# It's just semantics: How to get robots to understand the world the way we do

Jen Jen Chung[1,2], Julian Förster[2], Paula Wulkop[2], Lionel Ott[2], Nicholas Lawrance[2,3], and Roland Siegwart[2]

[1]The University of Queensland [2]ETH Zürich [3]CSIRO
jenjen.chung@uq.edu.au, {fjulian, pwulkop, liootte}@ethz.ch,
nicholas.lawrance@csiro.au, rsiegwart@ethz.ch

**Abstract.** Increasing robotic perception and action capabilities promise to bring us closer to agents that are effective for automating complex operations in human-centered environments. However, to achieve the degree of flexibility and ease of use needed to apply such agents to new and diverse tasks, representations are required for generalizable reasoning about conditions and effects of interactions, and as common ground for communicating with non-expert human users. To advance the discussion on how to meet these challenges, we characterize open problems and point out promising research directions.

## 1 Introduction

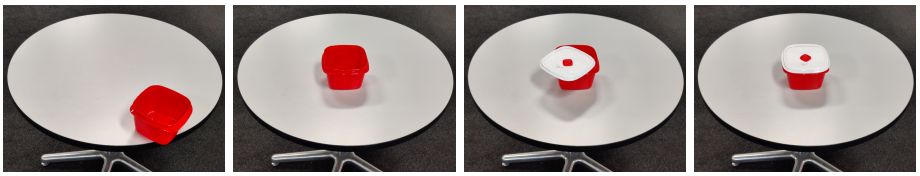
There is growing excitement surrounding the possibility of service robots soon becoming a staple in hospitals, malls, construction sites or even in the home. This is perhaps not entirely unwarranted given the astonishing technological progress that we've observed over just the last few years across all areas of robotics. There are now myriad options for autonomous point-to-point navigation in changing and dynamic human spaces [1–3], similarly so for safe physical interaction between robots and humans or other objects in the environment [4, 5]. On top of this, there are also any number of methods that generate the internal scene representations needed to support these tasks [6–10]. Nevertheless, even the most optimistic among us would likely agree that a scenario where we could simply ask a robot (or a team of robots) to "tidy the room" and expect a satisfactory outcome is still a fairly distant dream. The individual skills needed to complete such a task—pick and place of objects [11–15], opening and closing drawers and cupboards [16–18], navigation in clutter [19, 20], etc.—may all be available, but there is still a missing link that's needed to translate the semantic notion of "tidy" into a potentially extensive and complex sequence of robot-legible actions.

The field of high-level symbolic planning [21] pushes towards this idea by offering frameworks that operate over *predicates* (binary symbols of relations or properties, e.g. "teddy on floor") [22–24]. By describing the state of the world as a set of predicates and defining the actions available to the robot in terms of preconditions and effects (e.g. the precondition for a grasp action may be

the predicate "object in reach" and its resulting effect is "object in gripper"), such planners can accept commands such as "teddy in cupboard", to output an action sequence {navigate to teddy, grasp teddy, navigate to cupboard, place teddy} that completes the task. While the logic may be straightforward, even such a simple example hides a surprising amount of complexity. Most notably, the semantic concepts of "*on* the floor" or "*in* the cupboard" lead to the same interpretability problem as in the "tidy" example. It may be easier to conceive of a metric computation (a grounding) to evaluate the "on" and "in" predicates compared to "tidy", however, such a check still needs to be implemented. Indeed, one is needed for every predicate the robot encounters. Currently available solutions either hand-code these rules, which is untenable in the long run, or learn the groundings [25–28], which overfits to a specific task. Alternatively, end-to-end approaches short-circuit symbolic planning altogether by directly learning actions from semantic commands without a (human-legible) intermediate representation [29], which precludes a shared human-robot understanding. Moreover, given the incomplete, fuzzy, representation- and context-dependent nature of most predicates (see Figure 1 for an example of how even "on" can get messy), there is often no clear or easy way to consistently translate these to the metric representation of the environment in which robots operate.

