




Beyond Mimicry: Toward Lifelong Adaptability in Imitation Learning

Blue Sky Ideas Track

Nathan Gavenski 
King’s College London
London, United Kingdom
nathan.schneider_gavenski@kcl.ac.uk

Felipe Meneguzzi 
University of Aberdeen & PUCRS
Aberdeen, United Kingdom
felipe.meneguzzi@abdn.ac.uk

Odinaldo Rodrigues 
King’s College London
London, United Kingdom
odinaldo.rodrigues@kcl.ac.uk




ABSTRACT

Imitation learning stands at a crossroads: despite decades of progress, current imitation learning agents remain sophisticated memorisation machines, excelling at replay but failing when contexts shift or goals evolve. This paper argues that this failure is not technical but foundational: imitation learning has been optimised for the wrong objective. We propose a research agenda that redefines success from perfect replay to compositional adaptability. Such adaptability hinges on learning behavioural primitives once and recombining them through novel contexts without retraining. We establish metrics for compositional generalisation, propose hybrid architectures, and outline interdisciplinary research directions drawing on cognitive science and cultural evolution. Agents that embed adaptability at the core of imitation learning thus have an essential capability for operating in an open-ended world.

KEYWORDS

Imitation Learning, Generalisation, Compositional Learning

ACM Reference Format:

Nathan Gavenski , Felipe Meneguzzi , and Odinaldo Rodrigues . 2026. Beyond Mimicry: Toward Lifelong Adaptability in Imitation Learning : Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 5 pages.

1 INTRODUCTION

When a child observes an adult stacking blocks, they do not memorise every motion in a trajectory. Instead, they extract reusable behavioural primitives (grasping, lifting, placing) and the compositional rules that govern their combination. Later, when faced with blocks of different shapes, sizes, or materials, a child can build entirely new structures never seen in demonstrations. This compositional understanding, fundamental to human imitation learning (IL), remains conspicuously absent from modern IL systems.

The disconnect between human and artificial imitation reveals a foundational flaw in the field’s approach to learning from demonstrations. Current IL methods excel at reproducing demonstrated trajectories with impressive fidelity [7, 9]. Behavioural cloning (BC) achieves high accuracy on training distributions [19], inverse reinforcement learning infers underlying objectives [8], and adversarial approaches can handle imperfect demonstrations [24]. Yet, these

successes mask a critical limitation: when contexts shift beyond the training distribution performance degrades significantly [10]. An IL agent that masters a navigation layout cannot easily adapt to a configuration change. These failures are not edge cases, but the norm whenever demonstration and evaluation conditions vary. The brittleness stems from optimising for the wrong objective: sample efficiency rather than adaptability. However, no amount of memorisation enables the behaviour required for more nuanced tasks. The real challenge, we argue, is not simply to learn from fewer demonstrations, but to actually understand the compositional structure of tasks embedded in them.

This paper introduces Compositional Repertoire Learning (CRL), a research agenda that redefines success in IL from trajectory reproduction to compositional generalisation. Rather than treating demonstrations as sequences of actions to memorise, CRL requires agents to extract behavioural primitives and compositional rules that can be recombined to solve novel contexts. This shift in orientation of the field is not merely technical but philosophical, fundamentally changing what it means for an agent to imitate successfully. The core insight driving CRL is that demonstrations contain two types of information: surface trajectories and deep compositional structure. Current methods capture only superficial knowledge, learning associations between states and actions. CRL requires agents to capture the structure of the solution, understanding that they must grasp *primitives*, i.e., steps that achieve sub-goals, so they can be recombined according to task-dependent rules. This structural understanding enables agents to generalise far beyond their training distribution, adapting to novel contexts through creative recombination rather than nearest-neighbour retrieval.

2 BACKGROUND

Most IL works model environments as *Markov Decision Processes* (MDPs), which are tuples of the form $\mathcal{M} = \langle S, A, T, r, \gamma \rangle$, where S is the state space, A is the action space, $T : S \times A \rightarrow S$ is a transition function, $r : S \times A \rightarrow \mathbb{R}$ is an immediate reward function, and γ is the discount factor [21].

