Decision-making in brains and robots - the case for an interdisciplinary approach

Sang Wan Lee[1,2,3]* and Ben Seymour[4,5,6]*

[1]Department of Bio and Brain Engineering, Program of Brain and Cognitive Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

[2]KAIST Institute for Artificial Intelligence, Daejeon, Republic of Korea

[3]KAIST Institute for Health, Science, and Technology, Daejeon, Republic of Korea

[4]Computational and Biological Learning Laboratory, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK

[5]Brain Information Communication Research Laboratory Group, Advanced Telecommunications Research Institute International, Kyoto, Japan

[6]Center for Information and Neural Networks, National Institute for Information and Communications Technology, 1-4 Yamadaoka, Suita, Osaka, Japan, 565-0871

(*: corresponding authors. Email: sangwan@kaist.ac.kr, bjs49@cam.ac.uk)

Abstract

Reinforcement Learning describes a general method for trial-and-error learning, and has emerged as a dominant framework both for optimal control in autonomous robots, and understanding decision-making in the brain. Despite their common roots, however, these two fields have evolved largely independently. In this perspective we consider how each now face problems that could potentially be addressed by insights from the other, and argue that an interdisciplinary approach could greatly accelerate progress in both.

Introduction

Humans and autonomous robots share a common problem: how to survive in a complex, changing and dangerous world with little *a priori* knowledge, but with the requirement to continuously satisfy their appetites and avoid damage over long lifetimes. The key to survival is learning, and using what is learned to guide further decisions in as safe and efficient way as possible. But this is a difficult problem, with numerous challenges related to the practicalities of behaving in real physical environments full of noise and unpredictability. In recent years, Reinforcement Learning (RL) has emerged as a dominant theoretical framework to understand decision-making through an interaction with the world [1], and has underpinned research in both neuroscience and robotics. But despite a shared problem, the strategies which engineering science has developed differ in many respects with those in which brain appears to adopt. This raises the question as to whether brain-like algorithms can be shown to work in real control problems far beyond the highly controlled experiments from which the models were derived. In turn, it is also possible that brain-inspired solutions might be capable of improving aspects of robot control. In this perspective, we outline a series of areas where an interdisciplinary approach may have particular promise.

Reinforcement Learning. The goal of RL is to learn how to make decisions in a way that maximises future returns or minimises cost given an uncertain, but at least partly predictable, environment. In particular, RL addresses the credit assignment problem, which is ubiquitous in real-world decision-making problems. This credit assignment problem arises from the fact that many prediction and control situations ('Markov Decision Processes') involve complex sequences of states that precede an important outcome (i.e. a reward or punishment), making it difficult to know at what point the state or action took a turn for the better or worse. RL solves this by continually holding a prediction of the expected sum of future outcomes (i.e. the value), and updating this value as one moves through time (more formally, it solves the Bellman equation) [1]. If this new information changes the predicted value, then a prediction error is generated, and propagated back to earlier states to update their predictions so as to be more accurate in the future. In this way, RL algorithms can deal with sequential (higher-order) learning in a way that classic psychological learning models (such as the Rescorla-Wagner model)

cannot. Due to its ability to develop a control policy with minimal supervision, RL seemingly offers a practical tool for many optimal control problems [2].

Applications of RL to robotics in earlier days of research focused on simple robot control tasks, such as playing the game of backgammon in a simulation environment [3], repetitive pick-and-place [1], recycling pushing boxes with a wheeled mobile robot [4], positioning and inserting a peg [5], or learning frying pan skills with a robot arm [6]. Subsequently, advanced RL techniques have been applied to solve more challenging issues, including humanoid robot control [7], autonomous car driving [8], learning a control policy in highly complex environments [9]–[11], learning large repertoires of skills with robot arms [12], autonomous helicopter control [13], drone control [14], and robot swarm control [15]. This illustrates how the broad application of RL has extended from motor control to action control (i.e. decision-making).

Although RL's roots lie in early models of associative learning [16], the overt application to neuroscience emerged somewhat later, in terms of models of classical (Pavlovian) conditioning [1]. The idea that the brain might be physically implementing RL algorithms [17], [18] eventually found evidence in primate experiments showing what appeared to be direct encoding of a reward prediction error by dopamine neurons [19]. Evidence in humans subsequently came from functional neuroimaging experiments, where RL prediction errors were found for both reward and punishment learning [20], [21]. This stimulated enormous interest in RL models of human decision-making, and led to a number of key findings, including of distinct circuits for learning to make predictions and actions (Pavlovian and instrumental conditioning, respectively) [22], multiple circuits for action learning involving encoding state and action space with different levels of complexity [23], and of different strategies used to balance exploration with exploitation [24], [25].

But for many engineering applications, the practical utility of RL was limited in the face of the high dimensionality of sensors and actuators, and created a need for effective techniques at encoding state and action representations. A key breakthrough was the application of deep learning to approximate value functions, and this rapidly demonstrated success with agents operating in a high dimensional input space [9]–[11], as well as in a continuous action space

[12], [26], [27]. Notably, RL agents were subsequently shown to outperform humans [9], [10], and to learn without expert knowledge [11]. This provided a dramatic leap in both performance and efficiency across a wide range of real-world problems: from highly complex simulated games [9], [11] to robot manipulation [12], [28].

Although understanding autonomous control in robotics and neuroscience have proceeded somewhat independently, they may often be dealing with the *same* control problem (e.g. navigation, collision avoidance), especially since a common aspired to application of robotics is to support humans in their natural environment. With this convergence comes a number of areas where cross-fertilization of knowledge between the two might provide valuable new insights, as we describe below (also see **Figure 1**).