So far we've only considered the robot-agnostic elements of predicate understanding: how to determine whether a predicate holds given just the state of the world. A second challenge arises from the description of the robot actions, whose preconditions and effects by necessity include robot-specific predicates. Take the grasp action as an example, previously we suggested that having the target "in reach" would be a suitable precondition. However, whether or not this predicate holds depends entirely on the robot's available workspace, which varies from setup to setup. "In reach" also doesn't cover the full set of conditions needed to determine an object's graspability, since we would also need to factor in elements such as the robot's grasping mechanism (fingers, jamming, suction, etc.), its sensing capabilities and its payload capacity. A complete and correct set of symbolic action definitions would need to be based on robot-specific object affordances, and these, too, must somehow be obtained.

Given the difficulty of generating analytical translations between semantic predicates and the general metric space of a robot, it seems that the most promising solutions will likely require some degree of learning. Self-supervised strategies [30] may help to alleviate the need for manually labeled data and these can be of particular value when learning robot-centric affordances. Notably, the fidelity of available physics simulators [31] now lets us learn robot skills which transfer remarkably well to real-world execution [32, 33]. Of course, we're still left with all the robot-agnostic elements, which ultimately require some degree of user input, either via labeling [34, 35] or by demonstration [36]. As with any works needing user input, the core challenges are how to most efficiently gather this limited and precious resource, and how to most effectively make use of it. Works that deal with activity classification [37] and behavioral cloning from observation [38] can supplement these efforts by exploiting the huge amount of on-



(a) (b) (c) (d)

Fig. 1: (a)-(b) Two examples of where the predicate "container on table" holds. Indeed, we can generally agree that this predicate holds for any position or orientation of the container so long as it is supported from below by the table surface. (c) Does the predicate "lid on container" hold in this case? Typically when we say "put the lid on the container" we would like the outcome shown in (d). But if we were simply identifying the lid, we may describe it in both (c) and (d) as "the lid on the container".

line image and video data, while active learning [39] can help reduce the required amount of labeled data. However, regardless of the data source, additional effort will invariably be needed to verify the correctness of learned semantic predicates and methods for further knowledge refinement will almost surely be required.

This paper presents some of the key challenges in consolidating the semantic nature in which humans interpret the world and the actions that we take within it, with the fundamentally metric representations available to our robots. The motivating thesis of this investigation is that by enabling a mutual understanding of the world (or at least a way to map from a human's internal representation of it to a robot's and vice versa), we can interact with robots as smoothly as we would with other humans. Potentially arriving at a point where we can get them to tidy our rooms with a single command.

## 2 Contextual Scene Understanding

Semantics provides a compressed representation of the world at the level of objects and their relationships. They allow us to reason effortlessly over tasks that would otherwise be extremely tedious to describe within a metric representation. Assuming that we are stuck with robots that must, at the end of the day, operate using metric representations, the question is then, how do we convert our semantic, relational and contextual understanding of the environment into such a metric space? Deep learning-based algorithms for computer vision [40–42] and natural language processing [43, 44] have already given us a taste of what's possible at the conjunction of high-powered compute and enormous amounts of data. While this presents an obvious starting point, we posit here that a robot—an artificial *embodied* agent—performing long term reasoning and planning in shared human environments will need to undergo continual online adaptation of its world representation. This is not only so that it updates its scene understanding according to newly discovered semantic predicates, but also so that it can render its interactions with the world (reasoning about its own capabilities and the effects of its actions) within the same representation as new objects or new capabilities are introduced. From a symbolic planning perspective, this means that we need efficient strategies for discovering and evaluating predicates, as well as methods for generating agent-specific action definitions which can be corrected and completed over time given new interaction experiences.

### 2.1 Spatial Predicates: Object-to-object Relationships

Thankfully, a large number of predicates that are needed to describe typical tasks undertaken in human spaces can be evaluated spatially or temporally. These involve prepositions that describe the relationship between two objects (e.g. "on") or two actions (e.g. "after"). The latter is generally well-handled by symbolic planners that apply temporal logic [45, 46], while we can conceivably learn the rules for spatial predicates given sufficient training examples of relevant sensory data, such as comparing point clouds or bounding boxes of objects [47].

Unfortunately, and perhaps not surprisingly, this is not the end of the story for spatial predicates. Take for instance the examples shown in Figure 1. While the container is unquestionably "on" the table in all instances, whether or not the lid is considered "on" the container in Figure 1c and Figure 1d may differ depending on the intention of the task. Although it's possible to exhaustively learn the different subtleties of each predicate from scratch, this is not a satisfying solution and may be fatally unrealistic depending on the training data and memory requirements. A more principled approach may be to build levels of specificity into the representation which would allow directly extending and specializing from established predicates. That way, when a robot is told that it failed to "put the lid on the container" as in Figure 1c, it doesn't search across all possible configurations of the lid and container, but rather searches only within those configurations for which its original understanding of "on" holds.