Mirror Mechanism and Behavioural Cloning Humans possess a neural mechanism that allows them to map observed actions onto their own motor systems [4]. This mirror mechanism extracts the underlying intention of action sequences and, afterwards, adapts and replicates the learned behaviour in similar situations. For example, when watching someone stack blocks, humans learn goal-directed primitives, such as aligning, positioning, grasping, rather than just the actual hand and finger trajectories.

In computer science, a well-known attempt to capture the mirror mechanism is called “*Behavioural Cloning*” [19], where a policy π learns parameters θ such that $\pi_\theta : S \mapsto A$ by optimising: $\arg \min_\theta \sum_{(s,a) \in D_{\text{train}}} \mathbb{E}(a, \pi_\theta(s))$, where D_{train} is a set of teacher π_ψ demonstrations. However, this simplification into supervised learning without capturing the compositional structure leads to a fundamental limitation: mimicry, which we explain next.

Forms of Imitation There are two forms of imitation [6]: *mimicry* and *adaptive behaviour*. Mimicry is associated with *emulation*, i.e., agents retrieve the state in the demonstration that is most similar to the current state and then replay the associated action. This causes *behaviour-seeking mode* [23], associated with two types of failures: compounding errors [20, Ch. 26], where small deviations cascade into large failures, and memorisation [20, Ch. 19], where agents retrieve stored actions rather than computing appropriate responses. Adaptive behaviour generates novel responses while maintaining behavioural intent. Agents understand not just what to do but also why, decomposing demonstrations into reusable primitives and compositional rules, enabling near-optimal responses to unseen states through recombination rather than retrieval. Both forms of imitation are essential. Mimicry works in controlled environments, while adaptive behaviour enables generalisation for more dynamic tasks where variation is the rule rather than the exception.

Limitation of Current Evaluation Current evaluation paradigms fail to measure true generalisation. Current metrics, such as Average Episodic Reward (AER), hide critical failures through averaging. An agent might achieve high AER by perfectly memorising most trajectories while failing on novel situations. Furthermore, agentic applications (sequential decision problems, episodic in nature) create complex dependencies. An agent starting from a state not present in the training dataset might reach familiar territory, then rely on memorisation. These metrics cannot distinguish this trajectory-level memorisation from compositional understanding. Most critically, states not seen during training and testing create hidden failure points invisible during evaluation but catastrophic in deployment. These limitations demand new paradigms that directly measure generalisation under a compositional lens.

Compositional Generalisation Compositional generalisation, borrowed from linguistics [12], provides the theoretical foundation for understanding adaptive capabilities. Like humans understanding novel sentences through grammatical rules, agents must understand behavioural combinations through compositional principles. Compositional generalisation refers to five forms through which a model can generalise: (i) *systematicity*, which is generalisation via systematically recombining known parts and rules; (ii) *productivity*, the ability to extend predictions beyond the length seen in training data (akin to machine learning generalisation); (iii) *substitutivity*, which is generalisation via the ability to replace components with synonyms; (iv) *localism*, whether the model compositions operate locally or globally; and (v) *overgeneralisation*, referring to the model’s ability to pay attention to or being robust to exceptions.

In this work, we are particularly interested in the first three forms: systematicity, productivity, and substitutivity. To exemplify them, we use an excerpt from Kirk et al. [14]: “[For a block-stacking environment], an example of *systematicity* is the ability to stack blocks in new configurations once the basics of block-stacking are

mastered. Similarly, *productivity* can be measured by how many blocks the agent can generalise to, the complexity of blocks, and the complexity of the stacking configuration. *Substitutivity* can be evaluated by the agent’s ability to generalise to blocks of new colours, understanding that the new colour does not affect the physics of the block.” Thus, systematicity, productivity, and substitutivity allow us to capture dimensions of generalisation that go beyond mere accuracy on held-out states.

Related Approaches Beyond compositional generalisation, other noteworthy related approaches inspire CRL. For example, **Goal-conditioned IL** policies ($\pi_\theta \mid s, g$) enable flexible behaviour by conditioning on desired outcomes [13]. Hindsight experience replay relabels failed trajectories with achieved goals [3]. However, goal-conditioning does not address how to deal with novel goal-context combinations. Similarly, **Hierarchical and Skill-based IL** learns reusable behavioural components [18, 22]. These approaches address primitive extraction but typically assume fixed composition rules. CRL extends this by explicitly measuring compositional recombination across contexts.