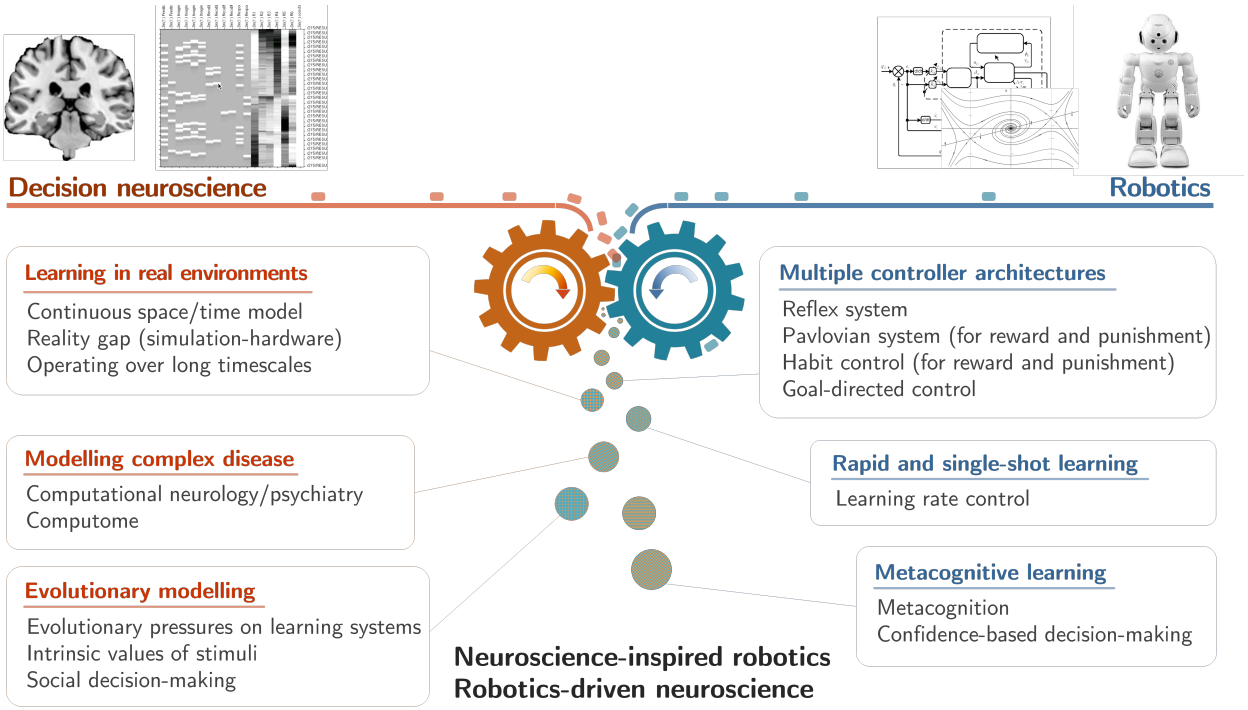

Figure 1. Overview of neuroscience - robotics approach for decision-making. The figure details key areas for interdisciplinary study.

Why decision neuroscience needs robotics

Despite significant progress in resolving the sophistication of computational models of brain-based learning and decision circuits, there is still a gulf between what they can currently explain and what they need to. Below we outline 3 key challenges:

Learning in real environments. Many models of human/animal decision-making are applied to experiment data that involves highly controlled environments, but as they become more complex, their validity outside of these contexts becomes more difficult to assume for several interacting reasons. First, almost all current models are discrete-state/time models, and extending them to continuous space is non-trivial [29]. This is because they require some way of dealing with the potential huge dimensional state- and action space, for example by partitioning or function approximation, whilst maintaining temporal smoothness of action [26], [30], [31]. Second, models that work in simulation are often not robust when operating in hardware - a phenomenon called the 'reality gap'. This is because of multiple sources of sensory noise and motor noise (flexibility, friction etc) that accompany any physical system, and the nature of this noise is difficult to predict. Third, uncertainty arises when building control systems that need to operate over very long periods (i.e. over development and lifetimes), because there are many changes in the noise and dynamics of the body/environment that happen over multiple time scales. Good evidence suggests that human decision-making adapts over such timescales (e.g. risk-taking, impulse inhibition, decision meta-cognition). These factors interact and emphasize the fact that without additional evidence, one should be cautious in assuming that current neuroscience models of decision-making will always work effectively across real-world situations and environments.

Modelling complex disease*.* One of the main objectives of decision neuroscience is to understand how its disruption might lead to disease. Computational neurology and computational psychiatry aim to understand how disease symptomatology emerges from underlying disturbance of specific computational elements i.e. how differences in the architecture or parameterisation of computational operations cause the disease state. [32]–[34]. However, it seems likely that many diseases do not involve a single computational element, but rather a set of interacting elements (a sort of 'computome') that *interact* together generate disease phenotype (**Figure 2**). But as the complexity of the underlying models increases, making reliable

predictions about how overall behaviour is changed when a specific computation element, or set of elements is altered, becomes much more difficult.

