Proceedings of the Twelfth Annual Conference on Advances in Cognitive Systems ACS-2025 (21-32)

Toward An Architecture for Robots in the Era of Foundation Models

Mohan Sridharan M.SRIDHARAN@ED.AC.UK
School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

Robots are increasingly being used to assist humans in different application domains. The ready availability of high-fidelity hardware and data has led to the development of deep networks and foundation models that are now considered to be state of the art for many problems in robotics. However, these methods and models are resource-hungry and opaque, and they are known to provide arbitrary decisions in previously unknown situations, whereas practical robot application domains require transparent, multi-step, multi-level decision-making and ad hoc collaboration under resource constraints and open world uncertainty. This essay argues that to leverage the full potential of robots, we need to revisit the fundamental principles that can be traced back to the early pioneers of AI who had a deep understanding of cognition and control in humans. We also need to embed these principles in the architectures we develop for robots, using deep networks as one of many tools that build on this foundation. In addition, this essay briefly illustrates the benefits of this approach by drawing on my work on core problems in robotics such as visual scene understanding and planning, changing-contact manipulation, and ad hoc multiagent collaboration.

1. Motivation and Claims

Robots are increasingly being deployed in application domains such as navigation, healthcare, and manufacturing. Although aided by the availability of high-fidelity hardware, this deployment has largely been due to recent advancements in the form of deep networks and foundation models (FMs) such as Large Language Models (LLMs), Vision Language Models (VLMs), and Vision Language Action models (VLAs), which are considered state of the art for perception, reasoning, manipulation, and interaction problems in robotics (Black et al., 2025; Doshi et al., 2024; Huang et al., 2023; Schick et al., 2023; Zhao et al., 2023). There is a lot of hype (and fear) associated with these methods and models, with claims being made about their "planning", "commonsense reasoning", and "artificial general intelligence" (AGI) capabilities. As a result, we are witnessing a rapid decline in the diversity of formulations being pursued to address problems in robotics.

To motivate the exploration of different formulations, consider the key requirements of integrated robot systems sensing and (inter)acting in the physical world, which include:

- making multi-step, multi-level decisions based on multimodal sensor inputs (e.g., vision, speech, and touch) in the absence of comprehensive domain knowledge;
- operating under open world uncertainty, where the true optimal decisions may be unknowable and probabilities may not always meaningfully model the uncertainty;

- operating under (often strict) constraints on resources such as computation, storage, training examples, and power;
- rapidly and incrementally revising (as needed) existing models for various tasks such as perception, planning, and manipulation; and
- supporting transparency in decision making, and expressing decisions in terms of human concepts such as beliefs and goals to promote understanding.

Next, consider the (by now) well-known characteristics of modern deep network methods and foundation models (Guan et al., 2023; Kambhampati et al., 2024; Lu et al., 2024).

- They are excellent statistical predictors for well-defined tasks, but they are inconsistent and may make arbitrary decisions in truly novel situations;
- Despite the development of architectures with different network structures, they are based on a narrow set of representations and update processes;
- They are resource-hungry systems, making substantial demands in the form of computation, data, storage, and energy; and
- They are *batch learning* systems whose operation remains opaque; even when we can attribute decisions to specific nodes or layers, we are often unable to ascribe meaning to this finding.

There is thus a fundamental mismatch between the requirements of integrated robot systems and the characteristics of the AI methods currently being developed and used in robotics. Attempts to address this mismatch have led to *hybrid methods*, e.g., neurosymbolic (NeSy) AI methods (Besold et al., 2022; Smet et al., 2023). They have also focused on enhancing autonomy in FMs by developing "Agentic AI" and "Agentic LLMs" (Plaat et al., 2025; Wang et al., 2024). In addition, they have sought to discover cognitive design patterns in LLMs, toward AGI (Wray et al., 2025). In all such work, prior knowledge encoded in logics, probability theory, or other design choices impose constraints on a deep network *backbone*. The associated representational and processing commitments continue to limit expressivity, efficiency, transparency, and reproducibility, whereas we still have not explored and understood the consequences of a broader set of design choices. Furthermore, they exacerbate the need for large computing centers, leading to a negative impact on sustainability.

