characterize ToMnet as engaging in behavior reading, rather than attributing mental states, does not diminish the significance of this development. Quite the contrary: one of the extraordinary implications of this study it that it demonstrates that an agent can pass traditional ToM tests using behaviour reading alone. Psychologists have hypothesized that this is possible, but have not demonstrated it empirically, as they lack full knowledge of an organism's learning history. AI research such as this can provide greater insight into the learning and cognitive mechanisms potentially involved in ToM in hu-

mans and other animals. In this case, perhaps one need not represent mental states in order to achieve sophisticated social behaviour after all.

While we see this example as constituting a success story for the use of rich psychological terms by AI researchers, the model it provides has not yet been consistently adopted by the field. While we do not wish to single out individual publications for practices that are widespread, we would note a quick glimpse at almost any leading machine learning publication will reveal frequent reference made to rich psychological concepts like understanding, motivation, and even creativity, without the same care and attention exhibited by studies like those mentioned above. This risks miscommunication and cross-purpose in interdisciplinary attempts to understand the basis of psychological capacities and in comparing the capacities of artificial and biological intelligences.

A key point to acknowledge here is that those working in artificial intelligence frequently have a kind of privileged access to the inner workings of the cognitive systems they build, an advantage not enjoyed in the field of animal behavior. Thus, perhaps AI researchers have special insight into cognitive mechanisms. This advantage should not be overstated, however. For one, there are well-known issues with the interpretability of artificial neural networks, such that the precise cognitive structures of artificial systems may not be fully transparent even to their creators [16]. More fundamentally, however, there is considerable debate concerning the underlying cognitive structure of capacities such as theory of mind or perception even in adult humans, with extensive fine-grained behavioural measures being our only current reliable test for their presence [17]. Hence simply knowing *how* a machine arrives at a given output does not automatically warrant the conclusion that the relevant process can be understood under the same psychological concept as that applied to adult humans. We therefore suggest that machine learning experts continue to engage closely with the standards of behavioural testing and theoretical wrangling found in other domains of cognitive science, and, in so doing, contribute their expertise and insights to these debates.

2 Normative dimensions of rich psychological terms

A second reason for diligence, we would suggest, is that some – and perhaps all – rich psychological term are 'thick concepts' [18], bringing with them normative considerations and hence the potential for ethical and legal ramifications. This is not the case for more practical concepts such as memory or learning, of course, but when we speak of a system as having intrinsic motivations, being an agent, or displaying imagination, there is a risk of implicitly licensing inferences about the moral or intellectual status of the system in question. Thus when authors speak of "intrinsically motivated [artificial] agents" [19], a natural but unwarranted (and doubtless unintended) assumption might be that the systems in question literally possess goals, desires, and perhaps even a form of moral responsibility. Similarly, when systems that learn with model-based reinforcement learning are described as having "imagination" [2], an incautious reader may leap to the

conclusion that the system in question possesses a capacity for the kinds of intellectual insight we associate with some of the most dramatic feats of human intelligence.

The machine learning community should not be held responsible for media sensationalism or misinterpretations of its claims, of course. However, we would suggest a pragmatic interest for researchers in avoiding terminology easily associated with human-level intellectual capacities or moral or agential status. Given the highly technical nature of most artificial systems, policymakers and the scientific media are often ill-placed to evaluate when such terms are used in an everyday, common sense way and when they are employed with a specific technical or scientific meaning. Gaps in understanding between the technical communities and the general public thus risks inviting moral panics, ill-considered legislation, and, more prosaically, disillusionment and squeezes in funding if and when technology fails to live up to misguided expectations. On these grounds, then, we suggest that it is in the interests of AI researchers to take particular care when invoking such rich psychological concepts.

3 Different kinds of intelligence

Finally, and most speculatively, we suggest that by overeagerly attributing familiar psychological notions to machine intelligence we may miss new and potentially illuminating information processing structures developed by AI researchers. To illustrate this point, compare current models of AI object recognition with their biological equivalents. The ability to spontaneously acquire new categories and apply them to inputs from external sensors is common to humans, animals, and many artificial systems.