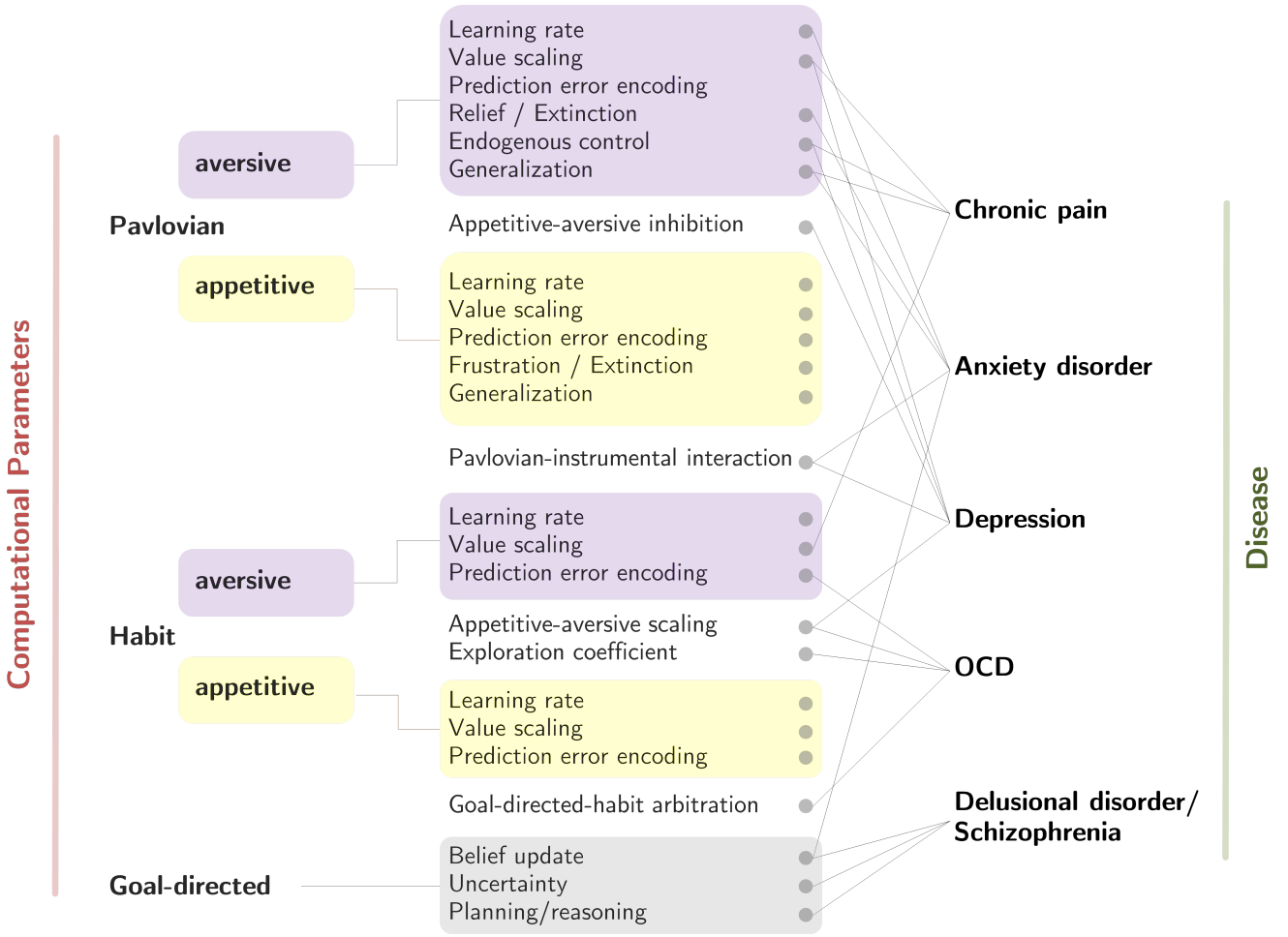

Figure 2. Computomics of disease. The diagram summarises some of the key parameters that are involved in human decision-making that can be proposed to be important for the potential development of neuropsychiatric disease. The general hypothesis is that these factors interact in difficult-to-predict ways to generate the overall disease phenotype.

For example, chronic lower back pain is thought to be primarily driven by brain-based factors (as opposed to over-activity of peripheral pain-sensing neurons (nociceptors)[35]. There are several

distinct factors proposed for how chronic pain might develop, including excessive pain value predictions [36], reduced perception of pain controllability (also a proposed factor in depression)[37], impaired relief prediction [38], negative expectation bias in perceptual inference [39], over-generalisation in both active and passive pain conditioning (also a proposed factor in anxiety)[40], and others. Each of these factors relate to distinct computational operations within the broad RL framework, and it is probable that they represent independent risk factors that when coexisting, interact to generate sufficient an effect to support the maintenance of pain. However, the complexity of this interaction means that how this happens is difficult to predict.

Evolutionary modelling. A key problem in decision neuroscience is to understand how and why certain behaviours and systems exist in the way that they do, especially since the very nature of human decision-making seems to make it prone to weaknesses such as impulsivity and compulsivity. Although laboratory experiments can address the *proximate* basis of much of the complexities of decision-making, understanding why neural systems are organised in the way they are (i.e. the *ultimate* basis) is critical to rationalising and predicting broader questions about the 'design' of these systems in the first place.

Evolutionary robotics provides a method for developing computational control architectures based on biologically-inspired evolutionary algorithms (i.e. Darwinian selection based on a fitness function), and so can be used to explore and mimic the selective pressures on neural control systems that have arisen during the course of animal evolution. This is a valuable approach because the vast complexity of the environment necessitates that decision systems have evolved primarily as *learning systems*, as opposed to hard-wired stimulus-response systems. However, the evolutionary pressures on learning systems are much less easy to predict, especially when multiple systems operate in parallel. However these pressures are critical in determining the complex architecture and meta-parameters of learning [41], [42]. Related to this is understanding how values (both reward and punishment) are bestowed through evolution on certain stimuli, such that they have become 'primary reinforcers' of behaviour [43]. This is especially important for certain states that have intrinsic value (e.g. novelty seeking, endogenous control of pain [44], since it allows animals to navigate complex decisions in which the ultimate benefit of a decision may be long-term, beyond the bounds of learnable experience.