This essay builds on a recent paper (Sridharan, 2025) to advocate that we revisit some key principles that can be traced back to the early pioneers of AI, and are relevant to the design of cognitive architectures (Langley, 2017), but are not fully leveraged in robotics research (Section 2). It also describes how embedding these principles enables the exploration of a broader space of choices in the design of robot architectures, and summarizes the corresponding benefits (Section 3).

2. Key Principles

The early pioneers of AI were deeply inspired by, contributed to, and had a sound understanding of related disciplines such as Psychology, Neuroscience, and Philosophy. Much of their work in AI was inspired by insights into *natural intelligence*, i.e., cognition and control in humans and other biological systems, leading to observations such as:

- Human behavior is jointly determined by internal cognitive processes and the environment. We jointly explore the underlying perception, reasoning, control, and learning problems using *different representations and processes at different abstractions* (Sloman, 2012; Turing, 1952), automatically directing *attention* to relevant representations and processes as needed (Broadbent, 1957; Triesman & Gelade, 1980).

- Unlike the "batch learning" and optimization approach currently prevalent in AI and other disciplines, humans acquire skills *incrementally*, *interactively*, and *compositionally* through *adaptive satisficing* under resource constraints and open world uncertainty; humans seek to make *rational* decisions instead of optimal ones (Simon, 1956; Gigerenzer, 2021).

- Human skills, particularly our sensorimotor skills, have evolved over a long time for some very hard and specific set of engineering problems. Any attempt to replicate these skills in robots needs to pursue an integrated systems approach (Minsky, 1986; Moravec, 1990); just replicating some of our hardware, e.g., our arms and hands, will not lead to the desired sensorimotor capabilities, e.g., dexterous robot manipulation, for a different set of tasks.

These observations do not preclude the use of deep network or FMs; in fact, some of them have been (re)discovered and used to improve the performance of deep networks. Instead, they direct us to focus on certain key principles in the design of robot architectures, with deep networks being one of many available tools. Here, we focus on three sets of such principles.

1. **Refinement, Compositionality, Attention.** The first set of principles advocate representing actions and change in the domain in the form of transition diagrams at different abstractions, with the fine(r)-granularity description(s) being a *refinement* of the coarse(r)-granularity description(s). Refinement is also related to *compositionality*, the hierarchical representation of knowledge at different resolutions. These principles have played a key role in computing and other disciplines over many decades (Fodor, 1975; Freeman & Pfenning, 1991; Dietterich, 1998). Research has identified that such representations lead to a good computational model for human cognition (Knoblich & Flach, 2001; Piantadosi et al., 2016), and for tasks in computer vision and robotics (Fidler & Leonardis, 2007; Zabkar & Leonardis, 2016).

   To truly adapt these principles to robotics, we need to move beyond discovering decompositions in deep networks (Prasad et al., 2024) or encoding these principles in deep network architectures. Instead, we need to establish a suitable representation (i.e., a *vocabulary*) and processes that update this representation at each level of abstraction, and define a formal relationship between the abstractions. The relevant representations and processes can then be chosen automatically for any given task and domain using the principle of *selective attention* (Broadbent, 1957) and decision heuristics (more information below). Even a limited exploration of attention has led to a performance improvement with deep networks (Doshi et al., 2024). Furthermore, if we expand the range of representations and update processes, it will enable the robot to acquire domain knowledge and make decisions based on different information sources. It will also enable the robot to support different descriptions of its decisions such that they make contact with human concepts such as goals and beliefs.

2. **Ecological Rationality (ER) and Decision Heuristics.** The second set of principles build on Herb Simon's definition of *Bounded Rationality* (Simon, 1956) and the related algorithmic theory of heuristics (Gigerenzer, 2020). Unlike the focus on optimal search in many disciplines (e.g., finance, computing) in the presence of *risk* over a set of known scenarios, ER studies decision making under *open world uncertainty*, i.e., when the space of possible scenarios is not known in advance. It characterises the behavior of a human or an AI system as a joint function of the internal cognitive processes and the environment, using *adaptive satisficing* and *decision heuristics* such as tallying, sequential search, and fast and frugal (FF) trees to make rational decisions instead of optimal ones.