Automatically learned predicate extensions can also provide an avenue for efficient knowledge consolidation. For instance, a sequence of $n$ objects one "on" the other would need $n-1$ "on" predicates to define. Alternatively, a person would typically describe such an arrangement with a single predicate "stack". Learning to associate jointly observed sets of predicates can potentially fast-track planning and will in general enable more compact state representations. One consideration to note is that some conditions may not be explicitly observable when simply associating predicate sets. For example, the order of the objects in a stack may or may not strictly matter. While this has little impact on being able to determine if a "stack" predicate holds, it can become very important for a planner that needs to know how to generate a stack from a random set of objects. Put another way, "on" is only a necessary condition for "stack", additional effort may still be needed to learn the full set of sufficient conditions for these higher-level predicates.

### 2.2 Affordances: Object-to-agent and Scene-to-agent Relationships

A second set of predicates that are needed are those that relate specifically to the agent and the actions that it can execute. High-level symbolic actions are defined by preconditions and effects, the former is the set of predicates that must hold in order for the action to be executable while the latter is the set of predicate changes that the action causes. Generally speaking, action preconditions relate to the agent-specific affordances of the particular object(s) targeted by that action, for example, recall the discussion on graspability from Section 1. One option for

learning object affordances is to simply rely on manually labeled data such as [35], however the limitations of this may quickly become apparent in an open world setting. In contrast, methods that can efficiently update and add to an existing knowledge base from the agent's own experiences are considerably more powerful and more appealing, especially given our use case in robotics.

Some affordance predicates may be immediately observable even without requiring physical interaction with the object. For example, the geometry of an object can already disqualify it from being considered graspable. Beyond geometry, other remotely observable object properties can provide good proxies that allow affordance generalization over unseen objects. Over time, a robot could learn to associate features, such as surface texture, with how easy an object has been to grasp given a particular end-effector. In an ideal scenario, such associations would remove the need for physical interaction and would allow the agent to directly attach an affordance to a new object just based on remote observation. The challenge is then to identify the salient object features that reveal these correlations and also to decide how optimistically (or conservatively) to generalize affordances to newly encountered objects.

Ultimately, however, the only way we can truly verify if an object has a particular affordance is to physically interact with it to see if the associated action succeeds. To further complicate things, there are agent-, object- and scene-level reasons why an action may fail. An object may not be graspable because (i) it's too heavy for the robot to lift; (ii) its current orientation does not allow grasping; or (iii) other objects in the scene are blocking the robot from grasping it. From the perspective of a symbolic planner, it can be quite useful to distinguish between these different cases as they each point to potential recovery mechanisms. In the first case, the ability to recover lies solely with the agent (increase the robot payload); in the second case the robot may be able to move the object into a graspable orientation; while in the last case, the robot will need to interact with other objects in the scene to change the graspability predicate of the target object. Uncovering the true reason behind why a particular object affordance doesn't hold will likely require a combination of knowledge extrapolation (to hypothesize what went wrong) and sophisticated exploration strategies to avoid the computational blowup of an exhaustive search. Predicate extension methods, such as those described in Section 2.1, will then also be needed to update the state and action representations to enable successful replanning.

### 2.3 Challenges and Benefits of Embodiment

Since affordance predicates are inherently tied to the robot's capabilities, their evaluation is at least directly measurable by the robot (did the action succeed or not?). Of course, knowing that an affordance holds is not the same as knowing how to effect that affordance. Recent works propose methods that exploit simulation to learn this directly from point cloud data [16, 17]. However, these approaches stop short of a full evaluation and only explore object interactions with a disembodied gripper, relying instead on downstream motion planners to

deal with kinematic and dynamics constraints that arise from the full realization of the robot. It remains to be seen whether learning a more generalized (robot-agnostic) set of object affordances, which are then refined to each specific setup, or directly learning a robot-centric representation results in better overall functionality. As is typical for learning algorithms, ease of training, ease of use, generalizability and overall performance will all factor into this comparison, and perhaps unique to robot-affordance learning, we would also be very interested in how readily the learned representation can be adapted given live, real-world interaction data.