One area where an evolutionary robotics approach may have particular value is in aspects of social decision making (e.g. sexual selection, cooperation and altruism, and observational learning and teaching). That is, many social decisions involve considerable complexity since action values depend on inferring and understanding the action and learning systems of others [45], and likely involve a complex combination of learning and intrinsic values that is difficult to predict without formal evolutionary simulation.

Taken together, these three challenges place limits on developing our understanding of decision mechanisms in neuroscience, and appeal to robotics-based approaches. Advanced simulation platforms and hardware implementation could offer a stepwise increase in demonstrating the computational validity and general applicability of brain-based models. And, in principle evolutionary robotics provides simulation platforms that allow interrogation of how such complex systems might emerge and how different aspects of decision-system might trade off against others, revealing the nature of traits that can create risk factors for neuropsychiatric disease.

Why robotics needs decision neuroscience

There are many previous examples in robotics in which biological inspiration has led to improvements in design and control systems, for example in motor control and artificial vision. It is therefore plausible that decision neuroscience could also offer novel insights, and below we describe three areas where this may be realistic.

***Multiple controller architectures*.** In contrast to most robot-control algorithms, humans and animals have multiple control systems that govern decisions and actions (**Figure 3**). First is a system for innate responses, which emit simple hard-wired behaviours in response to inherently salient stimuli (generalised reflexes such as approach or withdrawal, or specific reflexive actions such as leg-flexion to a shock). Second, these responses can also be transferred to predictive neutral stimuli through classical (Pavlovian) conditioning, which allows the behaviours to be emitted early. Furthermore, there are separate systems for reward and punishment, dealing with a full spectrum of positive and negative outcomes through specifically tuned anticipatory

responses [46]. Third, animals learn stimulus-response actions ('habits'), in which rewards or punishments reinforce actions that lead to them, such that after repeated pairing the actions are emitted 'automatically' when cued. Here too, there is also evidence for separate reward and punishment systems, tracking both best-case and worst-case scenario actions [47], [48]. Finally, a cognitive ('goal-directed') control system can guide actions by learning a model of environment, and leverages the learned knowledge base to quickly adapt to the change in the environment structure, such as a latent state-space or a reward structure. This allows great behavioral flexibility over performance when an environment structure changes. While this sophisticated control strategy is considered to be optimal in many scenarios, it requires a high computational cost and takes a longer time to process than the stimulus-response control system.

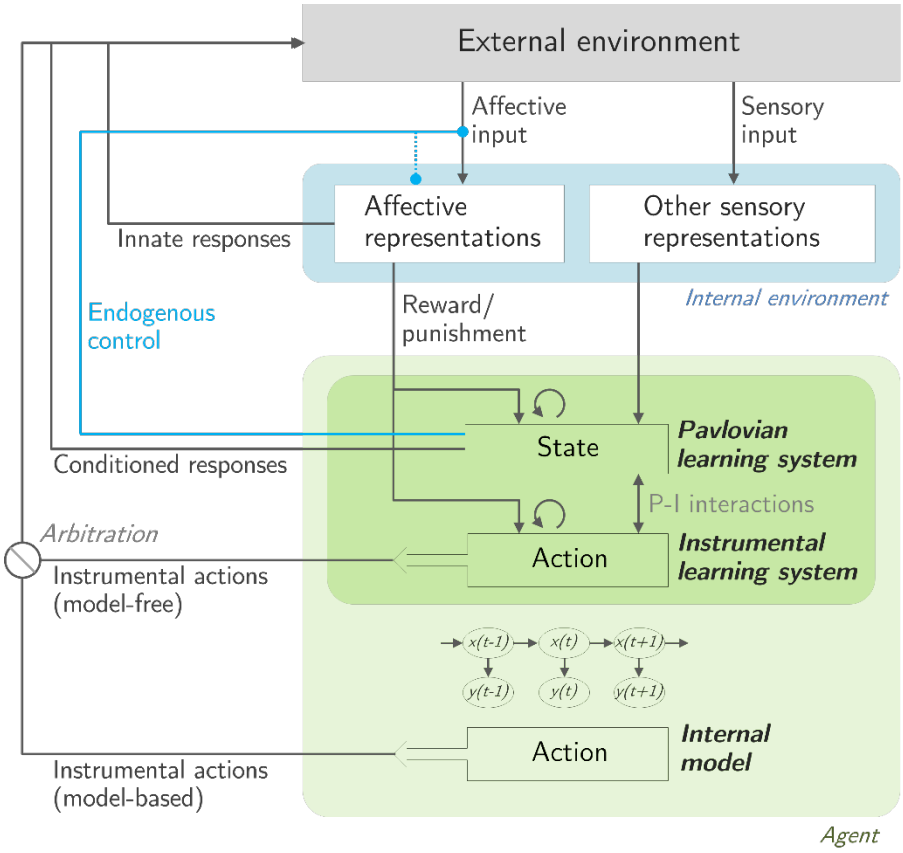

Figure 3. Overview of human decision making systems in the brain. Ultimately the brain is thought to implement a nested hierarchy of control systems: that range from implementation evolutionary acquired response systems that are quick and computationally efficient, to more

experience-based learning that involves progressively higher levels of computational sophistication.