   Unlike the use of heuristics as a "hack" or to explain biases (e.g., in the *heuristics and biases* program in Psychology), ER considers decision heuristics as a strategy to ignore part of the information in order to make decisions more quickly, frugally, and accurately than complex methods with many free parameters (Gigerenzer & Gaissmaier, 2011). Also, unlike modern AI methods that are largely *prescriptive* (describing what should be done), it is both *descriptive* (describing what people or agents do) and prescriptive. It uses an adaptive toolbox of classes of decision heuristics, and an algorithmic approach involving out-of-sample and out-of-population testing to identify heuristics that match domain characteristics. Such decision heuristics are well-suited to make decisions under open world uncertainty, where optimal decisions are unknowable and probabilities are not always a good model of the uncertainty. Their design also automatically supports process-level explanations of the decisions made.

3. **Interactive Learning and Memory Consolidation.** The third set of principles jointly refer to different types of learning such as supervised (or unsupervised) learning and learning from reinforcement (Laird et al., 2017). The difference lies in how this learning is achieved. Modern AI systems increasingly focus on learning a single model or policy that determines decisions across different categories, situations, platforms, and/or domains. Such an approach is considered to be essential for *generalization* without realizing that there is a mismatch between the underlying design choices and the desired functional capabilities, creating problems that we then struggle to address. For example, the learned model or policy is hard to understand, explain, or revise in a meaningful manner. Such approaches are appropriate for tasks or domains in which the space of possible options or situations is known a priori and there are no strict resource constraints; they are not really suitable for decision making *in the wild*, i.e., under open-world uncertainty (Katsikopoulos et al., 2021a).

   Interactive learning, on the other hand, focuses on learning as needed to adapt to any given domain and set of tasks. It advocates reasoning with prior knowledge and decision heuristics to trigger, inform, and constrain the learning. It also enables *cumulative learning* through *memory consolidation*, revising the learned knowledge and discovering high-level (i.e., more abstract) concepts and theories offline (Stickgold, 2005; Wolpert et al., 2011) to update the existing knowledge for subsequent reasoning. Such an approach is known to aid in incremental knowledge acquisition, information storage, and information retrieval in humans (Baddeley, 2012). It also, not surprisingly, leads to simpler models that are amenable to incremental and rapid revisions, even in situations that were previously unknown to the robot.

3. Architectural Examples

This section provides examples of embedding the principles outlined above in robot architectures to address problems in reasoning, control, collaboration, and learning.

Refinement for knowledge representation and reasoning. Refinement of an agent's action theories has been defined using situation calculus (Banihashemi et al., 2018), with a smooth transfer of information and control between two abstractions. However, this work makes the strong assumption of a bisimulation relation between these action theories, which limits expressivity for robot domains. There has also been related work on task and motion planning (TAMP) in robotics (Garrett et al., 2021; Kokel et al., 2023). This work combines discrete-space task planning and continuous-space motion planning at different resolutions, e.g., using first-order propositional logic to compute a sequence of abstract tasks to achieve a given goal, and using probabilistic motion planners (Srivastava et al., 2013) to compute a sequence of movement actions to complete each task in the abstract task plan. This can also involve learning feature-based state and action abstractions towards generalized TAMP for continuous control tasks (Curtis et al., 2022). However, existing methods do not fully: (a) support the bidirectional flow of relevant information between the different abstractions; (b) handle uncertainty, particularly the effect of non-stationarity and future state uncertainty on the associated models; and (c) address the discontinuities in the interaction dynamics, i.e., the sudden changes in forces and the resultant acceleration experienced by the robot when it makes or breaks contact with objects and surfaces (Garrett et al., 2021).

