A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms

Oliver Kroemer
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Scott Niekum
Department of Computer Science
The University of Texas at Austin
Austin, TX 78712, USA

George Konidaris
Department of Computer Science
Brown University
Providence, RI 02912, USA

Editor: Daniel Lee

Abstract
A key challenge in intelligent robotics is creating robots that are capable of directly interacting with the world around them to achieve their goals. The last decade has seen substantial growth in research on the problem of robot manipulation, which aims to exploit the increasing availability of affordable robot arms and grippers to create robots capable of directly interacting with the world to achieve their goals. Learning will be central to such autonomous systems, as the real world contains too much variation for a robot to expect to have an accurate model of its environment, the objects in it, or the skills required to manipulate them, in advance. We aim to survey a representative subset of that research which uses machine learning for manipulation. We describe a formalization of the robot manipulation learning problem that synthesizes existing research into a single coherent framework and highlight the many remaining research opportunities and challenges.

Keywords: manipulation, learning, review, robots, MDPs

1. Introduction
Robot manipulation is central to achieving the promise of robotics—the very definition of a robot requires that it has actuators, which it can use to effect change on the world. The potential for autonomous manipulation applications is immense: robots capable of manipulating their environment could be deployed in hospitals, elder- and child-care, factories, outer space, restaurants, service industries, and the home. This wide variety of deployment scenarios, and the pervasive and unsystematic environmental variations in even quite specialized scenarios like food preparation, suggest that an effective manipulation robot must...
be capable of dealing with environments that neither it nor its designers have foreseen or encountered before.

Researchers have therefore focused on the question of how a robot should learn to manipulate the world around it. That research has ranged from learning individual manipulation skills from human demonstration, to learning abstract descriptions of a manipulation task suitable for high-level planning, to discovering an object’s functionality by interacting with it, and many objectives in between. Some examples from our own work are shown in Figure 2.

Our goal in this paper is twofold. First, we describe a formalization of the robot manipulation learning problem that synthesizes existing research into a single coherent framework. Second, we aim to describe a representative subset of the research that has so far been carried out on robot learning for manipulation. In so doing, we highlight the diversity of the manipulation learning problems that these methods have been applied to as well as identify the many research opportunities and challenges that remain.

Our review is structured as follows. First, we survey the key concepts that run through manipulation learning, which provide its essential structure (Sec. 2). Section 3 provides a broad formalization of manipulation learning tasks that encompasses most manipulation problems but which contains the structure essential to the problem.

The remainder of the review covers several broad technical challenges. Section 4 considers the question of learning to define the state space, where the robot must discover the relevant state features and degrees of freedom attached to each object in its environment. Section 5 describes approaches to learning an environmental transition model that describes how a robot’s actions affect the task state. Section 6 focuses on how a robot can learn a motor control policy that directly achieves some goal, typically via reinforcement learning.
(Sutton and Barto, 1998), either as a complete solution to a task or as a component of that solution. Section 7 describes approaches that characterize a motor skill, by learning a description of the circumstances under which it can be successfully executed, and a model of the resulting state change. Finally, Section 8 surveys approaches to learning procedural and state abstractions that enable effective high-level learning, planning, and transfer.

2. Common Concepts in Learning for Manipulation

Manipulation tasks have significant internal structure and exploiting this structure may prove key to efficient and effective manipulation learning. Hence, before formalizing the manipulation learning problem, we will first discuss this internal structure.

2.1 Manipulations as Physical Systems

Every manipulation involves a physical robot interacting with its environment. As a result, all manipulations are subject to the laws of physics and the structure that they impose. This fairly obvious statement has wide-reaching implications for manipulation learning algorithms. Basic physical concepts (e.g., distinct objects cannot occupy the same space, and gravity applies a mass-dependent force to objects) provide strong prior knowledge for manipulation tasks. Concepts from physics, such as irreversible processes and object masses, are so fundamental that we generally take them for granted. However, these concepts provide invaluable prior knowledge and structure that can be exploited by learning algorithms.
and thus make learning manipulation skills tractable. Most of the concepts discussed in
the remainder of this section are to some extent the result of manipulations being physical
processes.

2.2 Underactuation, Nonholonomic Constraints, and Modes in Manipulations

Manipulation tasks are almost always characterized as underactuated systems. Even if
the robot is fully actuated, inanimate objects in the environment will contribute a set of inde-
pendent state variables to the state space, but not increase the robot’s action space. The
resulting discrepancy between the number of actuators and the number of DoFs means that
the system is underactuated. To change the state of the objects, the robot must first move
into a state from which it can alter the object’s state and then apply the desired manipula-
tion, e.g., make contact with an object and then push the object. These requirements can
be represented as a set of nonholonomic constraints that define how the robot can move
through the full state space based on different interactions with the environment.

Manipulation tasks can be modelled as hybrid systems, wherein the system dynamics
are continuous within each of a number of discrete dynamical modes. The dynamics are thus
piecewise continuous. In manipulation tasks, the mode switches often correspond to making
or breaking of contacts, with different contacts applying corresponding constraints and
allowing the robot to interact with various objects. The conditions for transition between
modes often correspond to subgoals or obstacles depending on which state variables the
robot should change or keep constant. Unfortunately, modes also make manipulation tasks
inherently discontinuous. Hence, small changes in the state can have a significant effect on
the outcome of a manipulation. It is therefore important that robots monitor their skill
executions for unexpected and undesired mode transitions.

2.3 Interactive Perception and Verification

Robots can perceive certain latent object properties by observing the outcomes of different
manipulation actions. This process is known as interactive perception. Many properties of
objects, such as material or kinematic constraints, can only be determined reliably through
interactive perception. If the goal of the task is to alter a latent property, then the robot
will need to use interactive perception to verify that the manipulation was successful, e.g.,
pulling on a door to ensure it was locked. Even if a property’s value can be approximated
using passive perception, interactive perception often provides a more accurate estimate.
In some cases, the estimate from the interactive perception can be used as the ground truth
value for learning the passive perception. Interactive perception thus provides the basis for
self-supervised learning. As the perception depends on the action, interactive perception
is often combined with active learning to actively select actions that maximize learning
performance.

2.4 Hierarchical Task Decompositions and Skill Reusability

Manipulation tasks exhibit a highly hierarchical structure. For example, the task of cleaning
a house can be divided into subtasks, such as cleaning the dishes, vacuuming the floors,
and disposing the trash. These subtasks can then be further divided into smaller subtasks,
such as grasping a plate or trashbag. Even basic skills, such as grasping, can be further divided into multiple goal-oriented action phases. This hierarchy divides the primary task into smaller, more tractable problems. The robot can learn skill policies for performing the lowest level tasks and then use these skills as an action basis for performing the next level of tasks. The robot can thus incrementally learn a hierarchical policy of skills, with the resulting policy hierarchy reflecting the task hierarchy. The complexity of the learning challenges at each level of the hierarchy is reduced, enabling faster skill learning.

The hierarchy is also important because it results in a modular structure. Subtasks and skills can often be interchanged to perform tasks in different ways depending on the scenario. Modularity also allows for some components to be predefined and others learned, e.g., an agent may be provided with a basic grasping reflex. More importantly, similar tasks will often appear multiple times within the hierarchy. For example, when cutting vegetables, each slice of the vegetable is a separate and slightly different task. These tasks are however similar enough that the robot should generalize across them rather than treat them as distinct with unique skills. We refer to such sets of similar tasks as *task families*. By exploiting the similarity of these tasks, the robot can efficiently learn skills across entire task families and can thus be reused multiple times. The ability to incorporate this modularity into the robot’s controllers and models is conditioned on having a suitable task decomposition. Discovering this structure autonomously by the robot is thus a major topic for manipulation research.

### 2.5 Object-Centric Generalization

One common structural assumption for manipulation tasks is that the world is made of up objects and that the robot’s goals will typically be to modify some aspect or attribute of a particular set of objects in the environment. Consequently, *generalization via objects*—both across different objects, and between similar (or identical) objects in different task instances—is a major aspect of learning to manipulate. Object-centric representations of manipulation skills and task models are often sufficient to generalize across task instances, but generalizing across different objects will require both motor skills and object models that adapt to variations in object shape, properties, and appearance. In some cases this can be done implicitly—e.g., with a compliant gripper that automatically adapts its shape to that of an object during grasping. One powerful approach to generalizing across objects is to find an abstract representation under which we can consider a family of objects to be equivalent or identical, even though they vary substantially at the pixel or feature level, and adapt accordingly.

### 2.6 Discovering Novel Concepts and Structures

Robots working in unstructured environments will often come across new types of objects. Performing new tasks with these objects may require adapting existing skills or learning entirely new skills. Learning in open-world environments is thus not just a matter of the robot filling in gaps in its knowledge base. Instead, the scope of the knowledge bases will continue to expand, sometimes in unforeseen ways. The ability to handle novel concepts is an important aspect of robot autonomy, as it allows robots to handle unforeseen situations. To operate efficiently, robots will need to be capable of generalizing and transferring knowledge
from prior experiences to structure the learning processes for these new concepts. This transfer learning may require more abstract reasoning depending on the similarity of the new concept to previous ones. Fortunately, as explained in this section, manipulation tasks exhibit a significant amount of structure that an autonomous robot can leverage when learning manipulation tasks.

3. Formalizing Manipulation Learning Tasks

Robot learning problems can typically be formulated as individual Markov Decision Processes (MDPs),\(^1\) described as a tuple:

\[
(S, A, R, T, \gamma),
\]

where \(S \subseteq \mathbb{R}^n\) is a set of states; \(A \subseteq \mathbb{R}^m\) is a set of actions; \(R(s, a, s')\) is a reward function, expressing the immediate reward for executing action \(a\) in state \(s\) and transitioning to state \(s'\); \(T(s'|s, a)\) is a transition function, giving the probability distribution over states \(s'\) reached after executing action \(a\) in state \(s\); and \(\gamma \in [0, 1]\) is a discount factor expressing the agent’s preference for immediate over future rewards. Here the goal of learning is to find a control policy, \(\pi\), that maps states to actions so as to maximize the return, or discounted sum of future rewards \(\sum_{i=0}^{\infty} \gamma^i r_i\), for that specific problem.

This common formulation captures a wide range of tasks, enabling researchers to develop broadly applicable general-purpose learning algorithms. However, as we have discussed in the previous section, the robot manipulation problem has more structure; a major task for robot manipulation researchers is to identify and exploit that structure to obtain faster learning and better generalization. Moreover, generalization is so central to manipulation learning that a multi-task formulation is needed. We therefore model a manipulation learning task as a structured collection of MDPs, which we call a task family. Rather than being asked to construct a policy to solve a single task (MDP), we instead aim to learn a policy that generalizes across the entire task family.

A task family is a distribution, \(P(M)\), over MDPs, each of which is a task. For example, a door-opening task family may model each distinct door as a separate task. Similarly, a vegetable-cutting task family may model each individual slice as a task. The action space is determined by the robot and remains the same across tasks, but each task may have its own state space and transition and reward functions. The reward function is formulated as a robot-dependent background cost function \(C\) —shared across the entire family—plus a reward \(G_i\) that is specific to that \(i^{th}\) task:

\[
R_i = G_i - C.
\]

The state space of the \(i^{th}\) task is written as:

\[
S_i = S_r \times S_{e_i}.
\]

---

\(^1\) We should note that the Partially-Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998) is a more accurate characterization of most robot learning problems. While these have been employed in some cases (Hsiao et al., 2007; Vien and Toussaint, 2015a b, for example), the difficulty of solving POMDPs, especially in the learning setting, means their usage here is so far uncommon.
where $S_r$ is the state of the robot and $S_{e_i}$ is the state of the $i^{th}$ environment. $S_{e_i}$ can vary from raw pixels and sensor values, to a highly pre-processed collection of relevant task variables. Many task environments will consist of a collection of objects and the variables describing the aspects of those objects that are relevant to the task. It is therefore common to model the environment as a collection of object states, resulting in a more structured state space where $S_{e_i}$ is partly factorized into a collection of relevant object states:

$$S_{e_i} = S_{w_i} \times \Omega^1_i \times \ldots \times \Omega^{k_i}_i,$$

where $S_{w_i}$ is the state of the general environment, $\Omega^j_i$ is the state of the $j$th relevant object in task $i$, and task $i$ contains $k_i$ objects. The number of relevant objects may vary across, and occasionally within, individual tasks. Modeling tasks in this factorized way facilitates object-centric generalization, because policies and models defined over individual objects (or small collections of objects) can be reused in new environments containing similar objects. This factorization can be clearly seen in symbolic state representations, wherein the modularity of proposition-based (e.g., $\text{CupOnTable}=\text{True}$) or predicate-based (e.g., $\text{On(Cup,Table)}=\text{True}$) representations allows the robot to consider only subsets of symbols for any given task. For manipulation tasks, we often employ predicate-based representations for their explicit generalization over objects.