Each system is best in different situations: the innate/Pavlovian systems exploits evolutionarily learned knowledge and allow rapid responses before much learning has happened, in effect providing an evolutionary prior on action space. Goal-directed learning allow sophisticated modelling of the world, supporting planning and flexibility [49]. And habit learning provides stability in the face of inherent unpredictability and over extended time, alongside considerable computational efficiency [50]. Given this multi-controller architecture, the brain therefore needs to decide how to integrate the different systems, and the brain appears to use a number of strategies, ranging from a simple scaling of system outputs (between reward and punishment systems), to direct competition between systems based on uncertainty by a "meta-controller" (between stimulus-response and goal-directed systems) [51]–[53].

Rapid and single-shot learning. Control algorithms for robots, especially in AI settings, usually require a large amount of experience. On the other hand, humans learn fast, often after a single experience. Consequently recent insights into the neuroscience of rapid learning in biological systems may provide practical insight to robotics.

First, it can improve optimality of robot learning. For example, a recent neuroscience study found that when there is only few examples available, or when interactions with environment are limited, humans have a strong tendency to increase their learning rates; they strive for quickly making sense of unknown parts of environment by compromising safe learning, rather than performing incremental learning [54]–[56]. A robotic system that can flexibly adjust its own learning speed would resolve a tradeoff between performance and speed.

Second, implementing human-style rapid learning in robots would improve human-robot interaction. Single-shot inference or jumping-to-conclusion is associated with suboptimal behavior that a rational robotic agent cannot predict. For example, a neuroscientific basis of rapid learning and inference[55] offers great potential for building robots that can interact with

suboptimal entities like humans. This can also foster social trust of the human users in service robotics environment.

Single-shot learning can be implanted into robots by taking up the idea of hippocampal episodic memory controller. It is originated from a theoretical idea in decision neuroscience, which has emerged as an effective alternative in the presence of a large amount of measurement noise in the environment or in the very early stage of learning [50]. The episodic memory control can guide decisions based on a single past episode that the agent remembers [57], [58], thereby allowing itself to kick-start learning when the environment becomes too complex or too noisy, or even when the it needs to transfer from one task to another.

Metacognitive learning*.* Reinforcement learning algorithms are optimistic (or over-confident) in that their current predictions are expected to be correct even if they are sampling from the part of the environment they haven't learned about enough. Learning without estimation of its own prediction performance (i.e., over-confident learning) may lead to resorting to a suboptimal policy (local minima problem), especially in a complex and dynamic environment.

Metacognition refers to an ability to evaluate the agent's own thought processes, such as perception [59], valuation [60], [61] and learning, inference [62], and to report the level of confidence/uncertainty about her choice leading to an outcome that she predicted [63]. The change in the confidence level depends on her ability to learn, e.g., a slow learner builds up confidence slowly, but it may also depend on various environmental contexts. For example, low task difficulty or low environmental noise would make the learning agent confident, leading to more decisive actions, whereas losing confidence in opposite cases would lead to a more cautious and defensive strategy. The metacognitive learning thus allows for rapid adaptation to the context change while maintaining robustness against environmental noise. This raises an optimistic expectation that the metacognitive learning agent would compensate for the weakness of model-based reinforcement learning that it inevitably suffers from a large amount of prediction error when the environment is highly noisy [52].

Growing evidence from decision neuroscience about computational mechanisms of metacognition and confidence-based decision-making [60], [61] may lend an insight into metacognitive learning algorithm design for robots [64]. It also has an enormous potential for further augmenting robot intelligence in many different ways. First, this ability would guide valuation taking into account uncertainty. Second, it could help resolve exploration-exploitation tradeoffs. For example, because a metacognitive agent has the ability to distinguish what it has learned from what it hasn't figured out yet, it can determine when to explore to learn more about the task and when to exploit to achieve a goal. Naturally a successful implementation of metacognitive reinforcement learning would bring about performance boost of decision-making models [65], [66].

Taken together, neuroscience offers insight into a set of solutions that might benefit robot control systems. Each involves adding complexity - multi-controller architectures, additional systems for rapid inference, and supervisory systems - but each of these seem to evolved in humans to deal with the practicalities and reality of living in the real-world. Although our current understanding of how the brain achieves are yet at a point to be directly implementable in robots, the principles they embody could inspire development of comparable strategies that confer flexibility and efficiency in robotics.

Conclusion

In summary, we have highlighted three key issues where decision neuroscience can be informed by robotics: learning in real environments (i.e. addressing the reality-gap in neuroscience models), modelling complex disease (making predictions when there are multiple risk factors), and evolutionary modelling (understanding why neural systems are structured in the way that they are). And we have highlighted three potentially valuable insights for robotics that come from decision neuroscience: adopting multi-controller architectures (to enhance safety and efficiency), rapid/single-shot learning (to enhance performance with very limited information exposure), and metacognition (to enhance robustness in the face of significant change). Together, they illustrate how an enhanced dialogue between neuroscience and robotics could be mutually beneficial.

Indeed an increasing number of studies are beginning to cross this boundary. Recent robotics and AI studies, for instance, draw on the idea of model-based reinforcement learning from decision neuroscience, showing that an agent possessing the human's ability to carry out simulations over predictive models of the environment can handle various physics-based navigation tasks [67], [68]. On the other hand, recent neuroscience studies adopted recurrent neural networks, a class of models frequently used for approximate optimal control of robots, to test theoretical ideas, such as that the structured representations of spatial information in the hippocampus improve the efficiency of goal-directed reinforcement learning [69], and that the midbrain dopamine system promotes meta-reinforcement learning in the prefrontal cortex [53].