Underlying this shared functional capacity, however, are key differences at the level of mechanism. For example, in order to reliably recognize a given object, artificial systems typically require a much larger set of training examples than humans or animals, and the kind of one-shot learning that is common in nature is extremely challenging for current computational architectures [20]. Similarly, contemporary artificial approaches to object recognition are easy to fool via the use of adversarial examples, while no clear parallel exists for biological perception (however, see [21] for some parallels). Artificial systems are also in many cases poor at extrapolating from natural images to cartoons or other pictorial forms of representation, a striking difference from the human case.

Yet there are also advantages of current machine systems over their biological counterparts, notably in speed and accuracy at classification among real world examples. Indeed, there is no reason some possible novel biological or artificial system to which we would comfortably ascribe general intelligence should have perceptual systems more like those found in animals and humans than those in current machine vision systems.

In light of this, we should take seriously the possibility that *new* rich psychological terms might be better suited to application in artificial systems. While it might be tempting to think of our repertoire of everyday psychological concepts

as static, even in the human case, science has regularly led us to innovate in this regard: notions like implicit bias, associative learning, and motivated reasoning have found their way into everyday explanations of behavior, but were initially coined by psychological researchers. Similar theoretical innovations might allow us to develop a set of new high-level psychological concepts able to capture the specific functional dynamics of artificial systems and allow for useful generalizations about their role in systems.

To get a sense of how such psychological neologisms might be put to work, consider the differences mentioned above between current machine vision systems and human perception. As noted, both forms of sensing have their own advantages and disadvantages. While human vision rarely results in the kinds of radical errors associated with adversarial examples (we are unlikely to mistake a toaster for an emu), in certain contexts it may be more prone to false negatives, as demonstrated by the recent outperformance of humans by AI in many medical imaging tasks [22]. In describing the sensory capacities of an artificial system, then, it might be more helpful – and more explanatory – to eschew reference to perception per se in favor of a new concept that more accurately captures the kinds of distinctive epistemological advantages and disadvantages of machine vision. By expanding our psychological vernacular to accommodate the burgeoning variety of high-level cognitive processes in artificial intelligence, we not only stand to gain in our ability to accurately characterize artificial systems, but might also in turn find applications for these new concepts in categorizing and understanding other kinds of non-human cognitive processes; these might include those of biological creatures quite different from ourselves, such as eusocial insects or cephalopods [23].

4 Conclusion

In summary, we claim that care should be taken before we employ the kind of psychological vocabulary currently applied to humans and animals to artificial systems, on the grounds that it may (i) impair scientific communication and understanding, (ii) invite premature conclusions of ethical or legal significance, and (iii) lead us to miss opportunities to expand our understanding of the varieties of minds.

It is commonly noted that previous 'AI winters' have been produced not so much by scientific stagnation as mismatches in expectations between researchers and broader academic and social communities. Ensuring such expectations are realistic by, among other things, reining in current tendencies to invoke rich psychological concepts to describe artificial systems will, we believe, help foster a more sustainable dialogue between researchers, funding bodies, and society at large. This is particularly important given that recent scientific developments have rendered it conceivable that artificial systems may soon literally possess some of the rich psychological capacities currently ascribed to them in only a narrow technical sense, and many readers—including informed ones—may struggle to determine which of these senses is being employed on a given occasion.

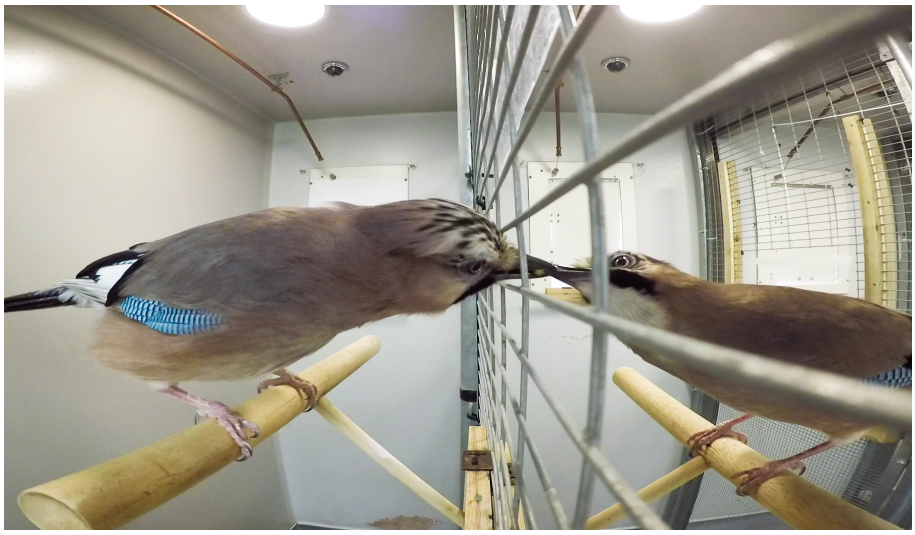

Figure 1: A male Eurasian jay shares food with a female partner according to her specific satiety. The attribution of desire-states may be an evolutionary and developmental precursor to Theory of Mind [24]. Image provided by Ljerka Ostojić (photo credit Piero Amodio and Ben Farrar).