Manipulation tasks also present modularity in their transition functions, i.e., the robot will only be able to affect a subset of objects and state variables from any given state. To capture the underactuated nature of manipulation tasks, we can model the tasks as hybrid systems with piecewise continuous dynamics. Each of the continuous dynamical subsystems is referred to as a mode, and the state will often contain discrete variables to capture the current mode. Mode switches occur when the robot enters certain sets of states known as guard regions, e.g., when the robot makes or breaks contact with an object. The robot can thus limit the state variables that it may alter by restricting itself to certain modes.

In some cases there is also structure in the action space, exploited through the use of higher-level actions, often called skills. Such skills are typically modeled using the options framework (Sutton et al., 1999), a hierarchical learning framework that models each motor skill as an option, $o = (I_o, \beta_o, \pi_o)$, where:

- $I_o : S \to \{0, 1\}$ is the initiation set, an indicator function describing the states from which the option may be executed.
- $\beta_o : S \to [0, 1]$ is the termination condition, describing the probability that an option ceases execution upon reaching state $s$.
- $\pi_o$ is the option policy, mapping states in the option’s initiation set to low-level motor actions, and corresponding to the motor skill controller.

The robot may sometimes be pre-equipped with a set of motor skills that it can reuse across the family of tasks; in other settings, the robot discovers reusable skills as part of its learning process.

One of the key challenges of learning a policy, or domain knowledge, across a task family is the need to transfer information from individual tasks to the whole family. A robot can learn the policy to solve a single task as a function of the task state. However,
transferring these functions across the family is not trivial as the tasks may not share the same state space. Transfer may be aided by adding extra information to the task—for example, information about the color and shape of various objects in the task—but since that information does not change over the course of a task execution, it does not properly belong in the state space. We model this extra information as a context vector $\tau$ that accompanies each task MDP, and which the robot can use to inform its behavior. Like the state space, the context can be monolithic for each task or factored into object contexts. To generalize across a task family, the robot will often have to learn policies and models as functions of the information in the context vector. For example, a cutting skill needs to adapt to material properties of different vegetable, or a door-opening skill needs to adapt to the masses and sizes of different doors.

In summary, our model of a family of manipulation tasks therefore consists of a task family specified by a distribution of manipulation MDPs, $P(M)$. Each manipulation MDP $M_i$ is defined by a tuple $M_i = (S_i, A, R_i, T_i, \gamma, \tau_i)$, where:

- $S_i = S_r \times S_{e_i}$, is the state space, where $S_r$ describes the robot state, and $S_{e_i}$ the environment state. Often the environment state is primarily factored into a collection of object states: $S_{e_i} = S_{w_i} \times \Omega^i_1 \times ... \times \Omega^i_{k_i}$, for a task with $k_i$ objects;
- $A$ is the action space, common across tasks, which may include both low-level primitive actions and a collection of options $O$;
- $R_i = G_i - C$ is the reward function, comprising a background cost function $C$ (common to the family) and a task-specific goal function $G_i$;
- $T_i$ is the transition function, which may contain exploitable structure across the sequence of tasks, especially object-centric structure;
- $\gamma$ is a discount factor, and
- $\tau_i$ is a vector of real numbers describing task-specific context information, possibly factored into object context: $\tau_i = \tau^i_1 \times ... \times \tau^i_{k_i}$, for $k$ objects.

**Overview of Learning Problems for Manipulation:** The learning problems posed by manipulation tasks can typically be placed into one of five broad categories, which we will discuss in the subsequent sections. An overview of the topics covered in this review are shown in Fig. 3.

When learning to define the state space (Sec. 4), the robot must discover the state features and degrees of freedom attached to each object in its environment. This information is assumed to be given in the traditional reinforcement learning and planning settings. That is not the case in robotics, and in particular in learning for manipulation, which involves interacting with objects that the robot’s designers do not have a priori access to. Learned representations of object states can be transferred across the task family as components of each task’s state space.

When learning a transition model of the environment (Sec. 5), the robot must learn a model of how its actions affect the task state, and the resulting background cost, for use in planning. This is closely connected to learning to define the state space. If the learned transition models and reward functions are object-centric, then they can be ported
across the task family, resulting in a natural means of object-centric generalization across tasks.

When **learning motor skills** (Sec. 6), the robot attempts to learn a motor control policy that directly achieves some goal, typically via reinforcement learning (Sutton and Barto, 1998). Here, the goal ranges from learning task-specific solution policies, to policies that can produce a solution to any task in the task family given the context vector, to useful motor skills that constitute a component of the solution policy but are not themselves a complete solution.

Given a learned motor skill, we may also **learn to characterize that motor skill** (Sec. 7), where the robot learns a description of the circumstances under which it can be successfully executed (often called preconditions, and corresponding to an option’s initiation set $I_o$), and a model of the resulting state change (often called an effect).

Finally, **learning compositional and hierarchical structure** (Sec. 8) aims to learn hierarchical knowledge that enables the robot to become more effective at solving new tasks in the family. Here, the goal is to learn component motor skills—completely specified options—and models of their operation to construct more abstract representations of the learning task.

**4. Learning Object and Environment Representations**

Modelling manipulation tasks and generalizing manipulation skills requires representations of the robot’s environment and the objects that it is manipulating. These representations serve as the basis for learning transition models, skill policies, and skill pre- and post-conditions, as discussed in later sections.
This section explains how the object-based state and context spaces of manipulation tasks can be defined and learned. We will also explain how the robot can discover objects and estimate their properties using passive and interactive perception. Since many of the extracted object properties and features may be irrelevant to learning certain components of manipulation tasks, we conclude the section by discussing how to select and learn relevant features.

4.1 Object Representations

As discussed in Section 2, a robot’s physical environment has considerable structure that it can exploit. In particular, the world can be divided into objects, each of which can be described by a collection of features or properties. Examples include movable objects such as mugs, tables, and doors, and stationary objects such as counters and walls. The robot can create a modular representation by segmenting the environment into objects and then estimating the values of their properties. This representation supports the reuse of skills by allowing the robot to efficiently generalize between similar objects across different tasks.

Object representations capture how objects vary both within tasks and across tasks of the same family. Within-task variations are captured by the state space—those features that a manipulation action can change; across-task variations are captured as part of the context space—attributes that are fixed in any specific task but aid generalization across pairs of tasks. For example, when stacking assorted blocks, the shapes and sizes of the blocks are fixed for a given task and are thus part of the context. Different stacking tasks may however use different sets of blocks, and thus the context changes across the tasks. Generalizing manipulation skills usually implies adapting, implicitly or explicitly, to variations in both context and state. For example, a versatile pick-and-place skill should generalize over the shapes of different objects (fixed in any specific task) as well as their positions in the environment (modifiable by actions during a task).

4.1.1 Types of Object Variations

Several types of within- and across-task object variations are common in the literature. **Object pose** (Pastor et al., 2009; Levine and Koltun, 2013; Deisenroth et al., 2015) variations are the most common, and must be included in the state space when they can be manipulated (e.g., by a pick-and-place skill). However, in some cases, these can be fixed within a task but vary across the task family and thus belong in the context (Da Silva et al., 2014; Kober et al., 2011; Jiang et al., 2012; Levine and Koltun, 2013) (e.g., the height of a surface in a wiping task (Do et al. 2014)). **Object shape** may vary within a task, via articulated (Niekum et al., 2015a; Sturm et al., 2010; Katz et al., 2010; Martín-Martín et al., 2016; Sturm et al., 2011), deformable (Schenck et al., 2017; Li et al., 2016; Schulman et al., 2013; Schenck and Fox, 2018a; Li et al., 2018; Seita et al., 2018; Li et al., 2018), or divisible (Lenz et al., 2015b; Wörgötter et al., 2013; Yamaguchi and Atkeson, 2016b) objects. Rigid object shapes may also vary across tasks (Burchfiel and Konidaris, 2018; Brandl et al., 2014), offering both a challenge to and opportunity for generalization (Schenck and Fox, 2018a). Similarly, objects may vary in **material properties**, which can have significant effects on manipulation but typically only vary across-tasks (e.g., cutting objects of different materi-
A Review of Robot Learning for Manipulation

als (Lenz et al., 2015b)), though there is some work on altering the material properties of manipulated objects (Chitta et al., 2011; Schenck and Stoytchev, 2012; Isola et al., 2015).

Finally, objects may vary in their interactions or relative properties, including both robot-object (Bekiroglu et al., 2013; Kopicki et al., 2016) and object-object (Stoytchev, 2005; Sturm et al., 2010; Kroemer et al., 2018; Jund et al., 2018) interactions. Objects may interact with each other resulting on constraints between them (Jain and Kemp, 2013), and manipulation may result in mode switches that add or remove constraints (Niekum et al., 2015a; Baum et al., 2017; Kroemer and Peters, 2014; Baisero et al., 2015). Variance across, rather than within, tasks is also possible—e.g., the properties of the joint connecting a cabinet and its door will remain constant for a given cabinet, but different cabinets can have different constraints (Sturm et al., 2010; Hausman et al., 2015; Niekum et al., 2015a; Dang and Allen, 2010).

4.1.2 Object Representation Hierarchies

Object models can be represented hierarchically, with layers corresponding to point-, part-, and object-level representations, each decreasing in detail and increasing in abstraction, and affording different types of generalization. This representation hierarchy mirrors the geometric structure of objects and their parts. Geometric properties and features capture where the points, parts, and objects are, and non-geometric properties tend to capture the corresponding what information defining these elements. Representations at each level of the hierarchy can capture general inherent properties as well as semantic or task-oriented properties (Dang and Allen, 2012; Myers et al., 2015; Jang et al., 2017). In addition to representing individual objects, the robot can also represent interactions between objects at the different layers of the hierarchy.

Point-level Representations: Point-level representations are the lowest level of the hierarchy, and include point cloud, pixel, and voxel representations for capturing the partial or complete shapes of objects in detail (Schenck et al., 2017; Bohg and Kragic, 2010; Klingsensmith et al., 2014; Florence et al., 2018; Choi et al., 2018). Point-level representations provide the robot with the most flexible representations for capturing important details of objects and manipulation tasks.

Each element of these representations may be associated with additional features, such as the color or material properties corresponding to this point. Segmentation methods can be used to assign labels to individual points according to which part or object they belong (Schenck and Fox, 2018a; Myers et al., 2015). Interactions may be modeled at this level of the hierarchy as contact points (Kaul et al., 2016; Dang and Allen, 2012; Rosman and Ramamoorthy, 2011; Kopicki et al., 2016; Viña et al., 2013; Su et al., 2015; Veiga et al., 2015; Piacenza et al., 2017).

Generalization across tasks can be accomplished by establishing correspondences between objects and environments at the point level. The robot may identify individual key points of objects (e.g., the tip of a tool (Edsinger and Kemp, 2006)), or use non-rigid registration or geometry warping to determine correspondences of entire sets of points across task instances (Schulman et al., 2013; Hillenbrand and Roa, 2012; Stückler and Behnke, 2014; Rodriguez and Behnke, 2018; Amor et al., 2012). These correspondences can then be used
to directly map manipulation skills between tasks or to compute higher-level representations (Stückler and Behnke, 2014).

**Part-level Representations:** Representations at the part level correspond to sets of multiple contiguous points from the lower level of the hierarchy (Sung et al., 2017a; Detry et al., 2013; Dang and Allen, 2010; Byravan and Fox, 2016), typically focusing on parts associated with certain types of manipulations. For example, a mug can be seen as having an opening for pouring, a bowl for containing, a handle for grasping, and a bottom for placing (Fagg and Arbib, 1998). Each part can then be described by a set of features describing aspects such as shape, pose, color, material, type, and surface properties. Robots can use part-level features to represent interactions or relations between the parts of different objects, or to indicate the types of interactions and constraints imposed between the parts due to the interaction (e.g., a peg must be smaller than the hole for an insertion task).

Defining correspondences at the part level enables generalization across different types of objects (Tenorth et al., 2013; Sung et al., 2015; Detry et al., 2013). Many objects have similar parts that afford similar interactions, although the objects may be very different as a whole. For example, a coin and a screwdriver are distinct types of objects, but they both have short thin metallic edge that can be used to turn a screw. Similarly, many objects have graspable handles; identifying the handle correspondences thus allows for transferring grasping skills (Detry et al., 2013; Kroemer et al., 2012). Part-based representations therefore allow the robot to generalize between different classes of objects without having to reason about individual points of the object (Sung et al., 2015).