3 COMPOSITIONAL REPERTOIRE LEARNING

The path from mimicry to adaptability requires mathematical rigour that current IL frameworks lack. This section lays CRL’s foundation by building on three components that transform compositional generalisation from intuition into a measurable quantity: *goals* enable trajectory-level success measurement, *controllable contexts* provide systematic variation for testing, and *deterministic transitions* isolate algorithmic performance from environmental randomness.

3.1 Goal-conditioned Contextual MDPs

The heart of CRL lies in a deceptively simple insight: standard MDPs cannot capture compositional structure because they conflate task definition with environmental dynamics. When everything changes simultaneously, it is impossible to isolate what varies compositionally from what varies arbitrarily. We propose Goal-conditioned Contextual MDPs (GCMDBPs), extending contextual MDPs [11] with explicit goal structure and context metrics.

Definition 3.1. A *Goal-conditioned Contextual MDP* is a tuple $\mathcal{M}_{gc} = \langle S, A, C, G, O, T, \phi, d \rangle$, where S is the state space, A is the action space, C is the set of controllable contexts, O is the set of observations, G is the set of goals, $T : O \times A \mapsto O$ is a deterministic transition function, $\phi : S \times C \mapsto O$ is a projection function that maps state-context pairs to observations, and $d : C \times C \rightarrow \mathbb{R} \geq 0$ is a contextual distance metric.

This formulation departs from standard problems in three critical ways. First, by separating goals from rewards, we enable trajectory-level evaluation capturing compositional generalisation (as defined in Section 2) beyond local action accuracy. Second, by using controllable contexts, GCMDBPs avoid random variation, which obscures compositional structure. Third, by introducing deterministic transition functions, it avoids stochasticity attribution problems.

Explicit goal specifications (declarative success criteria rather than numeric reward signals) provide the missing abstraction layer, defining success at the trajectory level to capture whether agents understand task structure beyond surface mimicry.

Definition 3.2. Let $\tau = (o_0, \dots, o_n)$ be a trajectory, i.e., an ordered sequence of observations produced by the agent’s interactions. G specifies success criteria for the agent and can be of one of the following types. (i) *terminal goals*: the agent succeeds if the final observation in a trajectory satisfies one of the designated goal states; (ii) *landmark goals*: the agent succeeds when observations in a trajectory satisfy all states in one of the designated sets of goal states; or (iii) *ordered landmark goals*: the agent succeeds when observations in a trajectory satisfy all states in one of the designated sequences of goal states (order preserving).

Goals enable deliberate failure analysis [1]. When agents fail, examining reached subgoals reveals whether failures stem from the inability to execute primitives [2], to combine them properly, or to adapt combinations to novel contexts.

Controllable contexts expose generative parameters, creating variation. For example, when agents fail on procedurally generated levels, it is hard to determine if they lack compositional understanding or encountered novel mechanics. Before we introduce them, let us say that two contexts c and c' are equivalent, in symbols $c \simeq c'$, if for all states s , we have that $\phi(s, c) = \phi(s, c')$.

Definition 3.3. A context set C is controllable when satisfying: (i) *distribution separation*, when for no $c \in C_{\text{train}}$ and $c' \in C_{\text{test}}$ do we have $c \simeq c'$; (ii) *observation independence*, the agent should not assume that the projection function ϕ is injective (i.e., it does not assume equal observations in different contexts to be from the same state); and (iii) *compositional structure*, where for all $c \in C_{\text{train}}$ and for all $c' \in C_{\text{test}}$ such that $c \neq c'$, $0 < d(c, c') \leq \delta_C$, where $\delta_C = \max\{d(c, c') \mid c \in C_{\text{train}}, c' \in C_{\text{test}}\}$.

In a block-stacking task, rather than random variation, controllable contexts allow control over problem characteristics, such as block count and colour. For navigation, map layout, room connectivity, and obstacle placement become explicit parameters. This enables systematic testing, varying single factors or specific combinations to reveal which compositional capabilities agents possess. For symbolic contexts representable as vectors of properties, we advocate the *Levenshtein* distance [16], which counts the number of insertion and deletion operations needed to transform between configurations. For example, transforming the specification of blocks from *red, blue, green* to *red, yellow, green* requires one deletion (*blue*) and one insertion (*yellow*), more accurately reflecting compositional changes, which Euclidean metrics struggle with.