5 Author Information

Affiliations: *Leverhulme Centre for the Future of Intelligence, University of Cambridge; Department of History and Philosophy of Science, University of Cambridge.* Marta Halina and Henry Shevlin. **Contributions:** Both authors contributed to writing this article. **Competing interests:** The authors declare no competing financial and non-financial interests.

References

1. Simon, H. A. Cognitive science: The newest science of the artificial. *Cognitive science* **4,** 33–46 (1980).
2. Racanière, S. *et al. Imagination-augmented agents for deep reinforcement learning* in *Advances in neural information processing systems* (2017), 5690–5701.
3. Kim, Y. *et al.* A bioinspired flexible organic artificial afferent nerve. *Science* **360,** 998–1003 (2018).
4. Rabinowitz, N. C. *et al.* Machine theory of mind. *arXiv preprint arXiv:1802.07740* (2018).

5. Shylaja, K., Vijayakumar, M., Prasad, E. V. & Davis, D. N. Artificial Minds with Consciousness and Common sense Aspects. *International Journal of Agent Technologies and Systems (IJATS)* **9,** 20–42 (2017).

6. Abate, T. *An Artificial Nerve System Gives Prosthetic Devices and Robots a Sense of Touch* `https://news.stanford.edu/2018/05/31/artificial-nerve-system-gives-prosthetic-devices-robots-sense-touch/`.

7. Premack, D. & Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* **1,** 515–526 (1978).

8. Lurz, R. W. *Mindreading animals: the debate over what animals know about other minds* (MIT press, 2011).

9. Karg, K., Schmelz, M., Call, J. & Tomasello, M. The goggles experiment: can chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour* **105,** 211–221 (2015).

10. Halina, M. There is no special problem of mindreading in nonhuman animals. *Philosophy of Science* **82,** 473–490 (2015).

11. Krupenye, C., Kano, F., Hirata, S., Call, J. & Tomasello, M. Great apes anticipate that other individuals will act according to false beliefs. *Science* **354,** 110–114 (2016).

12. Heyes, C. Apes submentalise. *Trends in cognitive sciences* **21,** 1–2 (2017).

13. Görür, O. C., Rosman, B. S., Hoffman, G. & Albayrak, S. Toward integrating Theory of Mind into adaptive decision-making of social robots to understand human intention (2017).

14. Winfield, A. F. T. Experiments in Artificial Theory of Mind: from safety to story-telling. *Frontiers in Robotics and AI* **5,** 75 (2018).

15. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95,** 245–258 (2017).

16. Selbst, A. D. & Barocas, S. The intuitive appeal of explainable machines. *Fordham L. Rev.* **87,** 1085 (2018).

17. Schaafsma, S. M., Pfaff, D. W., Spunt, R. P. & Adolphs, R. Deconstructing and reconstructing theory of mind. *Trends in cognitive sciences* **19,** 65–72 (2015).

18. Hare, R. M. *The language of morals* **77** (Oxford University Press, 1952).

19. Kulkarni, T. D., Narasimhan, K., Saeedi, A. & Tenenbaum, J. *Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation* in *Advances in neural information processing systems* (2016), 3675–3683.

20. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences* **40** (2017).

21. Elsayed, G. *et al. Adversarial Examples that Fool both Computer Vision and Time-Limited Humans* in *Advances in Neural Information Processing Systems* (2018), 3914–3924.

22. Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

23. Godfrey-Smith, P. *Other minds: The Octopus and the evolution of intelligent life* (William Collins London, 2016).

24. Ostojić, L., Shaw, R. C., Cheke, L. G. & Clayton, N. S. Evidence suggesting that desire-state attribution may govern food sharing in Eurasian jays. *Proceedings of the National Academy of Sciences* **110,** 4123–4128 (2013).