**Object-level Representations:** Object-level representations are important because robots generally select objects, not individual features, to manipulate (Janner et al., 2019; Deisenroth et al., 2015; Gao et al., 2016; Jang et al., 2018). Thus, robots must generalize between the sets of features attached to each object. Useful object-level representations group together object-specific properties such as an object’s pose, mass, overall shape, and material properties (for uniform objects). Semantic object labels can be used to distinguish different types of objects and how they should be manipulated (Jang et al., 2017). Similar to parts, object-level interaction features often define the relative poses, forces, and constraints between objects in manipulation tasks, e.g., relative poses for stacking blocks (Fang et al., 2016; Jain and Kemp, 2013; Uğur and Piater, 2015; Sundaralingam et al., 2019). The robot may also define different types of interactions at this level (e.g., object A is on or inside object B) or relations between objects (e.g., relative sizes and weights) to capture sets of interactions in a more abstract form (Schenck et al., 2012; Schenck and Fox, 2018a; Kulick et al., 2013; Schenck and Stoytchev, 2012; Fichtl et al., 2014). Generalization across objects typically requires establishing correspondences between distinct objects (Devin et al., 2018) that support similar manipulations.

Robots may also need to represent groups of objects—rather than modeling individual objects within the group, it is often more efficient and robust to use features that represent groups of objects as a whole entity. For example, clutter, or the background scene with which the robot should avoid colliding, is often represented in a similar manner to a single deformable or piece-wise rigid object. Identifying and reasoning about individual objects may not be necessary, and may even add additional complexity to learning, causing learning to generalize poorly or be less robust. Manipulating a specific object among a group may
require the robot to recognize it, or even actively singulate the object (Gupta et al., 2015; Hermans et al., 2012).

4.2 Passive and Interactive Perception

As embodied agents capable of manipulating their surroundings, robots can use actions to enhance their perception of the environment. Robot perception is therefore broadly divided into passive and interactive perception, with the key difference being whether or not the robot physically interacts with the environment.

The term passive perception refers to the process of perceiving the environment without exploiting physical interactions with it, i.e., non-interactive perception (Isola et al., 2015)—for example, recognizing and localizing objects in a scene based on a camera image (Burchfiel and Konidaris, 2018; Tremblay et al., 2018; Wang et al., 2019; Yamaguchi and Atkeson, 2016a). Passive perception allows the robot to quickly acquire a large amount of information from the environment with little effort. Passive perception does not require the environment or the sensor to be stationary; observing a human performing a manipulation task is still passive, as the robot does not itself perform the interaction. Similarly, moving a camera to a better vantage point is still a non-interactive form of perception, as the robot does not apply force to, or otherwise alter, the state of the environment (Saran et al., 2017; Kahn et al., 2015).

In interactive perception (Bohg et al., 2017), the robot physically interacts with its surroundings to obtain a better estimate of the environment. For example, a robot may push an object to better estimate its constraints or lift an object to estimate its weight (Barragán et al., 2014; Hausman et al., 2015; Katz and Brock, 2011). Robots can use a wide range of sensor modalities to observe the effects of its interactions, including haptic, tactile, vision, and audio (Lenz et al., 2015a; Chitta et al., 2011; Högman et al., 2016; Griffith et al., 2012; Thomason et al., 2016).

The need to perform actions means that interactive perception requires more time and energy than passive perception. The benefit of interactive perception is its ability to disambiguate between scenarios and observe otherwise latent properties (Tsikos and Bajcsy, 1991; Chitta et al., 2011; Gao et al., 2016), enabling the robot to reduce uncertainty. For example, the robot may not know whether two objects are rigidly connected or simply in contact; interactive perception allows it to test each hypothesis.

Different actions will result in different effects and hence robots can learn about their environment faster by selecting more informative actions (Barragán et al., 2014; Baum et al., 2017; Dragiev et al. 2013; Otte et al., 2014; Kulick et al., 2015; Saran et al. 2017; Kenney et al., 2009). For example, shaking a container will usually provide more information regarding its contents than a pushing action would (Schenck and Stoytchev, 2012; Sinapov et al. 2014; Schenck et al., 2014). Active learning approaches usually estimate the uncertainty of one or more variables in the environment and then select actions based on the resulting entropy, information gain, or mutual information (Otte et al., 2014; Högman et al., 2016; Hausman et al., 2015; van Hoof et al., 2014; Kulick et al., 2015).

The ability to test hypothesis means that interactive perception can also be used as a supervisory signal for learning to estimate properties using passive perception (Griffith et al., 2012; Pinto et al., 2016; van Hoof et al., 2014; Nguyen and Kemp, 2014; Wu et al.,
As an example of interactive learning, a robot may learn to predict the mass of objects from their appearance by first using interactive perception to determine the masses of a set of training objects. This form of self-supervised learning allows the robot to gather information autonomously and is thus crucial for enabling robots to operate in unfamiliar environments.

### 4.3 Learning About Objects and Their Properties

Having explained the different types of object variations and types of perception, we now discuss how the robot can learn about the objects around it from data.

#### 4.3.1 Discovering Objects

A common first step in learning is distinguishing the individual objects in the scene, which is a segmentation problem, typically accomplished using passive perception (Kraft et al., 2008; Alexe et al., 2012; Schwarz et al., 2018; Byravan and Fox, 2016; He et al., 2017). However, objects may often be close together in the scene, presenting the robot with ambiguous information about object identity. Here the robot can maintain a probabilistic belief over whether or not different parts belong to the same object (van Hoof et al., 2014; Kenney et al., 2009; Hausman et al., 2012), and use interactive perception or viewpoint selection to disambiguate the scene (Gualtieri and Platt, 2017; Hermans et al., 2012; van Hoof et al., 2014; Hausman et al., 2012).

#### 4.3.2 Discovering Degrees of Freedom

Once individual objects have been identified, the robot may need to identify their kinematic degrees of freedom (Niekum et al., 2015a; Jain and Niekum, 2020; Hausman et al., 2012; Katz et al., 2010; Baum et al., 2017; Sturm et al., 2010; Katz and Brock, 2011; Abbatematteo et al., 2019). These constraints and articulated connections are fundamental to establishing the objects’ state space for manipulation tasks, as well as for robust pose tracking (Desingh et al., 2018; Schmidt et al., 2015). The different types of joints are usually represented using distinct articulation models, e.g., revolute or prismatic, with their own sets of parameters. For example, a revolute joint model is specified by the position, direction, and limits of its rotation axis (Sturm et al., 2010; Barragán et al., 2014; Niekum et al., 2015a; Sturm et al., 2011). The robot must estimate these context parameters to accurately model the degrees of freedom. Passive perception can be used to estimate the kinematic chains of articulated objects, especially if the object is being moved by a person (Martín-Martín et al., 2016; Niekum et al., 2015a; Pillai et al., 2014; Brookshire and Teller, 2016; Jain and Niekum, 2020). Interactive perception is also well suited for estimating the articulation model parameters and the resulting motion constraints (Sturm et al., 2010; Barragán et al., 2014; Hausman et al., 2015). Given the high-dimensional parameter space of the articulation models, active learning approaches are often used to select informative actions for quickly determining the model parameters (Hausman et al., 2015; Baum et al., 2017), or transfer learning methods may be used (Abbatematteo et al., 2019).
4.3.3 Estimating Object Properties

Once the robot has identified an object in the environment, the next step is to estimate the object’s properties. Since some properties are only applicable to some classes of objects (Diuk et al., 2008; Bullard et al., 2018; Dragiev et al., 2013; Bhattacharjee et al., 2015; Schenck and Stoytchev, 2012; Wu et al., 2016; Chitta et al., 2011) the robot must first recognize the object class. For manipulation tasks, object classes are often derived from the actions and interactions that they afford the robot, e.g., container, graspable, and stackable, such that the interaction-relevant properties can be easily associated with the objects.

Here the robot may use passive and interactive perception to estimate object property values (Gao et al., 2016; Isola et al., 2015; Li et al., 2014; Varley et al., 2019). In addition to object recognition, passive perception is often used to estimate the position, shape, and material properties (Li et al., 2018; Bhattacharjee et al., 2015; Wu et al., 2016; Tremblay et al., 2018; Garcia Cifuentes et al., 2017; Schmidt et al., 2015; Burchfiel and Konidaris, 2017; Issac et al., 2016; Burchfiel and Konidaris, 2018; Wüthrich et al., 2015; Yamaguchi and Atkeson, 2016a). However, additional interactive perception can often be used to acquire more accurate estimates of these properties (Chitta et al., 2010; Björkman et al., 2013; Dragiev et al., 2013; Javdani et al., 2013; Petrovskaya and Khatib, 2011; Koval et al., 2017). Material and interaction properties are often difficult to estimate accurately using only passive perception. The estimates can be significantly improved by interacting with the objects and using tactile sensing (Sinapov et al., 2011b, 2014; Sung et al., 2017b; Gao et al., 2016; Lenz et al., 2015a; Hsiao et al., 2010; Varley et al., 2019). Exploratory actions can also be used to estimate the dynamic properties of objects, e.g., the center of mass, or the contents of containers (Guler et al., 2014; Schenck et al., 2014; Chitta et al., 2011).

4.4 Feature Learning and Selection

Even though an environment may contain a lot of objects and sensory stimuli, only a few object properties or sensory signals will usually be relevant for a given task. For example, when opening a bottle, the size of the cap is a relevant feature, but the color of the chair is irrelevant. Using a suitable set of relevant features simplifies the skill and model learning problems. It also increases robustness and generalization to new situations. If the set of object properties is sufficiently rich, then the robot may only need to select a suitable set of these as features for learning (Devin et al., 2018; Kroemer and Sukhatme, 2017; Montesano et al., 2008; Song et al., 2011; Stramandinoli et al., 2018). In many cases, the robot will however need to learn a set of features for the given task.

Unsupervised feature learning methods extract features from unlabeled training data. Dimensionality reduction methods may be used to capture correlations in data and discard noisy components of signals. For manipulation domains, dimensionality reduction approaches can be used to learn compact representations of complex objects or variations within classes of objects (Burchfiel and Konidaris, 2017; Bergström et al., 2012). Dimensionally reduction methods can also be used to reduce the robot’s action space of grasps to better fit common object shapes (Ciocarlie et al., 2007). Clustering methods are used to cluster together similar data samples. For manipulations, clustering method may for example be used to cluster together distinct types of objects or continuous effects of manipulations (Ugur and Piater, 2015).
Supervised approaches learn features as part of the overall model or skill learning process. Decision trees and neural networks are often used to learn features in a supervised setting. Deep learning in particular has become ubiquitous for feature learning in robotics. Different network structures and layers, such as auto-encoders, spatial soft max layers, convolutions, and segmentation masks, can be used to incorporate architectural priors for learning useful features. For example, segmentation masks can be used to capture the co-movement of nearby points in an image to capture points of objects or parts moving together (Byravan and Fox, 2016; Finn and Levine, 2017). Deep neural network models can be used to represent state spaces for manipulation tasks with high dimensional observation spaces. Representations of manipulation environments can be learned in a task-oriented manner to facilitate subsequent planning and control (Kurutach et al., 2018; Srinivas et al., 2018). Neural networks are also highly effective at combining data from multiple sensor modalities or information sources. For manipulation tasks, robots often use this approach to merge information between passive modalities (e.g., vision) and more interactive modalities (e.g., touch and haptics), or to incorporate additional task information (e.g., instructions) (Gao et al., 2016; Lee et al., 2019; Sung et al., 2017a).

5. Learning Transition Models

The goal of manipulation tasks is to alter the states of objects in the robot’s environment. Learning transition models that capture the changes in state as a result of the robot’s actions is therefore an important component of manipulation learning.

5.1 Representing and Learning Transition Models

Learning a transition model requires a suitable representation. The general form of a transition model for a family of MDPs is either a deterministic function $T : S \times A \rightarrow S$ or a stochastic distribution $T : S \times A \times S \rightarrow R$ over the next states given the current state and action. The transition models can also depend upon the context vector $\tau$ in order to explicitly generalize between contexts. A transition model is often employed iteratively to perform multi-step predictions.

**Continuous Models:** In manipulation tasks, robots are generally operating in continuous state and action spaces, e.g., the state is given by the continuous poses of objects and the actions are given by desired joint positions or torques. Robots therefore often learn low-level transition models for predicting the next state as a set of continuous values, even when the set of actions is discrete (Scholz and Stilman, 2010). Regression methods, such as neural networks, Gaussian processes, and (weighted) linear regressions, are commonly used for learning transition models (Schenck et al., 2017; Levine and Koltun, 2013; Deisenroth and Rasmussen, 2011; Atkeson et al., 1997a).

The simplest models, such as linear regression models, can be learned from small amounts of data (Stilman et al., 2008), but often underfit the inherent nonlinearities of most manipulation tasks. Hence, the generalization performance of these models is limited. Instead, time-dependent linear models are often used to learn local models for specific regions of the state space (Kumar et al., 2016), e.g., around the trajectory of a skill being learned. These models may be sufficient while maintaining low data requirements, and can be combined with prior models that capture more global information about the task. Gaus-
sian mixture models and neural networks have both been used to learn such informative priors (Fu et al., 2016).