The limitations mentioned above can be attributed to not leveraging the principles outlined above in building an integrated (cognitive) architecture that jointly addresses the underlying reasoning and learning problems. For example, we developed a refinement-based architecture that supported different representations (logics, probabilities) and processes (non-monotonic logical reasoning, probabilistic sequential decision making) for reasoning with any given domain's transition diagrams at two different resolutions (Sridharan et al., 2019). The fine-resolution description was defined as a refinement of the coarse-resolution description, which included theories of intention (Gomez et al., 2021), affordance (Langley et al., 2018; Sridharan et al., 2017), and explainable agency (Langley et al., 2017; Sridharan & Meadows, 2019; Sridharan, 2024). For any given goal, each abstract action in the plan created by logical reasoning in the coarse resolution was implemented as a sequence of fine-resolution transitions obtained by automatically identifying and reasoning probabilistically with the relevant part of the fine-resolution description. In addition, the use of decision heuristics helped learn and revise the model parameters to achieve more reliable and efficient operation compared with baselines based on deep networks or reasoning with comprehensive domain knowledge. Furthermore, we can consider including other representations and processes; we could obtain latent embeddings of perceptual inputs from deep networks and use a developmental learning approach to map target actions (e.g., grasp and push) to transitions between states defined in the latent space (Juett & Kuipers, 2019).

Decision heuristics for multiagent collaboration and robot manipulation. Although ER and decision heuristics have provided good performance on prediction problems in application domains such as finance, healthcare, and law (Brighton & Gigerenzer, 2012; Durbach et al., 2020; Gigerenzer, 2016; Katsikopoulos et al., 2021b), there is hardly any use of these methods in robot architectures, except in some related work in the cognitive systems community (Langley & Katz, 2022). This lack of uptake is potentially because the successes of decision heuristics do not receive the attention they deserve, and because their inherent simplicity makes researchers doubt their suitability for addressing complex practical problems.

As one example of the use of decision heuristics, consider the problem of agents (i.e., AI systems, robots, and humans) collaborating with other agents without prior coordination, i.e., *ad hoc teamwork* (AHT) (Mirsky et al., 2022). Methods considered state of the art for AHT use a large dataset and/or FMs to model the behavior of different agent types and to determine the ad hoc (AI) agent's behavior (Rahman et al., 2021; Liu et al., 2024). As discussed in Section 1, such methods do not support transparency or rapid adaptation to new situations, and the necessary resources (e.g., training examples, computation) are often not available in practical domains. We instead adapted our refinement-based architecture to pose AHT as a joint reasoning and learning problem. Each ad hoc agent chose its actions based on non-monotonic logical reasoning with prior domain knowledge (action theories at two abstractions) and an ensemble of FF trees learned rapidly to predict the behavior of other agents. We experimentally demonstrated the ability to collaborate in complex environments, adapting to previously unknown changes (e.g., in the environment or team composition) and providing better performance than state of the art baselines while using orders of magnitude fewer resources (e.g., 5K instead of 1M examples) (Dodampegama & Sridharan, 2023).

As a very different example of the use of decision heuristics, consider changing-contact robot manipulation, which involves a robot making and breaking contacts with different objects and surfaces; many robot and human manipulation tasks are such changing-contact tasks. The dynamics of these tasks are piecewise continuous, with abrupt transitions (i.e., sudden changes in force and acceleration) that can damage the robot or the domain objects. Unlike existing methods that attempt to explore the space of possible transitions in advance, and pose the problem of smooth motion as an (offline) optimization problem or learning problem (Khader et al., 2020), we drew inspiration from insights into human motor control (Kawato, 1999; Flanagan et al., 2003). Specifically, we enabled the robot to use a single initial demonstration of the desired motion trajectory, or run-time observations, to rapidly learn and revise simple *forward models* that predict the end-effector sensor observations in each upcoming time step. During run-time, any mismatch between the predicted values and the actual sensor measurements incrementally and automatically revised the predictive models and the gain parameters of a force-motion PD control law. Using experiments conducted in different simulation domains and on a physical robot manipulator, we demonstrated the ability to provide smooth motion during changing-contact manipulation tasks with changes in surfaces and contacts that the robot was not aware of before (Sidhik et al., 2024).