Finally, deterministic transitions eliminate confounding stochasticity. When agents fail in stochastic environments, we cannot distinguish between compositional limitations and unlucky events. Consider navigation under unsafe conditions: failure might reflect inability to adapt (generalisation failure) or unstable random transitions that would fail regardless (environmental stochasticity). Deterministic transitions eliminate this confounding factor. The execution of an action in a state produces a unique outcome, ensuring that failures reflect the agent’s limitations rather than chance. This enables reproducible debugging, allowing researchers to trace exact failure points, identify violated compositional rules, and design targeted improvements. Determinism transforms generalisation from a statistical property to a measurable one.

3.2 Comparison with POMDPs

A comparison between GCMDPs and Partially Observable MDPs (POMDPs) is warranted due to structural similarities. POMDPs excel when agents must maintain beliefs about hidden states and perform Bayesian inference. Yet, IL from expert demonstrations presents a different challenge: agents receive complete demonstrations with full state and context information. The problem is not inferring hidden states but understanding compositional structure. The projection $\phi : S \times C \rightarrow O$ captures how context shapes observation without hiding information. This transparency enables systematic generalisation testing that is impossible under POMDPs’ probabilistic framework. When a GCMDP agent fails, we know it lacks compositional understanding, not state information, a far more tractable debugging problem.

3.3 Measuring Compositional Generalisation

Current IL metrics give us success rate but reveal nothing about the nature of that success. An agent with 90% accuracy may be simply replicating trajectories in familiar contexts but failing catastrophically on novel ones, or it may indeed possess robust compositional understanding, albeit with occasional execution errors.

Consider a block-stacking task where an agent learns to execute stack commands that build towers by placing blocks in sequence (e.g., `stack(red, blue)` places the red block on top of the blue one). Suppose training includes blocks “red”, “blue”, and “green” in two-block configurations. At test time, we might present novel challenges: a new block (purple), longer sequences (red, blue, green, red) or symmetrical configurations (`stack(blue, red)`). Two agents with identical test accuracy might differ dramatically—one handles novel colours but fails on longer sequences, while the other shows the opposite pattern. Traditional metrics cannot capture these distinctions. Thus, we propose the *generalisation boundary*, adapted from behavioural science [4], that cuts through this ambiguity by explicitly measuring how far agents can deviate from training contexts while maintaining performance.

Definition 3.4. Let C be a set of controllable contexts, p a minimum performance threshold, and $SR_{\pi_\theta}(d_c)$ the success rate at contextual distance d_c (averaged over all contexts at that distance). We define the generalisation boundary as: $\delta_{\pi_\theta}(p) = \sup\{d_c \in [0, \delta_C] \mid SR_{\pi_\theta}(d_c) \geq p\}$, where δ_C is the maximum contextual distance.

This reveals generalisation profiles invisible to traditional evaluation. Two agents with identical test accuracy might have vastly different boundaries: one maintaining perfect performance until sudden failure, while the other degrades gracefully as contextual distance increases. High thresholds ($p = 0.9$) test robust systematicity, while lower thresholds ($p = 0.5$) reveal absolute limits of compositional understanding. Combined with compositional generalisation, we can measure contextual distances to different test cases. In our block-stacking example, we could change a block’s colour to test substitutivity, add a block to test productivity, or reorder the blocks to test systematicity. In addition, we can combine these forms of compositional generalisation to further test the agent. Agents may succeed at different distances for different compositional dimensions. An agent might handle substitutivity but fail systematicity, revealing specific compositional limitations.

3.4 Suitable Environments and Benchmarks

Not all environments enable compositional generalisation measurement. In Procgen [5], when agents fail on unseen levels, we cannot determine whether they lack compositional understanding or encountered entirely novel mechanics. Compatible benchmarks require: (i) fixed primitive semantics: a “grasp” action should function identically for red or blue blocks; environments changing primitives meanings conflate adaptation with relearning; (ii) isolated compositional dimensions: systematicity varies only structural combinations, productivity extends sequence length without novel primitives, substitutivity changes components with functional equivalents; (iii) interpretable failures: distinguishing nature of failures, perceptual (inability to recognise blocks), compositional (inability to plan sequences), or operational (inability to perform primitives); and (iv) modular evaluation: primitives, compositional rules, and integrated behaviours separately. Therefore, CRL calls for careful and well-thought-out benchmarks. As Mitchell [17] argues, current AI evaluation often cannot distinguish memorisation from genuine understanding; compositional generalisation measurement inherits this challenge but, through careful benchmark design, offers an opportunity for deeper insight into agent capabilities.