## 3 Summary and Outlook

Contextual scene understanding is challenging for a robot but it's a problem worth tackling as it enables robots to truly *reason* about the world to solve general, large-scale, long-horizon tasks given simple semantic commands. We've identified several major challenges related to predicate-based world representations and outlined promising directions for investigation. These include ideas for automatic predicate extension via specialization of existing predicates or consolidation of jointly observed predicates, as well as ideas for generalizing affordances to unseen objects and exploiting agent embodiment in robotic domains.

## References

1. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on Robotics **32**(6) (2016)
2. Van den Berg, J., Lin, M., Manocha, D.: Reciprocal velocity obstacles for real-time multi-agent navigation. In: IEEE International Conference on Robotics and Automation. (2008)
3. Gao, Y., Huang, C.M.: Evaluation of socially-aware robot navigation. Frontiers in Robotics and AI **8**(721317) (2021)
4. Alami, R., et al.: Safe and dependable physical human-robot interaction in anthropic domains: State of the art and challenges. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. (2006)
5. Zacharaki, A., Kostavelis, I., Gasteratos, A., Dokas, I.: Safety bounds in human robot interaction: A survey. Safety Science **127** (2020)
6. Florence, P., Manuelli, L., Tedrake, R.: Self-supervised correspondence in visuomotor policy learning. IEEE Robotics and Automation Letters **5**(2) (2020)
7. Garg, S., Sünderhauf, N., Dayoub, F., Morrison, D., Cosgun, A., Carneiro, G., Wu, Q., Chin, T.J., Reid, I., Gould, S., Corke, P., Milford, M.: Semantics for robotic mapping, perception and interaction: A survey. Foundations and Trends® in Robotics **8**(1–2) (2020)
8. Narita, G., Seno, T., Ishikawa, T., Kaji, Y.: PanopticFusion: Online volumetric semantic mapping at the level of stuff and things. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. (2019)
9. Grinvald, M., Furrer, F., Novkovic, T., Chung, J.J., Cadena, C., Siegwart, R., Nieto, J.: Volumetric instance-aware semantic mapping and 3D object discovery. IEEE Robotics and Automation Letters **4**(3) (2019)

10. Kothari, P., Kreiss, S., Alahi, A.: Human trajectory forecasting in crowds: A deep learning perspective. IEEE Transactions on Intelligent Transportation Systems (2021)
11. Bohg, J., Morales, A., Asfour, T., Kragic, D.: Data-driven grasp synthesis — a survey. IEEE Transactions on Robotics **30**(2) (2014)
12. Gualtieri, M., ten Pas, A., Saenko, K., Platt, R.: High precision grasp pose detection in dense clutter. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. (2016)
13. Mahler, J., Matl, M., Satish, V., Danielczuk, M., DeRose, B., McKinley, S., Goldberg, K.: Learning ambidextrous robot grasping policies. Science Robotics **4**(26) (2019)
14. Morrison, D., Corke, P., Leitner, J.: Learning robust, real-time, reactive robotic grasping. The International Journal of Robotics Research **39**(2-3) (2020)
15. Breyer, M., Chung, J.J., Ott, L., Siegwart, R., Nieto, J.: Volumetric grasping network: Real-time 6 DOF grasp detection in clutter. In: Conference on Robot Learning. (2021)
16. Mo, K., Guibas, L.J., Mukadam, M., Gupta, A., Tulsiani, S.: Where2Act: From pixels to actions for articulated 3D objects. In: IEEE/CVF International Conference on Computer Vision. (2021)
17. Wu, R., Zhao, Y., Mo, K., Guo, Z., Wang, Y., Wu, T., Fan, Q., Chen, X., Guibas, L., Dong, H.: VAT-Mart: Learning visual action trajectory proposals for manipulating 3D articulated objects. In: International Conference on Learning Representations. (2022)
18. Xu, Z., Zhanpeng, H., Song, S.: UMPNet: Universal manipulation policy network for articulated objects. IEEE Robotics and Automation Letters **7**(2) (2022)
19. Pierson, A., Vasile, C.I., Gandhi, A., Schwarting, W., Karaman, S., Rus, D.: Dynamic risk density for autonomous navigation in cluttered environments without object detection. In: International Conference on Robotics and Automation. (2019)
20. Regier, P.: Robot Navigation in Cluttered Environments. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn (2022)
21. Karpas, E., Magazzeni, D.: Automated planning for robotics. Annual Review of Control, Robotics, and Autonomous Systems **3** (2019)
22. Fikes, R.E., Nilsson, N.J.: Strips: A new approach to the application of theorem proving to problem solving. Artificial Intelligence **2**(3-4) (1971)
23. McDermott, D., Ghallab, M., Howe, A., Knoblock, C., Ram, A., Veloso, M., Weld, D., Wilkins, D.: PDDL: The planning domain definition language. Technical report, Yale Center for Computational Vision and Control (1998)
24. Garrett, C.R., Lozano-Pérez, T., Kaelbling, L.P.: FFRob: Leveraging symbolic planning for efficient task and motion planning. The International Journal of Robotics Research **37**(1) (2018)
25. Konidaris, G., Kaelbling, L.P., Lozano-Perez, T.: From skills to symbols: Learning symbolic representations for abstract high-level planning. Journal of Artificial Intelligence Research **61** (2018)
26. Ames, B., Thackston, A., Konidaris, G.: Learning symbolic representations for planning with parameterized skills. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. (2018)
27. Silver, T., Chitnis, R., Tenenbaum, J., Kaelbling, L.P., Lozano-Peréz, T.: Learning symbolic operators for task and motion planning. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. (2021)
28. Yuan, W., Paxton, C., Desingh, K., Fox, D.: SORNet: Spatial object-centric representations for sequential manipulation. In: Conference on Robot Learning. (2022)