Interactive learning for visual scene understanding and assistive robotics. To further illustrate the benefits of leveraging the interplay between reasoning and learning in robot architectures that embed the outlined principles, consider two other examples. These examples also illustrate how modern deep networks and FMs can be used effectively in such architectures.

The first example focuses on vision-based scene understanding, vision-based planning, and question answering, which are fundamental problems in computer vision and robotics. Methods considered state of the art for these problems are based on deep networks and FMs that are trained or tuned, for example, with a large dataset of images, potential questions, and answers to these questions. We, on the other hand, developed a refinement-based architecture to determine the occlusion of objects and the stability of object structures in images, arrange objects in desired configurations, and to answer questions about the decisions made. With this architecture, the robot first attempted to make the desired decisions (e.g., about stability and occlusion of objects) through non-monotonic logical reasoning with generic domain knowledge available a priori. When the robot could not make a decision (or made an incorrect decision on training examples), learning was triggered. The robot then automatically identified examples of relevant images and regions in these images to be used for learning models that were used to make the desired decisions. In addition, the examples used for learning were also used as input for decision tree induction driven by decision heuristics to acquire new knowledge (e.g., objects, actions, axioms) and consolidate existing knowledge to be used for subsequent reasoning. We experimentally demonstrated: (a) better performance than baselines based purely on deep networks, while using orders of magnitude fewer resources; (b) faster and more effective training of deep networks by using only the relevant examples; and (c) performance improvement directly attributable to reasoning and learning bootstrapping off of each other (Riley & Sridharan, 2019; Sridharan & Mota, 2023). We also demonstrated the ability to provide relational descriptions on-demand at different abstractions as explanations in response to different types of questions (causal, contrastive, counterfactual) (Sridharan, 2024).

The second example illustrates the effective use of FMs in architectures based on the principles outlined above. Specifically, we developed an architecture that enabled an *embodied (AI) agent*[1] to collaborate with other agents in completing assigned tasks in a home environment. Instead of making unsubstantiated and incorrect claims about the planning or commonsense reasoning capabilities of FMs, our architecture was similar (in spirit) to the work on *LLM-Modulo frameworks* (Guan et al., 2023; Kambhampati et al., 2024). It used an LLM to provide a generic prediction about the sequence of tasks likely to be assigned in the near future based on any recent history of task execution. The current and anticipated tasks were considered as joint goals by the robot, which incorporated decision heuristics with planning methods based on logics to compute action sequences that would enable it to achieve these goals in collaboration with the other agents. We experimentally demonstrated substantial improvement in the accuracy and computational efficiency of task completion compared with baselines that just used FMs (or deep networks) or knowledge-based reasoning, and baselines that did not reason about anticipated tasks tasks (Singh et al., 2025; Fu et al., 2025).

In summary, the objective of this essay was to promote appreciation of some fundamental principles that can be traced back to the early pioneers of AI, but are not being leveraged in the design of modern architectures for robots. Since many of the corresponding choices also arise in the design of cognitive systems and architectures, we hope that the examples provided above will encourage researchers in this community to explore and understand the capabilities of different robot architectures that embed these principles, leading to more robust solutions for open problems in robotics.

Acknowledgements

The ideas described in this paper were informed by research threads pursued in collaboration with Hasra Dodampegama, Michael Gelfond, Gerd Gigerenzer, Rocio Gomez, Konstantinos Kat-

1. ”Embodied agent” refers to an agent in a physics-based realistic simulation or in the physical world.

sikopoulos, Pat Langley, Ales Leonardis, Ben Meadows, Tiago Mota, Heather Riley, Saif Sidhik, Özgür Şimşek, Jeremy Wyatt, and Shiqi Zhang. This work was supported in part by the U.S. ONR Awards N00014-13-1-0766, N00014-17-1-2434, N00014-20-1-2390, and N00014-24-1-2737, AOARD award FA2386-16-1-4071, and U.K. EPSRC award EP/S032487/1. All conclusions described in this paper are those of the author alone.

References

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*, 1–29.

Banihashemi, B., Giacomo, G. D., & Lesperance, Y. (2018). Abstraction of Agents Executing Online and their Abilities in Situation Calculus. *International Joint Conference on Artificial Intelligence*. Stockholm, Sweden.