Nonparametric models, including Gaussian processes and locally weighted regression, can flexibly learn detailed transition models (Deisenroth et al., 2015; Kopicki et al., 2017). The generalization performance of these types of models depends heavily on their hyperparameters and is often fairly limited. However, local models can be learned from relatively small amounts of data. For example, an accurate model for planar pushing of a block may be learned from less than a hundred samples (Bauza and Rodriguez, 2017).

More complex hierarchical models, such as neural networks and decision trees, allow the robot to learn task-specific features that provide better generalization (Lenz et al., 2015a; Schenk et al., 2017; Fu et al., 2016; Janner et al., 2019; Sanchez-Gonzalez et al., 2018; Finn and Levine, 2017; Kurutach et al., 2018; Srinivas et al., 2018; Ye et al., 2018). For example, convolutional encoder-decoder networks can be used to learn models for predicting the transitions of individual points in the scene based on the selected actions (Byravan and Fox, 2016; Finn and Levine, 2017). However, unless the models are pre-trained, learning intermediate features significantly increases the amount of data required for training these models. Broadly speaking, the best representation for learning a continuous transition model often depends on the amount of training data available and the model expressiveness required by the task.

Predictive state representations (PSRs) allow the robot to model the distributions over future observations based on the history of past observations (Littman and Sutton, 2002; Singh et al., 2004; Boots et al., 2013; Stork et al., 2015). These models thus capture latent state information, as part of a POMDP formulation, without explicitly modeling or inferring the latent state. The robot can thus work directly with the observed variables.

More advanced parametric models provide additional structure, up to including entire physics engines (Wu et al., 2015; Scholz et al., 2014; Wu et al., 2016; Li et al., 2015). These engines provide a significant amount of prior structure and information, which improves generalization and data efficiency. However, even given the engine, it is difficult for a robot to map a real world scenario into a simulation environment. The simulation may also still fail to accurately capture certain interactions due to their lack of flexibility. Physics engines are therefore sometimes used to provide a prior or to initialize a more flexible representation by combining analytical models with flexible data-driven models (Ajay et al., 2018).

**Discrete Models:** Discrete transition models are used to learn transitions for tasks with discrete state and action spaces, typically capturing high-level tasks. For example, a robot may learn a discrete transition model for flipping a pancake, in which the transition probabilities indicate the probability of the pancake being flipped, dropped, or not transitioning. The model can then be used to predict the outcome of attempting multiple flips. Discrete models can also capture low-level transitions where the state space has been discretized. This results in the continuous states corresponding to a single discrete state being considered as identical. It also removes the smoothness assumption between states, such that state transition information is not shared between discrete states. As a result, a very fine-grained discretization will lead to poor generalization and a coarse discretization may not capture enough detailed transition information to perform the task. A suitable discretization should therefore cluster together states that will result in similar transitions.
given the same actions (Bergström et al., 2012). Adaptive or learned discretizations are a form of feature learning (Guan et al., 2018).

The most basic representations for discrete models are tabular models or finite state machines. Discrete models treat each state as atomic, with a transition distribution independent of all other states. This approach however severely hampers generalization and transfer. More structured models use collections of symbols to represent discrete state spaces. Proposition-based models define a set of propositions (Konidaris et al., 2018), e.g., Obj1Grasped and Obj2Clear, which are either true or false and thus define the state. A set of N propositions thus defines a set of $2^N$ states. The factorization provided by the propositions results in a more compact representation of the state space, and also allows transition information to be shared between these states. In particular, symbolic models are often used to express compact (and sometimes stochastic) action models known as operators. Each such operator has a precondition that must be true for the action to be executable, e.g., WaterReservoirFull=True and CoffeeJugFull=False for a coffee making action. The effects of the action are then given by lists of propositions that become true (e.g., CoffeeJugFull) and that become false (e.g., WaterReservoirFull), with all other propositions being unchanged by the transition. As the preconditions and effects only refer to a small number of propositions, the operator generalizes across the values of all other propositions that are not mentioned.

These models may be generalized to relational or first-order models (Lang et al., 2012), which support extremely compact action models and enable generalization across objects. This approach allows predicate-based models (e.g., Full(CoffeeJug) for a full jug of coffee and Full(WaterReservoir) for a full water reservoir) to be used for families of tasks with different numbers of objects and states. This extended generalization can unfortunately also have some unintended effects if distinct situations are treated as the same. For example, a dishwasher loading action would treat a paper plate the same as a ceramic plate unless a suitable predicate, e.g., IsCeramic(Plate), were included to avoid this type of overgeneralization.

**Hybrid Models:** Hybrid models combine elements of discrete and continuous models, often resulting from the hierarchical structure of manipulation skills and subtasks. The discrete components of the state are often used to capture high-level task information while the continuous components capture low-level state information.

Robots will often learn continuous models for specific subtasks. For example, a robot may learn separate models for opening a door, moving its body through free space, and stacking newspapers. These models can then be used to predict the effects of actions within each skill. In turn, these models can then become a part of the robot’s overall world model to predict the effects of sequences of skills. This approach allows the robot to exploit the modularity of the distinct subtasks for the discrete high-level transition model. Many works on learning continuous models for individual skills or subtasks can thus be seen as learning part of a larger hybrid model.

One common use of hybrid models is to represent modes in manipulation tasks (Hauser, 2018; Siméon et al., 2004; Kroemer et al., 2015; Lee et al., 2017; Choudhury et al., 2017; Zhou et al., 2018). In these cases, the dynamics are piece-wise continuous with jumps between the modes. These jumps are often defined by guard sets wherein the state transitions to a certain mode if the continuous state enters the guard set. In manipulation tasks, the jumps
between modes typically result from contacts being made or broken between objects. The mode switches thus capture the changing dynamics and constraints caused by the changing contacts. Planning approaches for manipulation domains often explicitly take this multimodal structure into account for planning (Toussaint et al., 2018; Jain and Niekum, 2018). Learning a hybrid model entails learning both the dynamics within each mode and the conditions for switching between modes. The dynamics within a mode can be learned using standard continuous model approaches. The guard regions can be modeled as explicit sets of states or using classifiers. More flexible models, such as nonparametric and neural network models, may also be able to implicitly capture hybrid dynamics of manipulation tasks (Finn and Levine, 2017). Exploiting the modularity of their hybrid system is however more difficult when using an implicit model of the mode structure.

5.2 Uncertainty in Transition Models

State transitions are often represented using probabilistic models that allow the robot to represent multiple potential outcomes. For example, a robot may choose to move to a trash can and drop an object inside because the resulting state distribution has a lower variance than attempting to throw the object into the trash. When working with probabilistic models, it is important to distinguish between two sources of uncertainty: aleatoric and epistemic uncertainty.

**Aleatoric uncertainty:** A stochastic process exhibits a randomness in its state transitions. For example, when throwing a block into the air, the side on which the block will land is random. Similarly, due to slight variations in the execution and surface properties, the exact location of an object being placed on a table may follow a Gaussian distribution. Attempting to open a door may only result in the door opening 98% of the time and remaining closed the other 2%.

Stochasticity resulting in aleatoric uncertainty is common in manipulation tasks, for both discrete and continuous state spaces (Kopicki et al., 2017; Lang et al., 2012). Probabilistic models are thus used to capture the process noise and approximate the distributions over next states (Babaeizadeh et al., 2018; Kroemer et al., 2015).

**Epistemic uncertainty:** In addition to the inherent stochasticity of manipulation, the outcome of an action may be uncertain due to the robot’s limited knowledge of the process. For example, consider the task of placing a closed container into a bowl of water. Whether the container will float or sink is deterministic given its mass and volume. However, the robot’s prediction over the two outcomes may still be uncertain if the robot does not know the mass of the container or because it lacks sufficient training samples to accurately predict the mass threshold at which the container will sink. The probabilistic prediction is thus the result of epistemic uncertainty, or uncertainty over the model parameters, rather than aleatoric uncertainty from inherent stochasticity. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced given additional information, e.g., using interactive perception to better estimate the mass of the container or acquiring additional training samples for estimating the threshold for the model.

Bayesian approaches explicitly capture epistemic uncertainty in their predictions by representing the probability distributions over different potential models (Deisenroth et al., 2015; Scholz et al., 2014). They also incorporate explicit prior beliefs over these model
distributions. By contrast, point estimates, such as maximum likelihood or maximum a posteriori MAP estimates, assume a single model given the current data and use this model to predict the next state. Linear Bayesian models are often too simple to capture the complexity of manipulation tasks. Kernel-based Gaussian processes have therefore become one of the most popular Bayesian approaches for learning transition models (Deisenroth et al., 2015; Högman et al., 2016). Gaussian processes, as well as linear Bayesian regression, both have explicit parameters modeling the stochasticity of the system being represented. The model uncertainty can also be used to guide exploration for learning an accurate model from fewer samples (Wang et al., 2018).

5.3 Self-supervision and Exploration for Learning Transitions

Transition models are usually learned in a self-supervised manner. Given the current state, the robot performs an action and observes the resulting effect on the state. The robot can thus acquire state, action, and next state tuples for training the model. To generalize more broadly, the robot will also need to estimate the context parameters for each task and then incorporate these into the model as well.

The robot can adopt different exploration strategies for acquiring samples. Random sampling is often used to learn general models and acquire a diverse set of samples. A grid approach may be used to ensure that the action samples are sufficiently spread out, but these approaches generally assume that the state can be reset between actions (Yu et al., 2016). Active sampling approaches can be used to select action samples that are the most informative (Wang et al., 2018). If the model is being used to learn and improve a specific skill, the robot may collect samples within the vicinity of the skill using a small amount of random noise (see model-based skill learning in the next section). Another increasingly popular approach to exploring an environment is to use intrinsic motivation (Chentanez et al., 2005; Pathak et al., 2019). In this case, the robot actively attempts to discover novel scenarios where its model currently performs poorly or that result in salient events.

5.4 Transferring and Reusing Transition Models

Transition models are not inherently linked to a specific task, and can therefore often be transferred and reused between different manipulation tasks and even task families. For example, a model learned for pushing an object may be used as the basis for learning push grasps of objects. In order to directly reuse a model, the learned model and the new task must have the same state, action, and context spaces, or a mapping between the spaces may be necessary (Taylor et al., 2007). Given compatible spaces, the ability to transfer or reuse models depends on the overlap between data distributions for training the old and new task models. This issue is known as covariate shift (input varies) and dataset shift (input and output varies). If the previous model was trained on data distinct from the needs of the new task, then the benefit of using the previous model will be limited and may even be detrimental. Assuming a sufficiently rich model, the robot can acquire additional data from the new task and use it to update the model for the new region. In this manner, the robot’s learned model may become more applicable to other tasks in the future. Data can also be shared across models when learning skills in parallel.
6. Learning Skill Policies

All of the robot learning problems discussed thus far are ultimately in service of helping the robot learn a policy that accomplishes some objective. Thus, the final learning goal for the robot is to acquire a behavior, or skill controller, that will perform a desired manipulation task. A common representation for skill controllers is a stochastic policy that maps state-action pairs to probabilities (or probability densities in the continuous case). This section discusses the spectrum of different policy parameterizations that can be chosen, algorithms for learning skill policies from experience and demonstrations, methods for transferring knowledge across skills and tasks, and approaches that can provide safety and performance guarantees during learning.

6.1 Types of Action Spaces

The choice of action spaces is an important part of designing a manipulation policy. The robot ultimately needs to send a control signal to its actuators to perform actions in the physical world. These signals may, for example, define the desired pressure for a hydraulic or pneumatic actuator (Bischoff et al., 2014; Gupta et al., 2016), the heating for a shape memory alloy actuator (AbuZaiter et al., 2016; Zimmer et al., 2019), the tendon activation for a cable-driven robot (Schlagenhauf et al., 2018; Andrychowicz et al., 2020), or the torque for an electric motor (Nguyen-Tuong and Peters, 2011. Levine et al., 2015). Robots that are difficult to model, such as soft robots and robots with complex dynamics, may often benefit from learning policies that directly output these control signals. However, the outputs of policies often do not work directly at the level of actuator signals.

In practice, an additional controller is often placed between the policy and the actuator (Gullapalli et al., 1994). Using an additional controller allows the robot to leverage a large body of prior work on control for robotics (Spong et al., 2005). Example controllers include simple linear PID controllers as well as more complex model-based admittance and impedance controllers. The policy’s actions then define the desired values for the controller. The action space of the policy often defines the desired positions, velocities, or accelerations for movements or desired forces and torques for interactions.