4 FUTURE RESEARCH LANDSCAPE

CRL opens research paths that extend beyond incremental improvements to IL. These directions challenge fundamental assumptions about learning, adaptation, and intelligence while remaining grounded in achievable scientific progress.

Compositional Generalization at Scale: Current benchmarks fail to measure compositional capabilities, conflating multiple dimensions or lacking structure entirely. CRL demands specialised evaluation to reveal exactly which capabilities fail: systematicity, varying only in structural combinations; productivity, extending sequences without novel primitives; substitutivity, replacing objects with functional equivalents. Open-world deployment presents the ultimate challenge: agents must decompose novel situations, repurpose primitives, and recognise compositional limits. Research into compositional curricula that mimic human developmental progression from simple to complex behaviours could dramatically improve generalisation while revealing fundamental learning principles.

Hybrid Architectures: CRL enables natural integration with complementary paradigms. Foundation models provide semantic priors for primitive selection; planners excel at long-horizon reasoning. Combining these yields agents that learn primitives via imitation but compose them using principled algorithms, addressing IL’s inability to reason about novel combinations. The challenge: preserving each paradigm’s strengths while ensuring semantic alignment between learned primitives and symbolic representations.

Safety and Ethics in Adaptive Systems: Creative recombination risks discovering harmful behaviours never demonstrated. Compositional safety specifications must constrain entire combination spaces, not individual trajectories. Behavioural contracts could guarantee safe primitive combinations—proving no navigation-grasping combination causes collision requires compositional algebra, not trajectory enumeration. Ethical questions arise: should agent populations preserve behavioural diversity or converge? How do we ensure cross-cultural adaptation respects local norms?

Multi-agent Compositional IL: Teams could share and recombine behavioural primitives, enabling collective repertoire expansion through social learning. This mirrors cultural transmission mechanisms in human societies and opens questions about convergence versus diversity in behavioural populations. Social learning theory shows how discoveries propagate through populations, suggesting mechanisms for efficient primitive sharing.

Cognitive Science Synergies: Developmental psychology reveals a structured progression from mimicry to composition. Kolb’s experiential learning theory [15] suggests beneficial cycles: experiencing demonstrations, analysing patterns, extracting structures, and testing combinations, implying that agents should reflect rather than pure matching. Behavioural science research on exploration-exploitation trade-offs can inform how agents balance exploring new combinations of primitives versus exploiting known compositions. Social learning theory shows how discoveries propagate through populations, mirroring human cultural transmission.

The compositional structure from GCMDPs opens pathways we have not described here, such as: multi-agent solutions using IL, where teams can share and recombine behavioural primitives, and implications for explainable AI through interpretable primitive decomposition. By reconceptualising IL through the lens of compositionality, CRL transforms IL from a narrow technique for trajectory reproduction into a foundational paradigm for understanding how agent systems acquire, adapt, and transmit complex behaviours. The breadth of these possibilities underscores that CRL is not merely an incremental advance but a research agenda that will reshape how we approach learning from demonstration.

5 CONCLUSION

This paper calls for a research agenda re-orientation in IL: from narrow mimicry toward lifelong adaptability through Compositional Repertoire Learning (CRL). By formalising Goal-conditioned Contextual MDPs and introducing compositional generalisation dimensions, we outline a foundation for agents that learn once and adapt everywhere. CRL envisions architectures that allow for integrating imitation with planning, symbolic reasoning, and foundation models, supported by benchmarks that stress compositional generalisation (systematicity, productivity, and substitutivity).

The significance of this shift lies in enabling agents to operate beyond static demonstrations, adapt to unseen contexts, and recombine learned behaviours for novel goals. These capabilities are critical for advancing autonomy in robotics, human-AI collaboration, and open-world systems where variability and uncertainty are the norm. By embracing CRL, the agent community can establish IL as a cornerstone for next-generation AI, bridging data-driven learning with structured reasoning. This will enable the realisation of more sophisticated cognitive learning techniques, paving the way for more robust, adaptable, and proactive agents.