Discovering computational principles of such meta-level functions in these brain circuits offer great potential for furthering the design of brain-robot interfaces. It would not only allow the ability to read out latent states of the brain, such as a learning strategy or a task goal [70], but also inform when and how the brain creates a new mental state [71]. Given that a set of possible choices varies according to the brain's latent states, a brain-robot interface equipped with this capability potentially allows a robot agent to make more precise predictions about user's intention.

It is also possible that robotics and neuroscience can synergistically work above and beyond simple interdisciplinary approaches. For instance, most of robot-based high-throughput screening or neuroscience studies require completion of the following cycle: hypothesis development, assay preparation / task design, data collection / experiment, and evaluation / data analysis. To test a hypothesis of interest against alternative ones, the task design requires careful manipulation of task variables while controlling for potential confounding effects. Rapid progress in machine learning may provide powerful tools for finding an optimal task parameter set that allows us to effectively contrast a main hypothesis with competing hypotheses [72], [73]. One of important directions for future research concerns AI-based task parameter optimization, in which AI being incorporated into robot-based automation of experiments [74], [75].

Finally, a key emerging concept in human-robot interaction is that robots may be better suited to human interactive environments if they learn, act and behave in a similar way to humans. This

may help in terms of human's ability to understand and infer the robots goals and intentions, enhance empathy towards the robot, facilitate social decision-making such as cooperation and joint decision-making, and promote observational learning. Indeed a key part of the future application of human-assistive robots for people with cognitive impairment could conceivably include decision support, in which a robot infers goals of a human and uses its own decision-making algorithms to suggest optimal decisions. Notwithstanding this, there is an inherent logic in having robots think in the same way as humans as being a key facet to successful integrative environments, even if this means robots occasionally displaying suboptimal human-like traits that emerge naturally from brain-like architectures, such as mild impulsivity and anxiety, in certain situations.

Acknowledgments.

B.S. is funded by the Wellcome Trust (097490), Arthritis Research UK (21357), and the National Institute of Information and Communications Technology of Japan. S.W.L. is supported by (i) the ICT R&D program of MSIP/IITP. [2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion], (ii) Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451), (iii) Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00677, Development of Robot Hand Manipulation Intelligence to Learn Methods and Procedures for Handling Various Objects with Tactile Robot Hands), (iv) Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-06 and (v) the research fund of the KAIST (Korea Advanced Institute of Science and Technology) (Grant code: G04150045).

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, vol. 1, no. 1. MIT press Cambridge, 1998.

[2] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, 2005.

[3] G. Tesauro and Gerald, "Temporal difference learning and TD-Gammon," *Commun. ACM*, vol. 38, no. 3, pp. 58–68, Mar. 1995.

[4] S. Mahadevan and J. Connell, “Automatic programming of behavior-based robots using reinforcement learning,” *Artif. Intell.*, vol. 55, no. 2–3, pp. 311–365, Jun. 1992.

[5] V. Gullapalli, J. A. Franklin, and H. Benbrahim, “Acquiring robot skills via reinforcement learning,” *IEEE Control Syst.*, vol. 14, no. 1, pp. 13–24, Feb. 1994.

[6] J. Kober and J. R. Peters, “Policy Search for Motor Primitives in Robotics.” pp. 849–856, 2009.

[7] Pamplona Daniela; Bernardino Alexandre, “Smooth Foveal Vision with Gaussian Receptive Fields,” *IEEE-RAS Int. Conf. Humanoid Robot.*, 2009.

[8] A. Sallab, M. Abdou, E. Perot, and S. Yogamani, “Deep Reinforcement Learning framework for Autonomous Driving,” *Electron. Imaging*, vol. 2017, no. 19, pp. 70–76, Jan. 2017.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[10] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[11] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, “Mastering the game of Go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[12] S. Gu, E. Holly, T. Lillicrap, and S. Levine, “Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 3389–3396.

[13] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng, “An application of reinforcement learning to aerobatic helicopter flight,” *Adv. Neural Inf. Process. Syst.*, vol. 19, p. 1, 2007.

[14] J. Hwangbo, I. Sa, R. Siegwart, and M. Hutter, “Control of a Quadrotor With Reinforcement Learning,” *IEEE Robot. Autom. Lett.*, vol. 2, no. 4, pp. 2096–2103, Oct. 2017.

[15] M. Hüttenrauch, A. Šošić, and G. Neumann, “Guided Deep Reinforcement Learning for Swarm Systems,” in *AAMAS 2017 Autonomous Robots and Multirobot Systems (ARMS) Workshop*, 2017.

[16] B. WIDROW and M. E. HOFF, “ADAPTIVE SWITCHING CIRCUITS,” *Tech. report, STANFORD Electron. LABS*, 1960.

[17] K. J. Friston, G. Tononi, G. N. Reeke, O. Sporns, and G. M. Edelman, “Value-dependent selection in the brain: simulation in a synthetic neural model.,” *Neuroscience*, vol. 59, no. 2, pp. 229–43, Mar. 1994.

[18] A. G. Barto and M. Duff, “Monte Carlo matrix inversion and reinforcement learning,” *Adv. Neural Inf. Process. Syst.*, p. 687, 1994.

[19] W. Schultz, P. Dayan, and P. R. Montague, “A neural substrate of prediction and reward,” *Science (80-.).*, vol. 275, pp. 1593–1599, 1997.

[20] J. P. O’Doherty, P. Dayan, K. Friston, H. Critchley, and R. J. Dolan, “Temporal Difference Models and Reward-Related Learning in the Human Brain,” *Neuron*, vol. 38, pp. 329–337, 2003.