Both desired position and force information can be defined in the joint space or a Cartesian task space for the end effector. For manipulation tasks, it is often easier to generalize interactions with objects across the robot’s workspace using a Cartesian action space (Mason, 1981; Ballard, 1984). For example, with a Cartesian action space, applying an upward force on a grasped object would be the same action anywhere in the robot’s workspace. The controller is then responsible for mapping these desired signals into the joint space for actuation. For a joint-space action policy, the robot would need to learn different joint torques depending on the arm’s current configuration.

Although the Cartesian actions could be defined in a global or robot-centric coordinate frame, additional generalization can often be achieved by using task frames associated to individual objects or salient features of the environment (Ballard, 1984). These task frames may be predefined, selected, or learned. Using a given object-relative task frame allows the policy learning to focus on how to perform the task rather than where to perform it.

In addition to determining the desired values for inputs to the controller, the policy may also output additional values for adapting the controller. In particular, a policy may define
different controller gains as part of its action space (Buchli et al., 2011; Ren et al., 2018). In this manner, the policy can make a robot more compliant or stiffer at different points during the task execution. Otherwise the gains are usually predefined and fixed.

The inclusion of a controller also allows the robot to use a policy that operates at a lower frequency than the controller. While the low-level controllers may operate at 100s or 1000s of Hertz, the policies can operate at lower frequencies. For policies operating at lower frequencies, an additional interpolation step may be used to guide the controller between the desired values.

6.2 The Spectrum of Policy Structure

In robotic manipulation, specific parameterizations are often used that restrict the representational power of the policy; if these restrictions respect the underlying structure of the task, generalization and data efficiency are often improved without significantly impacting asymptotic performance. Thus, the choice of policy representation is a critical design decision for any robot learning algorithm, as it dictates the class of behaviors that can be expressed and encodes strong priors for how generalization ought to occur. This results in a spectrum of policy structures, ranging from highly general (but often sample-inefficient) to highly constrained (but potentially more sample-efficient) representations.

Nonparametric Policies: The most expressive policy representations are nonparametric, growing as needed with task or data complexity. This category includes nearest neighbor-based approaches, Gaussian processes (Rasmussen, 2004), Riemannian Motion Policies (Ratliff et al., 2018), and locally-weighted regression (Atkeson et al., 1997b; Vijayakumar and Schaal, 2000). These representations are the most flexible and data-driven, but they also typically require large amounts of data to produce high-quality generalization. This representational paradigm has been successful in manipulation tasks in which very little task knowledge is given a priori to the robot, such as an 88-dimensional octopus arm control problem in which Gaussian Process Temporal Difference learning is used (Engel et al., 2006), and 50-dimensional control of a SARCOS arm via locally weighted projection regression (Vijayakumar and Schaal, 2000).

Generic Fixed-size Parametric Policies: More commonly, fixed-complexity parametric policy representations are used, making stronger assumptions about the complexity and structure of the policy. Common parametric policy representations include look-up tables (Sutton and Barto, 1998), linear combinations of basis functions such as tile coding, the Fourier basis (Konidaris et al., 2011b, 2012), neural networks (Levine et al., 2016), decision tree classifiers (Quinlan, 1986; Huang and Liang, 2002), and support vector machines (Cortes and Vapnik, 1995; Ross et al., 2011). While any given choice of parameters has a fixed representational power, a great deal of engineering flexibility remains in choosing the number and definition of those parameters. Design decisions must also be made about which features will interact and which will contribute independently to the policy (e.g., the fully-coupled vs. independent Fourier basis (Konidaris et al., 2011b)), or more general decisions regarding inductive bias and representational power (e.g., neural network architectures).

The choice of a fixed parameterization makes particular assumptions about how generalization ought to occur across state-action pairs. For example, tabular representations can
represent any (discrete) function, but do not intrinsically support generalization to unseen states and actions. By contrast, policies comprised of linear combinations of basis functions (for both discrete and continuous state spaces) more naturally generalize to novel situations since each parameter (a basis weight) affects the policy globally, assuming basis functions with global support. However, the success of such generalization relies on the correctness of assumptions about the smoothness and “shape” of the policy, as encoded by the choice of basis functions. Some parametric forms make more specific structural assumptions—for example, by construction, convolutional neural networks (Krizhevsky et al., 2012; Levine et al., 2016; Agrawal et al., 2016; Mahler et al., 2017) support a degree of invariance to spatial translation of inputs.

**Restricted Parametric Policies:** The expressiveness of a policy is only limited by the underlying policy representation; however, sample complexity, overfitting, and generalization often get worse as representational power increases. This has lead to the development of specialized policy representations that have limited representational power, but exploit the structure of robotics problems to learn and generalize from less data. However, the quality of the generalization in such methods is highly dependent upon the accuracy of the underlying assumptions that are made. Such methods lie on a spectrum of how restrictive they are; for example, structured neural network architectures (e.g., value iteration networks (Tamar et al., 2016), schema networks (Kansky et al., 2017), etc.) impose somewhat moderate restrictions, whereas other methods impose much stronger restrictions, such as forcing the policy to obey a particular set of parameterized differential equations (Schaal, 2006). However, it is not always clear how to choose the right point along the spectrum from a completely general to highly-specific policy class. The following are some common examples of restricted policy classes.

Linear Quadratic Regulators (LQR) (Zhou et al., 1996) are commonly used to stabilize around (possibly learned) trajectories or points, in which the cost is assumed to be quadratic in state and the dynamics are linear in state. These restrictions are relaxed in Iterative LQR (Fu et al., 2016) and Differential Dynamic Programming (Tassa et al., 2014; Yamaguchi and Atkeson, 2015), in which nonlinear dynamics and nonquadratic costs can be approximated as locally linear and locally quadratic, respectively. LQR-based controllers have also been chained together as LQR-trees, allowing coverage of a larger, more complex space (Tedrake et al., 2010). LQR-like methods also assume that the state is fully observable; Linear Quadratic Gaussian (LQG) control methods further generalize LQR by using a Kalman filter for state estimation, in conjunction with an LQR controller (Levine et al., 2015; Platt et al., 2010). LQR/LQG controllers are optimal when the linear-quadratic assumptions are met, but require full knowledge of the dynamics and cost function, as well as a known trajectory around which to stabilize. Thus, generalization in this case simply translates to being able to stabilize around this known trajectory optimally from anywhere in the state space.

Dynamic Movement Primitives (DMPs) support the learning of simple, generalizable policies from a small number of demonstrations, which can be further improved via reinforcement learning (Schaal, 2006; Pastor et al., 2009). DMPs leverage the fact that many robotic movements can be decomposed into two primary parts—a goal configuration in some reference frame and a “shape” of the motion. For example, screwing in a screw requires a turning motion to get to the desired final configuration with respect to the hole.
DMPs use a set of differential equations to implement a spring-mass-damper that provably drives the system to an adjustable goal from any starting location, while also including the influence of a nonlinear function that preserves the desired shape of the movement. DMP controllers can be learned from very little data, but typically only generalize well in cases in which a good solution policy broadly can be described by a single prototypical motion shape primitive.

Whereas DMPs generate deterministic policies, other approaches treat the problem probabilistically, allowing stochastic policies to be learned. ProMPs are a straightforward probabilistic variant of DMPs that can produce distributions of trajectories (Paraschos et al., 2013). Gaussian Mixture Regression (GMR) (Calinon et al., 2007) models the likelihood of states over time as a mixture of Gaussians, allowing for a multimodal distribution of trajectories that can encode multiple ways of performing a task, as well as the acceptable variance in different parts of the task.

Goal-based Policies: At the far end of the spectrum, the most restrictive policy representations are primarily parameterized by a goal configuration. Given that the goal configuration is the primary parameter, these methods are typically either fixed strategies (such as splining to a goal point (Su et al., 2016), or between keyframes (Akgun et al., 2012)) or have a very small number of adjustable parameters, such as a PID controller or motion planner.

6.3 Reinforcement Learning

For any given policy representation, reinforcement learning (RL) (Sutton and Barto, 1998) can be used to learn policy parameters for skill controllers. In robotics domains, tasks addressed with RL are usually episodic, with a fixed number of time steps or a set of terminal states that end the episode (e.g., reaching a particular object configuration), but occasionally may be continuing tasks with infinite horizons (e.g., placing a continual stream of objects into bins as fast as possible).

There are many challenges in applying RL to robotics. Due to the time it takes to collect data on physical robots, the tradeoff between exploration and exploitation becomes significantly more important, compared to problems for which fast, accurate simulators exist. Furthermore, few problems in robotics can be strictly characterized as MDPs, but instead exhibit partial observability (Platt et al., 2011) and nonstationarity (Padakandla et al., 2019) (though it often remains practical to cast and solve these problems as MDPs, nonetheless). Since many tasks in robotics are multi-objective (e.g., pick up the mug, use as little energy as possible, and don’t collide with anything), it can be difficult to define appropriate reward functions that elicit the desired behavior (Hadfield-Menell et al., 2016). Due to the episodic nature of most robotics tasks, rewards tend to be sparse and difficult to learn from. Robotics problems also often have high-dimensional continuous state features, as well as multi-dimensional continuous actions, making policy learning challenging.

Here, we categorize RL algorithms using three primary criteria: (1) model-based or model-free, (2) whether or not they compute a value function, and in what manner they use it, and (3) on-policy or off-policy.

Model-Based RL: Accurate models of transition dynamics are rarely available a priori and it is often difficult to learn transition models from data. Nonetheless, there have been
notable model-based successes in robotic manipulation in which an approximate model is learned from data (Deisenroth and Rasmussen, 2011; Levine et al., 2015; Lenz et al., 2015a; Kumar et al., 2016; Finn and Levine, 2017; Schenck and Fox, 2018b). In these examples, the model is typically used to guide exploration and policy search, leading to greatly improved data efficiency. More generally, the primary benefits of model-based RL in robotics are that (1) in some domains, an approximate model is simpler to learn than the optimal policy (since supervised learning is generally easier than RL) (2) models allow for re-planning / re-learning on-the-fly if the task changes, and (3) models allow for certain types of data collection that are difficult or impossible in the real world (such as resetting the world to an exact state and trying different actions to observe their outcome). The primary disadvantage of model-based methods, aside from the difficulty of obtaining or learning models, is that incorrect models typically add bias into learning. However, some methods mitigate this by directly reasoning about uncertainty (Deisenroth and Rasmussen, 2011), whereas methods such as doubly robust off-policy evaluation (Jiang and Li, 2016) use a model as a control variate for variance reduction, sidestepping the bias issue. More details on model learning can be found in Section 5.

**Model-Free RL:** Model-free methods learn policies directly from experiences of the robot, without building or having access to a model. When the dynamics of the environment are complex, as they often are in contact-rich manipulation tasks, it can be significantly easier to learn a good policy than a model with similar performance. Model-free methods can learn how to implicitly take advantage of complex dynamics without actually modeling them—for example, finding a motion that can pour water successfully without understanding fluid dynamics. The downside of the model-free approaches is that they cannot easily adapt to new goals without additional experience in the world, unlike model-based approaches. Furthermore, they typically require a large (sometimes prohibitive) amount of experience to learn. For this reason, some recent approaches have sought to combine the benefits of model-based and model-free methods (Gu et al., 2016; Chebotar et al., 2017; Pong et al., 2018; Feinberg et al., 2018; Englert and Toussaint, 2016).

**Value Function Methods:** These methods aim to learn the value of states (or more commonly, state-action pairs)—the expected return when following a particular policy, starting from that state(-action pair). Value functions methods are known to be low variance, highly sample efficient, and in discrete domains (or continuous domains with linear function approximation), many such methods can be proven to converge to globally optimal solutions. However, value function methods are typically brittle in the face of noise or poor or underpowered state features, and do not scale well to high-dimensional state spaces, all of which are common in robot manipulation tasks. Furthermore, they are not compatible with continuous actions, severely limiting their application to robotics, except where discretization is acceptable. Actions are often discretized at higher levels of abstraction when they represent the execution of entire skills, rather than primitive actions (Kroemer et al., 2015). Finally, it is difficult to build in useful structure and task knowledge when using value function methods, since the policy is represented implicitly. Nonetheless, value function-based methods such as Deep Q-Networks (Mnih et al., 2015) and extensions thereof have been applied successfully to robotics tasks (Zhang et al., 2015; Kalashnikov et al., 2018).

**Policy Search Methods:** Policy search methods parameterize a policy directly and search for parameters that perform well, rather than deriving a policy implicitly from a value
function. Policy search methods have become popular in robotics, as they are robust to noise and poor features and naturally handle continuous actions, (the direct parameterization of the policy eliminates the need to perform a maximization over values of actions). They also often scale well to high-dimensional state spaces since the difficulty of policy search is more directly related to the complexity of a good (or optimal) policy, rather than the size of the underlying state space.