29. Shridhar, M., Manuelli, L., Fox, D.: CLIPort: What and where pathways for robotic manipulation. In: Conference on Robot Learning. (2022)
30. Nair, A., Bahl, S., Khazatsky, A., Pong, V., Berseth, G., Levine, S.: Contextual imagined goals for self-supervised robotic learning. In: Conference on Robot Learning. (2020)
31. Collins, J., Chand, S., Vanderkop, A., Howard, D.: A review of physics simulators for robotic applications. IEEE Access **9** (2021)
32. Peng, X.B., Andrychowicz, M., Zaremba, W., Abbeel, P.: Sim-to-real transfer of robotic control with dynamics randomization. In: IEEE International Conference on Robotics and Automation. (2018) 3803–3810
33. Zhao, W., Queralta, J.P., Westerlund, T.: Sim-to-real transfer in deep reinforcement learning for robotics: A survey. In: IEEE Symposium Series on Computational Intelligence. (2020)
34. Cohen, V., Burchfiel, B., Nguyen, T., Gopalan, N., Tellex, S., Konidaris, G.: Grounding language attributes to objects using Bayesian eigenobjects. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. (2019)
35. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3D semantic scene graphs from 3D indoor reconstructions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020)
36. Gopalan, N., Rosen, E., Konidaris, G., Tellex, S.: Simultaneously learning transferable symbols and language groundings from perceptual data for instruction following. Robotics: Science and Systems XVI (2020)
37. Rodríguez-Moreno, I., Martínez-Otzeta, J.M., Sierra, B., Rodriguez, I., Jauregi, E.: Video activity recognition: State-of-the-art. Sensors **19**(14) (2019)
38. Torabi, F., Warnell, G., Stone, P.: Behavioral cloning from observation. In: International Joint Conference on Artificial Intelligence. (2018) 4950–4957
39. Bıyık, E., Losey, D.P., Palan, M., Landolfi, N.C., Shevchuk, G., Sadigh, D.: Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. The International Journal of Robotics Research **41**(1) (2022)
40. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. (2012)
41. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. (2016)
42. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision. (2017)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. (2017)
44. Brown, T., Mann, B., Ryder, N., Subbiah, M., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems. (2020)
45. Belta, C., Bicchi, A., Egerstedt, M., Frazzoli, E., Klavins, E., Pappas, G.J.: Symbolic planning and control of robot motion [Grand Challenges of Robotics]. IEEE Robotics & Automation Magazine **14**(1) (2007)
46. Kress-Gazit, H., Fainekos, G.E., Pappas, G.J.: Temporal-logic-based reactive mission and motion planning. IEEE Transactions on Robotics **25**(6) (2009)
47. Mo, K., Qin, Y., Xiang, F., Su, H., Guibas, L.: O2O-Afford: Annotation-free large-scale object-object affordance learning. In: Conference on Robot Learning. (2022)