ACKNOWLEDGMENTS

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org).

REFERENCES

- [1] Leonardo Amado, Reuth Mirsky, and Felipe Meneguzzi. 2022. Goal recognition as reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9644–9651.
- [2] Leonardo R. Amado, Ramon F. Pereira, and Felipe Meneguzzi. 2023. Robust Neuro-Symbolic Goal and Plan Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11937–11944.
- [3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. *Advances in neural information processing systems* 30 (2017).
- [4] Albert Bandura. 1977. *Social learning theory*. Prentice-Hall, Englewood Cliffs ;.
- [5] Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. 2019. Leveraging Procedural Generation to Benchmark Reinforcement Learning. *arXiv preprint arXiv:1912.01588* (2019).
- [6] Meghan A. Deshais, Jacquelyn MacDonald, and Leslie Quiroz. 2022. *Generalized Imitation and Applied Behavior Analysis*. Springer International Publishing, 821–844. https://doi.org/10.1007/978-3-030-88538-0_35
- [7] Osher Elhadad, Felipe Meneguzzi, , and Reuth Mirsky. 2026. GRAIL: Goal Recognition Alignment through Imitation Learning. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, Chris Amato and Louise Dennis (Eds.). (to appear).
- [8] Nathan Gavenski, Felipe Meneguzzi, Michael Luck, and Odinaldo Rodrigues. 2024. A Survey of Imitation Learning Methods, Environments and Metrics. *arXiv preprint arXiv:2404.19456* (2024).
- [9] Nathan Gavenski, Juarez Monteiro, Adilson Medronha, and Rodrigo C. Barros. 2022. How Resilient Are Imitation Learning Methods to Sub-optimal Experts?. In *Intelligent Systems*, João Carlos Xavier-Junior and Ricardo Araújo Rios (Eds.). Springer International Publishing, Cham, 449–463.
- [10] Nathan Gavenski and Odinaldo Rodrigues. 2025. Quantifying Generalisation in Imitation Learning. In *Advances in Neural Information Processing Systems*, Vol. 38. Curran Associates, Inc.
- [11] Assaf Hallak, Dotan Di Castro, and Shie Mannor. 2015. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259* (2015).
- [12] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research* 67 (April 2020), 757–795. <https://doi.org/10.1613/jair.1.11674>
- [13] Leslie Pack Kaelbling. 1993. Learning to achieve goals. In *IJCAI*, Vol. 2. 1094–8.
- [14] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2023. A Survey of Zero-shot Generalisation in Deep Reinforcement Learning. *Journal of Artificial Intelligence Research* 76 (Jan. 2023), 201–264. <https://doi.org/10.1613/jair.1.14174>
- [15] David A Kolb. 1983. *Experiential learning*. Financial Times Prentice Hall, Upper Saddle River, NJ.
- [16] VI Lcvenshtcin. 1966. Binary coors capable or ‘correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, Vol. 10.
- [17] Melanie Mitchell. 2023. How do we know how smart AI systems are? *Science* 381, 6654 (2023), eadj5957. <https://doi.org/10.1126/science.adj5957> arXiv:<https://www.science.org/doi/pdf/10.1126/science.adj5957>
- [18] Scott Niekum, Sarah Osentoski, George Konidaris, Sachin Chitta, Bhaskara Marthi, and Andrew G Barto. 2015. Learning grounded finite-state representations from unstructured demonstrations. *The International Journal of Robotics Research* 34, 2 (2015), 131–157.
- [19] Dean A Pomerleau. 1988. Alvin: An autonomous land vehicle in a neural network. *Advances in neural information processing systems* 1 (1988).
- [20] Stuart J. Russell and Peter Norvig. 2022. *Artificial intelligence : a modern approach* (global edition ed.). Pearson Education Limited, Harlow, England.
- [21] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press, Cambridge, MA. <https://search.worldcat.org/title/1043175824>
- [22] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1-2 (1999), 181–211.
- [23] Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Steven Wu. 2021. Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 10022–10032. <https://proceedings.mlr.press/v139/swamy21a.html>
- [24] Faraz Torabi, Garrett Warnell, and Peter Stone. 2019. Recent Advances in Imitation Learning from Observation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 6325–6331. <https://doi.org/10.24963/IJCAI.2019/882>