[21] B. Seymour, J. P. O’Doherty, P. Dayan, M. Koltzenburg, A. K. Jones, R. J. Dolan, K. J. Friston, and R. S. Frackowiak, “Temporal difference models describe higher-order learning in humans,” *Nature*, vol. 429, no. 6992, pp. 664–667, Jun. 2004.

[22] J. P. O’Doherty, P. Dayan, J. Schultz, R. Deichmann, K. Friston, and R. J. Dolan, “Dissociable roles of ventral and dorsal striatum in instrumental conditioning,” *Science (80-.).*, vol. 304, pp. 452–454, 2004.

[23] I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman, “The successor representation in human reinforcement learning,” *Nat. Hum. Behav.*, vol. 1, no. 9, pp. 680–692, Sep. 2017.

[24] N. D. Daw, J. P. O’Doherty, P. Dayan, B. Seymour, and R. J. Dolan, “Cortical substrates for exploratory decisions in humans.,” *Nature*, vol. 441, no. 7095, pp. 876–9, Jun. 2006.

[25] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen, “Humans use directed and random exploration to solve the explore–exploit dilemma.,” *J. Exp. Psychol. Gen.*, vol. 143, no. 6, pp. 2074–2081, Dec. 2014.

[26] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv Prepr. arXiv1509.02971*, 2015.

[27] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. JMLR.org, p. I-387, 2014.

[28] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2016.

[29] K. Doya, “Reinforcement Learning in Continuous Time and Space,” *Neural Comput.*, vol. 12, no. 1, pp. 219–245, Jan. 2000.

[30] W. D. Smart, W. D. Smart, and L. P. Kaelbling, “Practical Reinforcement Learning in Continuous Spaces,” pp. 903--910, 2000.

[31] H. van Hasselt and M. A. Wiering, “Reinforcement Learning in Continuous Action Spaces,” in *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, 2007, pp. 272–279.

[32] K. J. Friston, A. D. Redish, and J. A. Gordon, “Computational Nosology and Precision Psychiatry.,” *Comput. psychiatry (Cambridge, Mass.)*, vol. 1, pp. 2–23, 2017.

[33] P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan, “Computational psychiatry.,” *Trends Cogn. Sci.*, vol. 16, no. 1, pp. 72–80, Jan. 2012.

[34] X.-J. Wang and J. H. Krystal, “Computational Psychiatry,” *Neuron*, vol. 84, no. 3, pp. 638–654, Nov. 2014.

[35] A. V. Apkarian, M. N. Baliki, and P. Y. Geha, “Towards a theory of chronic pain,” *Prog. Neurobiol.*, vol. 87, no. 2, pp. 81–97, Feb. 2009.

[36] J. W. Vlaeyen and S. J. Linton, “Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art.,” *Pain*, vol. 85, no. 3, pp. 317–32, Apr. 2000.

[37] S. Zhang, H. Mano, M. Lee, W. Yoshida, M. Kawato, T. W. Robbins, and B. Seymour, “The control of tonic pain by active relief learning,” *Elife*, vol. 7, Feb. 2018.

[38] M. N. Baliki, P. Y. Geha, H. L. Fields, and A. V. Apkarian, “Predicting Value of Pain and Analgesia: Nucleus Accumbens Response to Noxious Stimuli Changes in the Presence of Chronic Pain,” *Neuron*, vol. 66, no. 1, pp. 149–160, Apr. 2010.

[39] C. Büchel, S. Geuter, C. Sprenger, and F. Eippert, “Placebo Analgesia: A Predictive Coding Perspective,” *Neuron*, vol. 81, no. 6, pp. 1223–1239, Mar. 2014.

[40] A. Norbury, T. W. Robbins, and B. Seymour, “Value generalization in human avoidance learning,” *Elife*, vol. 7, May 2018.

[41] S. Elfwing, *Embodied evolution of learning ability*. Kungliga Tekniska högskolan, 2007.

[42] F. Mondada and D. Floreano, “Evolution of neural control structures: some experiments on mobile robots,” *Rob. Auton. Syst.*, vol. 16, no. 2–4, pp. 183–195, Dec. 1995.

[43] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg, “Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective,” *IEEE Trans. Auton. Ment. Dev.*, vol. 2, no. 2, pp. 70–82, Jun. 2010.

[44] S. Kakade and P. Dayan, “Dopamine: generalization and bonuses.,” *Neural Netw.*, vol. 15, no. 4–6, pp. 549–59.

[45] W. Yoshida, R. J. Dolan, and K. J. Friston, “Game Theory of Mind,” *PLoS Comput. Biol.*, vol. 4, no. 12, p. e1000254, Dec. 2008.

[46] B. Seymour, N. Daw, P. Dayan, T. Singer, and R. Dolan, “Differential encoding of losses and gains in the human striatum.,” *J. Neurosci.*, vol. 27, no. 18, pp. 4826–31, May 2007.

[47] E. Eldar, T. U. Hauser, P. Dayan, and R. J. Dolan, “Striatal structure and function predict individual biases in learning to avoid pain.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 17, pp. 4812–7, Apr. 2016.

[48] S. Elfwing and B. Seymour, “Parallel reward and punishment control in humans and robots: Safe reinforcement learning using the MaxPain algorithm,” in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2017, pp. 140–147.

[49] L. Piccolo, F. D. Libera, A. Bonarini, B. Seymour, and H. Ishiguro, “Pain and self-preservation in autonomous robots: From neurobiological models to psychiatric disease,” in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2017, pp. 263–270.

[50] P. D. Máté Lengyel, “Hippocampal Contributions to Control: The Third Way,” in *Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 889–896.