Besold, T. R., et al. (2022). Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. In P. Hitzler & M. K. Sarker (Eds.), *Neuro-Symbolic Artificial Intelligence: The State of the Art*. IOS Press, Amsterdam.

Black, K., et al. (2025). π_0: A Vision-Language-Action Flow Model for General Robot Control. *Robotics: Science and Systems Conference*. Los Angeles, USA.

Brighton, H., & Gigerenzer, G. (2012). How Heuristics Handle Uncertainty? In *Ecological Rationality: Intelligence in the World*. New York: Oxford University Press.

Broadbent, D. E. (1957). A Mechanical Model for Human Attention and Immediate Memory. *Psychology Review*, *64*, 205–215.

Curtis, A., Silver, T., Tenenbaum, J. B., Lozano-Perez, T., & Kaelbling, L. P. (2022). Discovering State and Action Abstractions for Generalized Task and Motion Planning. *AAAI Conference on Artificial Intelligence* (pp. 5377–5384).

Dietterich, T. (1998). The MAXQ Method for Hierarchical Reinforcement Learning. *International Conference on Machine Learning (ICML)*.

Dodampegama, H., & Sridharan, M. (2023). Knowledge-based Reasoning and Learning under Partial Observability in Ad Hoc Teamwork. *Theory and Practice of Logic Programming*, *23*, 696–714.

Doshi, R., Walke, H., Mees, O., Dasai, S., & Levine, S. (2024). Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion, and Aviation. *International Conference on Robot Learning*. Munich, Germany.

Durbach, I. N., Algorta, S., Kantu, D. K., Katsikopoulos, K. V., & Simsek, O. (2020). Fast and Frugal Heuristics for Portfolio Decisions with Positive Project Interactions. *Decision Support Systems*, *138*.

Fidler, S., & Leonardis, A. (2007). Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. *International Conference on Computer Vision and Pattern Recognition*.

Flanagan, J. R., Vetter, P., Johansson, R. S., & Wolpert, D. M. (2003). Prediction Precedes Control in Motor Learning. *Current Biology*, *13*, 146–150.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Freeman, T., & Pfenning, F. (1991). Refinement Types for ML. *ACM SIGPLAN Conference on Programming Language Design and Implementation* (pp. 268–277). Toronto, Canada.

Fu, T., Jauw, B., & Sridharan, M. (2025). Combining LLM, Non-monotonic Logical Reasoning, and Human-in-the-loop Feedback in an Assistive AI Agent. *IEEE International Conference on Robot and Human Interactive Communication*. Eindhoven, The Netherlands.

Garrett, C. R., Chitnis, R., Holladay, R., Kim, B., Silver, T., Kaelbling, L. P., & Lozano-Perez, T. (2021). Integrated Task and Motion Planning. *Annual Review of Control, Robotics, and Autonomous Systems*, *4*, 265–293.

Gigerenzer, G. (2016). *Towards a Rational Theory of Heuristics*, (pp. 34–59). London: Palgrave Macmillan UK.

Gigerenzer, G. (2020). What is Bounded Rationality? In *Routledge Handbook of Bounded Rationality*. Routledge.

Gigerenzer, G. (2021). Embodied Heuristics. *Frontiers in Psychology*, *12*, 1–12.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*, 451–482.

Gomez, R., Sridharan, M., & Riley, H. (2021). What do you really want to do? Towards a Theory of Intentions for Human-Robot Collaboration. *Annals of Mathematics and Artificial Intelligence, special issue on commonsense reasoning*, *89*, 179–208.

Guan, L., Valmeekam, K., Sreedharan, S., & Kambhampati, S. (2023). Leveraging Pre-trained Large Language Models to Construct and Utilize World Models for Model-based Task Planning. *International Conference on Neural Information Processing Systems*. New Orleans, USA.

Huang, W., et al. (2023). Grounded Decoding: Guiding Text Generation with Grounded Models for Robot Control. *International Conference on Neural Information Processing Systems*. New Orleans, USA.