Pure policy search approaches eschew learning a value function entirely. These are sometimes referred to as “actor-only” methods, since there is a direct parameterization of the policy, rather than a “critic” value function. Actor-only approaches include the gradient-based method REINFORCE (Williams, 1992), for use when the policy is differentiable with respect to its parameters. However, REINFORCE tends to suffer from high variance due to noisy sample-based estimates of the policy gradient (though a variance-reducing baseline can be used to partially mitigate this) and is only locally optimal. Actor-only policy search also includes gradient-free optimization methods (De Boer et al., 2005; Mannor et al., 2003; Hansen and Ostermeier, 2001; Davis, 1991; Theodorou et al., 2010), which are usable even when the policy is non-differentiable, and in some cases (e.g., genetic algorithms) are globally optimal, given sufficient search time. Unsurprisingly, these advantages often come at the cost of reduced sample efficiency compared to gradient-based methods.

In contrast to actor-only approaches, actor-critic policy search methods (Sutton et al., 2000; Peters and Schaal, 2008; Peters et al., 2010; Lillicrap et al., 2015; Mnih et al., 2016; Schulman et al., 2015, 2017; Wang et al., 2017; Wu et al., 2017; Haarnoja et al., 2018) use both a value function (the critic) and a directly parameterized policy (the actor), which often share parameters. These methods typically have most of the advantages of actor-only methods (the ability to handle continuous actions, robustness to noise, scaling, etc), but use a bootstrapped value function to reduce the variance of gradient estimates, thereby gaining some of the sample efficiency and low-variance of critic-only methods. The unique advantages of actor-critic methods have made them state of the art in many robotic manipulation tasks, as well as in the larger field of deep reinforcement learning.

**On-Policy vs Off-Policy Learning:** When using RL to improve a policy, on-policy algorithms are restricted to using data collected from executions of that specific policy (e.g., SARSA (Sutton and Barto, 1998)), whereas off-policy algorithms are able to use data gathered by any arbitrary policy in the same environment (e.g., Q-learning (Watkins and Dayan, 1992)). This distinction has several consequences which are notable in robotic domains for which data collection has significant costs. On-policy algorithms are not able to re-use historical data during the learning process—every time the behavior policy is updated, all previously collected data becomes off-policy. It is also often desirable to use collected data to simultaneously learn policies for multiple skills, rather than only a single policy; such learning is impossible in an on-policy setting. However, despite these advantages, off-policy learning has the significant downside of being known to diverge under some conditions when used with function approximation, even when it is linear (Sutton and Barto, 1998). By contrast, on-policy methods are known to converge under mild assumptions with linear function approximation (Sutton and Barto, 1998).

**Exploration Strategies:** A significant, but often overlooked, part of policy learning is the exploration strategy employed by the agent. This design decision can have enormous impact on the speed of learning, based on the order in which various policies are explored.
A Review of Robot Learning for Manipulation

In physical robotic manipulation tasks, this is a particularly important choice, as data is much more difficult to collect than in simulated domains. Furthermore, in the robotics setting, there are concerns related to safety and possible damage to the environment during exploration, as will be further discussed in Section 6.6.

The most common exploration strategies involve adding some form of noise to action selection, whether it be in the form of direct perturbation of policy parameters (Theodorou et al., 2010), Gaussian noise added to continuous actions (Lillicrap et al., 2015), or epsilon-greedy or softmax action selection in discrete action spaces (Mnih et al., 2015). Unsurprisingly, random exploration often fails to efficiently explore policy space, since it is typically confined to a local region near the current policy and may evaluate many similar, redundant policies.

To address these shortcomings, exploration strategies based on the psychological concept of intrinsic motivation (Chentanez et al., 2005; Oudeyer and Kaplan, 2009) have used metrics such as novelty (Huang and Weng, 2002; Bellemare et al., 2016; Hart, 2009; Ecoeff et al., 2019; Burda et al., 2018), behavioral diversity (Lynch et al., 2019), uncertainty (Martinez-Cantin et al., 2007), and empowerment (Mohamed and Rezende, 2015) to diversify exploration in a more effective way. However, these methods are heuristic-based and may still lead to poor performance. In fact, many essentially aim to use intrinsic motivation to uniformly explore the state space, ignoring structure that may provide clues about the relevance of different parts of the state space to the problem at hand. By contrast, other approaches have taken advantage of structure specific to robotic manipulation by directing exploration to discover reusable object affordances (Montesano et al., 2008), or learning exploration strategies that exploit the distribution of problems the agent might face via metalearning (Xu et al., 2018b). Finally, as discussed in the next subsection, the difficulty of exploration in RL is sometimes overcome by leveraging demonstrations of good behavior, rather than learning from scratch.

6.4 Imitation Learning

In contrast to reinforcement learning, which learns from a robot’s experiences in the world (or a model of it), imitation learning (Schaal, 1999; Argall et al., 2009) aims to learn about tasks from demonstration trajectories. This can be thought of as a form of programming, but one in which the user simply shows the robot what to do instead of writing code to describe the desired behavior. Learning from demonstration data has been extensively studied in several different settings, because it can enable the robot to leverage the existing task expertise of (potentially non-expert) humans to (1) bypass time-consuming exploration that would be required in a reinforcement learning setting, (2) communicate user preferences for how a task ought to be done, and (3) describe concepts, such as a good tennis swing, that may be difficult to specify formally or programmatically. It is worth noting that imitation learning and reinforcement learning are not mutually exclusive; in fact, it is common for imitation learning to be followed by reinforcement learning for policy improvement (Kober and Peters, 2009; Taylor et al., 2011).

Demonstrations in imitation learning are typically represented as trajectories of states or state-action pairs. There are several mechanisms by which a robot may acquire such demonstration trajectories, including teleoperation, shadowing, kinesthetic teaching, mo-
tion capture, which are discussed in greater detail in the survey by Argall et. al (Argall et al., 2009). More recently, keyframe demonstrations (Akgun et al., 2012), virtual reality demonstrations (Zhang et al., 2018; Yan et al. 2018), and video demonstrations in the so-called learning from observation setting (Liu et al., 2018) have also become more commonly used.

**Behavioral Cloning:** The simplest way to use demonstration data to learn a motor skill is to use it as supervised training data to learn the robot’s policy. This is commonly called *behavioral cloning*. Recall that a deterministic policy \( \pi \) is a mapping from states to actions: \( \pi : S \rightarrow A \), whereas a stochastic policy maps state-action pairs to probabilities (which sum to 1 at each state when marginalizing over actions): \( \pi : S \times A \rightarrow \mathbb{R} \). The demonstration provides a set of state-action pairs \((s_i, a_i)\), that can be used as training data for a supervised learning algorithm to learn policy parameters that should, ideally, be able to reproduce the demonstrated behavior in novel scenarios.

Behavioral cloning is often used as a stand-alone learning method (Ross et al., 2011; Bagnell et al., 2007; Torabi et al., 2018; Pastor et al., 2009; Cederborg et al., 2010; Akgun et al., 2012; Duan et al., 2017; Hayes and Demiris, 1994; Schaal et al., 2005), as well as a way to provide a better starting point for reinforcement learning (Pastor et al., 2011a), though this additionally requires the specification of a reward function for RL to optimize. Other recent work has focused on expanding the purview of behavioral cloning by unifying imitation learning and planning via probabilistic inference (Rana et al., 2017), utilizing additional modalities such as haptic input (Kormushev et al., 2011), learning to recognize and recover from errors when imitating (Pastor et al., 2011b,a), and scaling behavioral cloning to complex, multi-step tasks such as IKEA furniture assembly (Niekum et al., 2015b).

**Reward Inference:** Rather than learning a policy directly from demonstration data, an alternative approach is to attempt to infer the underlying reward function that the demonstrator was trying to optimize. This approach aims to extract the intent of the motion, rather than the low-level details of the motion itself. This approach is typically called inverse reinforcement learning (IRL) (Ng et al., 2000), apprenticeship learning (Abbeel and Ng, 2004; Boularias et al., 2012), or inverse optimal control (Moylan and Anderson, 1973; Englert et al., 2017; Englert and Toussaint, 2018a). The inferred reward function can then be optimized via reinforcement learning to learn a policy for the task.

The IRL paradigm has several advantages. First, if the reward function is a function of the objects or features in the world and not the agent’s kinematics, then it can be naturally ported from human to robot (or between different robots) without encountering the correspondence problem. In addition, reward functions are often sparse, thereby providing a natural means of generalizing from a small amount of training data, even in very large state spaces. In addition, the human’s behavior may encode a great deal of background information about the task—for example, that an open can of soda should be kept upright when it is moved—that are easy to encode in the reward function but more complex to encode in a policy, and which can be reused in later contexts. Unfortunately, IRL also presents several difficulties. Most notably, the IRL problem is fundamentally ill-posed—infinitely many reward functions exist that result in the same optimal policy (Ng et al., 2000). Thus, the differentiation between many IRL algorithms lies in the metrics that they use to disambiguate or show preference for certain reward functions (Abbeel and Ng, 2004,
Maximum Entropy IRL (Ziebart, 2010) addresses the problems of demonstrator suboptimality and ill-posedness by leveraging a probabilistic framework and the principle of maximum entropy to disambiguate possible reward functions. Specifically, they develop an algorithm that assigns equal probability to all trajectories that would receive equal return under a given reward function and then use this distribution to take gradient steps toward reward functions that better match the feature counts of the demonstrations (Ziebart et al., 2008), while avoiding having any additional preferences other than those indicated by the data. Rather than generating a point estimate of a reward function, which forces an algorithm to face the ill-posedness of IRL head on, Bayesian IRL (Ramachandran and Amir, 2007) instead uses Markov Chain Monte Carlo to sample from the distribution of all possible reward functions, given the demonstrations. Finally, in the more restricted case of linearly-solvable MDPs, the IRL problem is well-posed, avoiding these problems (Dvijotham and Todorov, 2010).

All of the IRL algorithms mentioned so far rely on reward functions specified as a linear combination of features. While this does not restrict the expressivity of reward functions in practice (more complex features can always be provided), it burdens the designer of the system to ensure that features can be learned from in a linear manner. By contrast, Gaussian Process and nonparametric IRL (Levine et al., 2011; Englert et al., 2017) and various neural network-based methods (Ho and Ermon, 2016; Finn et al., 2016; Wulfmeier et al., 2015; Fu et al., 2017) aim to partly relieve this burden by searching for reward functions that are a nonlinear function of state features. However, such flexibility in representation requires careful regularization to avoid overfitting (Finn et al., 2016).

Many of the aforementioned methods have an MDP solver in the inner loop of the algorithm. Computational costs aside, this is especially problematic for robotics settings in which a model is not available and experience is expensive to collect. Some recent IRL methods that have been shown to work in real robotic domains sidestep this obstacle by alternating reward optimization and policy optimization steps (Finn et al., 2016) or framing IRL as a more direct policy search problem that performs feature matching (Doerr et al., 2015; Ho and Ermon, 2016). If available, ranked demonstrations can be used to get rid of the need for inference-time policy optimization or MDP solving entirely, by converting the IRL problem to a purely supervised problem; furthermore, this approach allows the robot to potentially outperform the demonstrator (Brown et al., 2019a, b). Alternately, active learning techniques have been used to reduce the computational complexity of IRL (Brown et al., 2018; Cui and Niekum, 2018; Lopes et al., 2009), as well as strategies that make non-I.I.D. assumptions about the informativeness of the demonstrator (Brown and Niekum, 2019; Kamalaruban et al., 2019). Finally, outside of an imitation learning framework, goals for the robot are sometimes specified via natural language commands that must be interpreted in the context of the scene (Tellex et al., 2011).

**Learning from Observation:** A relatively new area of inquiry aims to learn from demonstrations, even when no action labels are available and the state is not exactly known. For example, a robot may visually observe a human performing a task, but only have access to raw pixel data and not the true underlying state of the world, nor the actions that the human took. This problem is referred to Learning from Observation (LfO), and
several recent approaches have addressed problems including unsupervised human-robot correspondence learning (Sermanet et al., 2018), context translation (Liu et al., 2018), adversarial behavioral cloning (Torabi et al., 2018), and IRL from unsegmented multi-step video demonstrations (Goo and Niekum, 2019a; Yang et al., 2015). In an extreme version of the LfO setting, the agents is expected to infer an objective from single-frame goal-state images, rather than a full trajectory of observations (Zeng et al., 2018; Xie et al., 2018).