[51] N. D. Daw, Y. Niv, and P. Dayan, “Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control.,” *Nat. Neurosci.*, vol. 8, no. 12, pp. 1704–11, Dec. 2005.

[52] S. W. Lee, S. Shimojo, and J. P. O’Doherty, “Neural Computations Underlying Arbitration between Model-Based and Model-free Learning,” *Neuron*, vol. 81, no. 3, pp. 687–699, Feb. 2014.

[53] J. X. Wang, Z. Kurth-Nelson, D. Kumaran, D. Tirumala, H. Soyer, J. Z. Leibo, D. Hassabis, and M. Botvinick, “Prefrontal cortex as a meta-reinforcement learning system,” *Nat. Neurosci.*, vol. 21, no. 6, pp. 860–868, Jun. 2018.

[54] T. E. J. Behrens, M. W. Woolrich, M. E. Walton, and M. F. S. Rushworth, “Learning the value of information in an uncertain world,” *Nat. Neurosci.*, vol. 10, no. 9, pp. 1214–1221, Sep. 2007.

[55] S. W. Lee, J. P. O’Doherty, and S. Shimojo, “Neural Computations Mediating One-Shot Learning in the Human Brain,” *PLoS Biol.*, vol. 13, no. 4, 2015.

[56] E. Payzan-LeNestour and P. Bossaerts, “Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings,” *PLoS Comput. Biol.*, vol. 7, no. 1, p. e1001048, Jan. 2011.

[57] D. Kumaran, D. Hassabis, and J. L. McClelland, “What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated.,” *Trends Cogn. Sci.*, vol. 20, no. 7, pp. 512–534, Jul. 2016.

[58] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, “Neuroscience-Inspired Artificial Intelligence.,” *Neuron*, vol. 95, no. 2, pp. 245–258, Jul. 2017.

[59] S. M. Fleming, E. J. van der Putten, and N. D. Daw, “Neural mediators of changes of mind about perceptual decisions,” *Nat. Neurosci. 2018 214*, vol. 21, no. 4, p. 617, Mar. 2018.

[60] B. De Martino, S. M. Fleming, N. Garrett, and R. J. Dolan, “Confidence in value-based choice.,” *Nat. Neurosci.*, vol. 16, no. 1, pp. 105–10, Jan. 2013.

[61] D. Bang and S. M. Fleming, “Distinct encoding of decision confidence in human medial prefrontal cortex.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 23, pp. 6082–6087, May 2018.

[62] F. Meyniel and S. Dehaene, “Brain networks for confidence weighting and hierarchical inference during probabilistic learning.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 19, pp. E3859–E3868, May 2017.

[63] S. M. Fleming and H. C. Lau, “How to measure metacognition,” *Front. Hum. Neurosci.*, vol. 8, p. 443, Jul. 2014.

[64] R. van den Berg, A. Zylberberg, R. Kiani, M. N. Shadlen, and D. M. Wolpert, “Confidence Is the Bridge between Multi-stage Decisions.,” *Curr. Biol.*, vol. 26, no. 23, pp. 3157–3168, Dec. 2016.

[65] L. Li, M. L. Littman, T. J. Walsh, and A. L. Strehl, “Knows what it knows: a framework for self-aware learning,” *Mach. Learn.*, vol. 82, no. 3, pp. 399–443, Mar. 2011.

[66] Z. Wang and M. E. Taylor, “Improving Reinforcement Learning with Confidence-Based Demonstrations,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 3027–3033.

[67] J. B. Hamrick, A. J. Ballard, R. Pascanu, O. Vinyals, N. Heess, and P. W. Battaglia, “Metacontrol for Adaptive Imagination-Based Optimization,” in *ICLR*, 2017.

[68] S. Chiappa, S. Racaniere, D. Wierstra, and S. Mohamed, “Recurrent Environment Simulators,” Apr. 2017.

[69] A. Banino, C. Barry, B. Uria, C. Blundell, T. Lillicrap, P. Mirowski, A. Pritzel, M. J. Chadwick, T. Degris, J. Modayil, G. Wayne, H. Soyer, F. Viola, B. Zhang, R. Goroshin, N. Rabinowitz, R. Pascanu, C. Beattie, S. Petersen, A. Sadik, S. Gaffney, H. King, K. Kavukcuoglu, D. Hassabis, R. Hadsell, and D. Kumaran, “Vector-based navigation using grid-like representations in artificial agents,” *Nature*, vol. 557, no. 7705, pp. 429–433, May 2018.

[70] D. Kim and S. W. Lee, “Model-based BCI : A novel brain-computer interface framework for reading out learning strategies underlying choices,” in *IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC)*, 2018, pp. 1–5.

[71] M. R. Song and S. W. Lee, “Meta BCI : Hippocampus-Striatum Network Inspired Architecture Towards Flexible BCI,” in *The 6th international winter conference on Brain-computer interface (IEEE BCI 2018)*, 2018, pp. 0–2.

[72] S. Yi, J. Lee, and S. W. Lee, “Maximally separating and correlating model-based and model-free reinforcement learning,” in *Computational and Systems Neuroscience (COSYNE)*, 2018.

[73] C. H. Lee, S. Y. Heo, and S. W. Lee, “Designing an Experiment without a Human Experimenter,” in *Computational and Systems Neuroscience (COSYNE)*, 2018.

[74] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare, “The Automation of Science,” *Science (80-.).*, vol. 324, no. 5923, pp. 85–89, Apr. 2009.

[75] Y. Sverchkov and M. Craven, “A review of active learning approaches to experimental design for uncovering biological networks,” *PLOS Comput. Biol.*, vol. 13, no. 6, p. e1005466, Jun. 2017.