Juett, J., & Kuipers, B. (2019). Learning and Acting in Perpersonal Space: Moving, Reaching, and Grasping. *Frontiers in Neurorobotics*, *13*, 718–727.

Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L. P., & Murthy, A. (2024). Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. *International Conference on Machine Learning*. Vienna, Austria.

Katsikopoulos, K., Simsek, O., Buckmann, M., & Gigerenzer, G. (2021a). *Classification in the Wild: The Science and Art of Transparent Decision Making*. MIT Press.

Katsikopoulos, K., Simsek, O., Buckmann, M., & Gigerenzer, G. (2021b). Transparent modeling of influenza incidence: Big data or a single data point from psychological theory? *International Journal of Forecasting*.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, *6*, 718–727.

Khader, S. A., Yin, H., Falco, P., & Kragic, D. (2020). Data-efficient model learning and prediction for contact-rich manipulation tasks. *IEEE Robotics and Automation Letters*, *5*, 4321–4328.

Knoblich, G., & Flach, R. (2001). Predicting the Effects of Actions: Interactions of Perception and Action. *Psychological Sciences*, *12*, 467–472.

Kokel, H., Natarajan, S., Ravindran, B., & Tadepalli, P. (2023). RePReL: A Unified Framework for Integrating Relational Planning and Reinforcement Learning for Effective Abstraction in Discrete and Continuous Domains. *Neural Computing and Applications*, *35*, 16877–16892.

Laird, J. E., et al. (2017). Interactive Task Learning. *IEEE Intelligent Systems*, *32*, 6–21.

Langley, P. (2017). Progress and Challenges in Research on Cognitive Architectures. *The Thirty-first AAAI Conference on Artificial Intelligence*. San Francisco, USA.

Langley, P., & Katz, E. P. (2022). Motion Planning and Continuous Control in a Unified Cognitive Architecture. *Annual Conference on Advances in Cognitive Systems*. Arlington, VA.

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable Agency for Intelligent Autonomous Systems. *Innovative Applications of Artificial Intelligence*. San Francisco, USA.

Langley, P., Sridharan, M., & Meadows, B. (2018). Representation, Use, and Acquisition of Affordances in Cognitive Systems. *AAAI Spring Symposium on Integrating Representation, Reasoning, Learning and Execution for Goal Directed Autonomy*. Stanford, USA.

Liu, X., Li, P., Yang, W., Guo, D., & Liu, H. (2024). Leveraging Large Language Model for Heterogeneous Ad Hoc Teamwork Collaboration. *Robotics: Science and Systems Conference*. Delft, Netherlands.

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2024). Are Emergent Abilities in Large Language Models just In-Context Learning? *Annual Meeting of the Association for Computational Linguistics* (pp. 5098–5139). Bangkok, Thailand.

Minsky, M. L. (1986). *The Society of Mind*. Simon and Schuster.

Mirsky, R., Carlucho, I., Rahman, A., Fosong, E., Macke, W., Sridharan, M., Stone, P., & Albrecht, S. (2022). A Survey of Ad Hoc Teamwork: Definitions, Methods, and Open Problems. *European Conference on Multiagent Systems*.

Moravec, H. P. (1990). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The Logical Primitives of Thought: Empirical Foundations for Compositional Cognitive Models. *Psychological Review*, *123*, 392–424.

Plaat, A., van Duijn, M., van Stein, N., Preuss, M., van der Putten, P., & Batenburg, K. J. (2025). *Agentic Large Language Models: A Survey*. Technical report, arXiv: `https://arxiv.org/abs/2503.23037`.

Prasad, A., Koller, A., Hartmann, M., Clark, P., Sabharwal, A., Bansal, M., & Khot, T. (2024). ADaPT: As-Needed Decomposition and Planning with Language Models. *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 4226–4252). Mexico City, Mexico: Association for Computational Linguistics.

Rahman, M. A., Hopner, N., Christianos, F., & Albrecht, S. V. (2021). Towards open ad hoc teamwork using graph-based policy learning. *International Conference on Machine Learning* (pp. 8776–8786).