Corrective Interactions: Rather than learning from full demonstrations in batch, it is often advantageous to solicit (potentially partial) corrective demonstrations or other forms of feedback over time. For example, a human could intervene in a pouring task and adapt the angle of the cup and the robot’s hand mid-pour. This provides a natural mechanism to collect data in situations where it is most needed—for example, situations in which mistakes are being made, or where the robot is highly unsure of what to do. Some approaches actively ask users for additional (partial) demonstrations in areas of the state space in which confidence is low (Chernova and Veloso, 2009) or risk is high (Brown et al., 2018), while others rely on a human user to identify when a mistake has been made (Niekum et al., 2015b). Higher level information can also be used to make more robust corrections, such as grounded predicate-based annotations of corrections (Mueller et al., 2018) and action suggestions in a high-level finite state machine (Holtz et al., 2018). The robot can also actively solicit assistance when needed, for example, via natural language (Knepper et al., 2015). Finally, rather than using corrective demonstrations, the TAMER framework uses real-time numeric human feedback about the robot’s performance to correct and shape behavior (Knox et al., 2013).

6.5 Skill Transfer

Given the high sample complexity of learning in complex robotics tasks, skills learned in one task are often transferred to other tasks via a variety of mechanisms, thereby increasing the efficiency of learning.

Direct Skill Re-use: One of the simplest ways to transfer a skill policy is to directly re-use it in a new task related to the one it was learned on. Typically, some amount of adaptation is required to achieve good performance in the new task. One simple way to perform such refinement is to initialize a new skill with an existing skill’s policy parameters and adapt them for the new task via reinforcement learning (Pastor et al., 2009). However, a naive realization of this approach only works when the original task and the new task have identical state representations. When this is not the case, transfer can still occur via a subset of shared state features that retain semantics across tasks. This is sometimes called an agent space (Konidaris and Barto, 2007), since these are typically agent-centric, generic features (e.g., laser scanner readings), rather than problem-specific features, or a deictic representation (Platt et al., 2019). Since an agent space only covers some subset of the features in a problem, transfer occurs via initialization of a value function, rather than policy parameters.

More generally, it is often useful to find a state abstraction—a minimal subset of state features required to perform some skill. Abstractions facilitate transfer by explicitly ignoring parts of the state that are irrelevant for a particular skill, which could otherwise serve as distractors (e.g., irrelevant objects), as well as allowing transfer to state spaces of different
sizes (as also seen in agent spaces, which are a type of abstraction). In some state spaces, such as the visual domain, the state space is not factored in a manner that makes abstraction easy (e.g., the texture of an object is not a separate state feature, but distributed across many pixels). One popular way of forcing a deep neural network to abstract away complex variables such as texture and color is domain randomization (Tobin et al., 2017; Bousmalis et al., 2018), which is discussed in greater detail later in this section.

**Parameterized Skills:** In certain task families, only some aspects of the task context change, while all other task semantics remain the same or are irrelevant. For example, it may be desirable to transfer a policy that can hit one goal location on a dartboard, in order to hit a different goal location; similarly, transfer learning could acquire a policy for completing a pendulum swing-up task with different pendulum lengths and masses. In these restricted cases, specialized parameterized skills can be learned that facilitate transfer via mechanisms that modulate the policy based on the aspect of the task context parameter that is changing (Calinon, 2018; Englert and Toussaint. 2018b).

Dynamic Movement Primitives (Schaal, 2006; Pastor et al., 2009) use a simple spring-mass-damper system to smoothly adjust to new initial and goal locations. Another approach uses manifold learning to smoothly modulate policy parameters based on a variable task parameter (Da Silva et al., 2014). Contextual policy search uses a hierarchical, two-level policy for low-level control and generalization across contexts, respectively (Kupcsik et al., 2013). Universal value function approximators (Schaul et al., 2015; Andrychowicz et al., 2017) track a value function for all \(\langle \text{state, action, goal} \rangle\) triplets, rather than only \(\langle \text{state, action} \rangle\) pairs, allowing policy similarities across nearby goals to be leveraged explicitly.

**Metalearning:** Rather than re-using a skill directly for initialization, metalearning approaches aim to “learn to learn”—in other words, learn something about a distribution of tasks that allows for more efficient learning on any particular task from that distribution in the future. Thus, metalearning facilitates transfer within a task family by performing learning across samples from a task distribution at training time, rather than performing transfer sequentially and on-line as problems are encountered, as in the direct re-use case.

Model Agnostic Metalearning (MAML) (Finn et al., 2017) searches for a set of policy parameters that can be adapted quickly for particular tasks drawn from some distribution. Reptile (Nichol et al., 2018) simplifies MAML by replacing a complex optimization with an approximate approach that only requires standard stochastic gradient descent to be performed on each sampled task individually. Another related metalearning approach learns an attention-based strategy that allows the robot to imitate novel manipulation tasks, such as block-stacking, from a single demonstration (Duan et al., 2017).

Other forms of metalearning have focused on reward functions rather than policies. While many problems in robotic manipulation have simple sparse reward formulations (e.g., +1 when peg is in hole, -1 otherwise), potential-based shaping rewards (Ng et al., 1999) can be added to any reward function to encourage faster learning (across some distribution of problems) without changing the optimal policy of the MDPs. More generally, modified reward functions can help to overcome multiple forms of agent boundedness (Sorg et al., 2010b) and can be found via gradient descent (Sorg et al., 2010a), genetic programming (Niekum et al., 2010), or other evolutionary methods (Houthooft et al., 2018).

Some metalearning approaches have been developed to directly modify the representation of the policy itself, or other parts of the learning algorithm, rather than only the
settings of policy or reward parameters. This includes methods that evolve the structures of neural networks (Stanley et al., 2009), or that learn skill embeddings (Hausman et al., 2018), structured exploration strategies (Gupta et al., 2018), transfer feature relevances between families (Kroemer and Sukhatme, 2017), reusable network modules (Alet et al., 2018), or that co-learn a differentiable model and trajectory optimizer (Srinivas et al., 2018).

**Domain Adaptation:** In contrast to parameterized skills, some task families retain all of their high-level semantics across instances, differing only in lower-level details. In these cases, *domain adaptation* techniques are commonly used to bridge the so-called domain gap between two (or more) domains. For example, in the “sim2real” problem, when switching from a simulated task to a physical robot, the low-level statistics of the visual scene and physics may change, while the high-level steps and goals of the task stay fixed. Grounded action transformations address the physics domain gap by iteratively modifying agent-simulator interactions to better match real-world data, even when the simulator is a fixed black box (Hanna and Stone, 2017). Domain randomization is a data synthesis technique used to address visual domain gaps by randomizing environmental attributes such as texture, color, and dynamics in simulation, as to force the system to be invariant to statistical changes in the properties (Tobin et al., 2017; Bousmalis et al., 2018; Peng et al., 2018). A different type of visual shift can result from changes in viewpoint, often encountered in the setting of learning from unstructured videos. Approaches to this problem have included unsupervised learning of correspondences between first-person and third-person task demonstrations (Stadie et al., 2017), as well as imitation-from-observation approaches that work from multi-viewpoint data (Liu et al., 2018). Other approaches attempt to learn disentangled representations that lead naturally to robust policies that are tolerant of domain shift (Higgins et al., 2017).

In some tasks, the state and action space may change in a manner that requires an analogy to be made between the domains, as in the case of transferring a manipulation policy between two robot arms with different kinematic structures by finding a shared latent space between the two policy representations (Gupta et al., 2017). Time Contrastive Networks take a self-supervised metric-learning approach to a similar problem, learning representations of behaviors that are invariant to various visual factors, enabling robot imitation of human behaviors without an explicit correspondence (Sermanet et al., 2018). Other methods perform a more direct mapping between states and actions (Taylor et al., 2007), but assume that there exists a complete mapping, while the aforementioned latent space approach is able to discover partial mappings.

**Sequential Transfer and Curriculum Learning:** While the aforementioned transfer paradigms primarily consider instances of transfer individually, it is sometimes advantageous to view multiple instances transfer as a sequential learning problem. For example, Progressive Neural Networks (Rusu et al., 2016) use a neural network to learn an initial task, transferring this knowledge to subsequent tasks by freezing the learned weights, adding a new column of layers for the new task, and making lateral layer-wise connections to the previous task’s neurons. Sequential transfer learning is also a useful paradigm for mastering tasks that are too difficult to learn from scratch. Rather than attack the final task directly, *curriculum learning* presents a sequence of tasks of increasing difficulty to the agent, which provides a smoother gradient for learning and can make positive rewards significantly less sparse in an RL setting. Such methods often use curricula provided by human designers.
A Review of Robot Learning for Manipulation

(Bengio et al., 2009; Sanger, 1994; Pinto and Gupta, 2016), though several recent methods seek to automatically generate a curriculum (Narvekar et al., 2016; Svetlik et al., 2017; Florensa et al., 2017).

Scheduled Auxiliary Control (Riedmiller et al., 2018) learns to sequentially choose from a set of pre-defined auxiliary tasks in a manner that encourages efficient exploration and learning of the target task. Guided Policy Search (Levine et al., 2015) first learns simple time-dependent controllers for individual task instances, which are then used to generate data to train a more complex neural network. Universal Value Function Approximators (UVFAs) (Schaul et al., 2015) can learn and transfer knowledge more efficiently in curriculum-like settings by estimating the value of a state and action conditioned upon the current goal, rather than for a single fixed goal. Hindsight Experience Replay (HER) provides additional goal settings for UVFAs by simply executing actions, and in hindsight imagining that every state that the agent ends up in was actually a goal (Andrychowicz et al., 2017).

Reverse curriculum learning (Florensa et al., 2017) begins at the goal and works backwards, trying to learn to get to the goal successfully from more and more distant states. This stands in contrast to standard curriculum learning that typically starts in some distant state, slowly moving the goal of the subtasks closer and closer to the true goal. This idea is also related to that of skill chaining (Konidaris and Barto, 2009b; Konidaris et al., 2012), which will be discussed in further detail in Section 8.1.

6.6 Safety and Performance Guarantees

Whether a policy is learned directly for a specific task or transferred from a previous task, an important but understudied question is “How well will the policy perform across the distribution of situations that it will face?”. This is an especially important question in robotic manipulation, in which many future applications will require behaviors that are safe and correct with high confidence: robots that operate alongside humans in homes and workplaces must not cause injuries, destroy property, or damage themselves; safety-critical tasks such as surgery (Van Den Berg et al., 2010) and nuclear waste disposal (Kim et al., 2002) must be completed with a high degree of reliability; robots that work with populations that rely on them, such as the disabled or elderly (Goil et al., 2013), must be dependable. While the term “safety” has taken on many different meanings in the literature, in this section, we take a broad view of safe learning to include all methods that aim to bound a performance metric with high confidence.

Performance Metrics: Most performance metrics for robotic manipulation tasks can be represented as cumulative reward under some reward function—for example, the preferences of completing a task accurately, finishing within a certain amount of time, or never (or with low probability) visiting certain states can all be described with particular reward functions and acceptable thresholds on cumulative reward, or return. Thus, for simplicity, we will discuss the problem of bounding performance as being equivalent to that of lower-bounding some function of return. However, it is worth noting that certain preferences, such as some temporal logic statements, cannot be expressed as a reward function without augmenting the state space; some cannot be expressed by any reward function at all (Littman et al., 2017).
The principle axis of variation regarding performance metrics is whether the *expected* return of a policy is being bounded, or whether a risk-aware function of return is used—for example, a robust worst-case analysis (Ghavamzadeh et al., 2016), a PAC-style bound (Strehl et al., 2006), or bounded value-at-risk (Chow et al., 2015; Brown et al., 2020). Risk-aware metrics are generally more appropriate for safety-critical tasks in which the robot could cause damage or injury in a single trial that goes poorly, whereas expected performance is often used in scenarios in which long-term performance (e.g., percentage of correctly manufactured items) matters more than the outcome of any single trial.

Finally, one important distinction to make is *when* a policy must obey performance bounds. Some robotics tasks, such as learning to manipulate a fragile object, demand performance guarantees during the exploration and learning phase, rather than only after deployment of the final policy. For example, in one recent work, demonstrations were used to constrain exploration in safety-critical manipulation tasks (Thananjeyan et al., 2019). By contrast, other tasks may lend themselves to safe practice, such that only the final policy needs to be accompanied by guarantees. For example, a legged robot may learn a walking gait while attached to a support rig so that it cannot catastrophically fall while learning.

**Classes of Guarantees and Bounding Methods:** Given that standard policy learning in robotics is a challenging open problem, in part due to limited real-world data collection abilities, it is not surprising that safe learning is even more difficult. Safe learning methods are typically significantly more data-hungry and/or require more accurate models than standard learning methods, as they need to provide guarantees about the quality of the policy. For this reason, very few safe learning methods (e.g., from the reinforcement learning community) have been used in robotics applications. This is a significant gap in the current literature and an opportunity for future work that can provide strong performance guarantees in low-data and poor-model robotics settings.