Riley, H., & Sridharan, M. (2019). Integrating Non-monotonic Logical Reasoning and Inductive Learning With Deep Learning for Explainable Visual Question Answering. *Frontiers in Robotics and AI, special issue on Combining Symbolic Reasoning and Data-Driven Learning for Decision-Making*, *6*, 20.

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems*. New Orleans, USA.

Sidhik, S., Sridharan, M., & Ruiken, D. (2024). An Adaptive Framework for Trajectory Following in Changing-Contact Robot Manipulation Tasks. *Robotics and Autonomous Systems*, *181*, 1–21.

Simon, H. A. (1956). Rational Choice and the Structure of the Environment. *Psychological Review*, *63*, 129–138.

Singh, S., Swaminathan, K., Dash, N., Singh, R., Banerjee, S., Sridharan, M., & Krishna, M. (2025). AdaptBot: Combining LLM with Knowledge Graphs and Human Input for Generic-to-Specific Task Decomposition and Knowledge Refinement. *International Conference on Robotics and Automation (ICRA)*. Atlanta, USA.

Sloman, A. (2012). Meta-morphogenesis and the Creativity of Evolution. *Workshop on Computational Creativity, Concept Invention, and General Intelligence at ECAI*. Montpellier, France.

Smet, L. D., Martires, P. Z. D., Manhaeve, R., Marra, G., Kimmig, A., & Readt, L. D. (2023). Neural Probabilistic Logic Programming in Discrete-Continuous Domains. *International Conference on Uncertainty in Artificial Intelligence* (pp. 529–538).

Sridharan, M. (2024). Integrated Knowledge-based Reasoning and Data-driven Learning for Explainable Agency in Robotics. In *Explainable Agency in Artificial Intelligence: Research and Practice*. CRC Press.

Sridharan, M. (2025). Back to the Future of Integrated Robot Systems. *AAAI Conference on Artificial Intelligence*. Philadelphia, US.

Sridharan, M., Gelfond, M., Zhang, S., & Wyatt, J. (2019). REBA: A Refinement-Based Architecture for Knowledge Representation and Reasoning in Robotics. *Journal of Artificial Intelligence Research*, *65*, 87–180.

Sridharan, M., & Meadows, B. (2019). Towards a Theory of Explanations for Human-Robot Collaboration. *Kunstliche Intelligenz*, *33*, 331–342.

Sridharan, M., Meadows, B., & Gomez, R. (2017). What can I not do? Towards an Architecture for Reasoning about and Learning Affordances. *International Conference on Automated Planning* *and Scheduling*. Pittsburgh, USA.

Sridharan, M., & Mota, T. (2023). Towards Combining Commonsense Reasoning and Knowledge Acquisition to Guide Deep Learning. *Autonomous Agents and Multi-Agent Systems*, *37*.

Srivastava, S., Riano, L., Russell, S., & Abbeel, P. (2013). Using Classical Planners for Tasks with Continuous Operators in Robotics. *International Conference on Automated Planning and Scheduling (ICAPS)*. Rome, Italy.

Stickgold, R. (2005). Sleep-dependent Memory Consolidation. *Nature*, *437*, 1272–1278.

Triesman, A. M., & Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology*, *12*, 97–136.

Turing, A. (1952). The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London B*, *237*, 37–72.

Wang, L., et al. (2024). A Survey on Large Language Model-based Autonomous Agents. *Frontiers of Computer Science*, *18*, 1–26.

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of Sensorimotor Learning. *Nature Reviews Neuroscience*, *12*, 739–751.

Wray, R. E., Kirk, J. R., & Laird, J. E. (2025). Applying Cognitive Design Patterns to General LLM Agents. *International Conference on Artificial General Intelligence*.

Zabkar, J., & Leonardis, A. (2016). Motor Memory: Representation, Learning, and Consolidation. *Biologically Inspired Cognitive Architectures*, *16*, 64–74.

Zhao, Z., Lee, W. S., & Hsu, D. (2023). Large Language Models as Commonsense Knowledge for Large-Scale Task Planning. *International Conference on Neural Information Processing Systems*.