Safe learning methods fall roughly into two categories: Exact methods and probabilistic high-confidence bounds. Formal verification-based approaches are exact methods that use a known model, along with a safety specification (e.g., a finite-state automata) to guarantee (or bound the probability, in the stochastic case) that following a policy will not lead to an unsafe outcome (Fu and Topcu, 2014; Fainekos et al., 2009; Chinchali et al., 2012; Chen et al., 2013; Alshiekh et al., 2018). However, some research has focused on control synthesis that obeys linear temporal logic constraints without access to a model (Sadigh et al., 2014). By contrast, probabilistically safe approaches aim to provide lower-bounds on expected return, rather than utilizing logical specifications of safety (Thomas et al., 2015b; Jiang and Li, 2016; Hanna et al., 2017; Thomas and Brunskill, 2016). While probabilistic methods appear to be promising for robotics applications (since they do not require a model), to the best of our knowledge, they have not been used in real robotics problems, potentially due to their high sample complexity.

Thus far, we have only considered performance guarantees in a standard MDP setting, in which either (1) the reward function is known, or (2) samples from the reward function are available. However, this description of the problem does not cover a common scenario that occurs when learning from demonstrations—having access to states and actions, but not rewards. In the inverse reinforcement learning setting, several approaches have examined how to bound the policy loss between the the robot’s policy (commonly a policy learning via RL under the inferred reward function) and the optimal policy under the ground-truth
A Review of Robot Learning for Manipulation

reward function of the demonstrator, even though it is unknown (Abbeel and Ng, 2004; Brown and Niekum, 2018; Brown et al., 2020).

7. Characterizing Skills by Preconditions and Effects

Executing a manipulation skills alters the state of the robot’s environment; if the robot is to use its skills to achieve a specific goal, it requires a model of those outcomes. This model is termed a *postcondition*, and describes the regions in the state space that the robot will find itself in after skill execution. The robot must also model the circumstances under which the skill can be executed—termed its *preconditions*. Knowledge of these two models for each skill can be used to determine whether, for example, a sequence of skills can be executed from a particular state (Konidaris et al., 2018). Pre- and postconditions are used to sequence skills (or actions) for task planning. The planner searches through sequences of actions with the constraint that the postconditions of each skill must fulfill the preconditions of the next skill.

7.1 Pre- and Postconditions as Propositions and Predicates

The pre- and postconditions of manipulation skills are typically represented abstractly as either *propositions* or *predicates* (Krüger et al., 2011; Kulick et al., 2013; Konidaris et al., 2018; Beetz et al., 2010; Shapiro and Ismail, 2003; Ugur and Piater, 2015) that are either true or false in any specific state. For example, the robot may represent the outcome of a particular navigation skill using the boolean proposition $\text{AtTableB}$. It could also use a predicate representation, $\text{At(TableB)}$, which supports more compact and efficient representations, and allows for easier generalization across objects. We therefore use predicate-based symbols for the remainder of this section although most of the explanation also applies to proposition symbols.

The *grounding* or anchoring of a predicate refers to the mapping between the (often continuous) low-level state and context and the predicate’s truth value; it defines the meaning of the symbol (Konidaris et al., 2018; Coradeschi et al., 2013). To avoid confusion, we will not use the term grounding for assigning objects to predicates as is done in the planning literature (Russell and Norvig, 2003).

**Classifier Representation:** The grounding of a predicate can be modelled as a binary classifier (Konidaris et al., 2018; Kroemer and Sukhatme, 2016; Konidaris et al., 2015). The classifier represents the mapping from the state and context to either true or false. If a predicate is defined for a subset of objects, then the state and context features of those objects are used as the input to the predicate classifier (Kroemer and Sukhatme, 2016). For example, the predicate $\text{Grasped(KnifeA, RHand)}$ would consider the state and context of the knife and hand to determine if the knife was being grasped by the hand. The robot may learn a probabilistic classifier to determine the likelihood that the predicate is true given the current state and context. In this manner, the robot may handle situations where the predicate is only sometimes valid, or when the classifier has been learned and the robot is uncertain of its true value in states it has not encountered before (Konidaris et al., 2018).

**Distribution Representation:** Alternatively, predicates can be modelled as probability distributions over the state space (Konidaris et al., 2018; Detry et al., 2011). The
distributions are defined in the state and context space. For a non-probabilistic approach, the distribution defines the set of states and contexts where the distribution is true, and the predicate is false otherwise. For a probabilistic approach, the distribution defines a probability density over the states and contexts given that the predicate is true (Konidaris et al., 2018; Detry et al., 2011). The distribution may be defined only for the objects assigned to the predicate. This distribution is useful for sampling states and contexts in which the predicate is true.

Modularity and Transfer: One precondition proposition and one postcondition proposition for each skill are sufficient for skill sequencing. These predicates can be monolithic representations that define the sets of states and contexts for that specific skill. However, by decomposing the conditions into modular predicates, the robot can share knowledge between different skills and tasks. These predicates can often be defined for subsets of objects, e.g., $\text{Full}(\text{mug, water})$ and $\text{Grasped}(\text{mug, hand})$. The predicates often define labels for individual objects, e.g., $\text{Container}(\text{obj1})$, and the relationships between the objects, e.g., $\text{Above}(\text{obj1, obj2})$. Modular predicates will often capture the contact-based manipulation modes between pairs of objects, e.g., $\text{Grasped}(\text{Obj1,Hand})$ or $\text{On}(\text{Obj2,Obj3})$. Previous works have explored methods for learning specific predicates or discovering suitable sets of predicates (Kulick et al., 2013; Montesano et al., 2008; Konidaris et al., 2018; Hjelm et al., 2014).

Learning the preconditions and postconditions is generally easier when reusing predicates from previous tasks rather than learning from scratch. However, some discrepancies may exist between tasks and thus require additional learning (Ugur and Piater, 2015). For example, a robot may require a specific type of grasp for performing a task. The predicate $\text{Grasped}(\text{obj3,hand})$ is thus not sufficient. The robot should instead learn a general predicate for transferring knowledge between tasks, and a task-specific predicate for incorporating additional constraints (Detry et al., 2017; ten Pas et al., 2017; Gualtieri and Platt, 2018; Lu and Hermans, 2019; Bohg et al., 2012; Hjelm et al., 2014). The latter predicate is generally easier to learn given the former predicate. For example, a robot may learn to identify stable grasps for a wide range of objects, and subsequently learn to identify a subset of grasps for specific objects or to grasp handles of cooking utensils without relearning to grasp from scratch (Detry et al., 2017; Gualtieri and Platt, 2018; ten Pas et al., 2017).

7.2 Learning Pre- and Postcondition Groundings

The robot can train the pre- and postcondition predicate classifiers using samples of states and contexts where the conditions were known to be true or false. The ground truth labels could be provided by a human supervisor, although this approach would limit the autonomy of the robot and may require substantial expertise from the user. In some cases a human can provide data for desired post- and preconditions—what the skill should achieve, from which conditions—but the actual conditions will ultimately depend on the robot’s capabilities.

Instead of relying on manual labeling, the robot can learn the condition labels from experience. The labels for the preconditions can be learned given a fixed postcondition—all states from which a skill execution leads to a state satisfying the postcondition are positive examples, and all other states are negative examples. The robot should use a probabilistic classifier to capture the stochasticity of the transitions. The robot can thus
obtain the precondition labels by executing the skill from different states and observing the resulting outcomes.

The postconditions are less trivial to define. A human may predefine a desired postcondition, but to achieve autonomy the robot must discover different postconditions on its own. Distinct postconditions can be learned by clustering the outcomes of the skill from different initial states and contexts (Dogar et al., 2007; Ugur et al., 2009; Ugur and Piater, 2015). Manipulation skills often have distinct effects, e.g., a box remained upright or toppled over, which can be extracted through clustering or by detecting salient events. Each distinct postcondition cluster will then be associated with its own preconditions. This approach is often employed in developmental learning to discover specific useful skills from more general skill policies, e.g., grasping and pushing from reaching (Oztop et al., 2004; Juett and Kuipers, 2018).

A more goal-oriented approach to specifying postconditions is to construct skills where the postcondition is either the goal of the task or another skill’s precondition (Konidaris and Barto, 2009b; Konidaris et al., 2011a). This approach directly learns pre- and postconditions for constructing skills that can be sequentially executed, and may avoid learning conditions that are irrelevant to the robot’s task set, but it introduces dependencies between the robot’s skills.

Finally, pre- and postcondition predicates can be grounded in sensory data with the assistance of natural language, for example, as part of an interactive dialogue between a human and a robot (Thomason et al., 2016). Conversely, knowledge of pre- and postconditions of skills can help to ground natural language commands to those skills, especially when the language description is incomplete (Misra et al., 2014). For example, the command “Stir the soup” may imply picking up a spoon first, which could be determined via the precondition to a stirring skill.

7.3 Skill Monitoring and Outcome Detection

Most skills will have distinct pre- and postconditions with some of the predicates changing as a result of the skill execution, e.g., executing a grasping skill on a book should result in Grasped(Book,RHand)=True once the book has been grasped (Dang and Allen, 2012; Garrett et al., 2018). The predicates may also change due to errors in the skill execution. For example, when executing a placing skill, the predicate values On(Book, Table)=False or InCollision(Book, Obstacle)=True would correspond to errors. To perform manipulation tasks robustly, the robot must monitor its skill executions and determine if and when it has achieved the intended outcome or whether an error has occurred.

Learning Goal and Error Classifiers: Goals and errors in skill executions can be modeled as distinct predicates with values that are switched for the postconditions. Detecting goals and errors in such executions can thus be modeled as a classification problem (Bekiroglu et al., 2013; Rodriguez et al., 2010). Rather than using only the current state, the robot can incorporate action and sensor information from the entire skill execution (Bekiroglu et al., 2013; Madry et al., 2014; Rodriguez et al., 2010), although it is often better to stop a skill early when an error occurs (Pastor et al., 2012; Su et al., 2016; Sukhoy et al., 2012). Transient events, such as vibrations from mode transitions or incipient slip can then be used to better detect the predicate switches (Park et al., 2016; Veiga et al., 2015; Su et al., 2016).
The robot can use a variety of sensor modalities, including vision and audio, to detect the predicate switches, e.g., a robot can use tactile sensing to determine if a grasp attempt succeeded \texttt{Grasped(obj)=True} (Calandra et al., 2018; Dang and Allen, 2012; Madry et al., 2014; Dang and Allen, 2013). A robot can also learn optimal locations for placing a camera, or other sensor, to reliably verify if a desired postcondition has been fulfilled (Saran et al., 2017).

**Detecting Deviations from Nominal Sensory Values:** Stereotypical executions of a manipulation skill will usually result in similar sensations during the execution. Larger deviations from these nominal values often correspond to errors (Yamaguchi and Atkeson, 2016b). Hidden Markov models can be used to track the successful progress of a skill’s execution based on sensory signals (Park et al., 2016; Lello et al., 2013; Hovland and MacCarragher, 1998; Bekiroglu et al., 2010). The robot may also learn the nominal sensory signals as a regression problem (Pastor et al., 2011a; Kappler et al., 2015; Sukhoy et al., 2012). Significant deviations from the expected sensory values would then trigger the stopping of the current skill. An error may also trigger a corresponding recovery action (Dang and Allen, 2013; Veiga et al., 2018; Yamaguchi and Atkeson, 2017). Unlike the outcome classifiers from the previous section, these models are trained using only data from successful trials.

**Verifying Predicates:** The pre- and postconditions of skills often change the values of predicates corresponding to modes and constraints between objects. For example, a placing skill can make \texttt{On(DishC,DishB)=True} and an unlocking skill can make \texttt{Locked(DoorA)=False}. The robot can verify these switches in the predicate values using interactive perception. In some cases the predicate can be verified by directly performing the next skill in the sequence, e.g., attempting to lift an object after it has been grasped (Pinto and Gupta, 2016; Kalashnikov et al., 2018). For contact constraints, the robot can often perform small perturbations to verify the predicate’s final value (Debus et al., 2004; Wang and Kroemer, 2019). In both examples, the robot can obtain a better estimate of the predicate by performing the additional skill and observing the effect. However, the additional skills can sometimes also change the predicates, and hence some care needs to be taken when verifying predicates.

### 7.4 Predicates and Skill Synthesis

Skills will often have additional high-level arguments that define how to execute the skill (Ugur et al., 2011). For example, a grasping skill may take in a grasping pose, or a scrubbing skill may take in a desired force (Bohg et al., 2014). In this manner, a higher-level policy may adapt the skill to its specific needs.

**Representing and Synthesizing Skill Parameters:** Similar to the predicate representations, these policy parameter arguments can be modeled as additional input features for the precondition classifier or as distributions over valid and invalid argument values (Detry et al., 2011; Jiang et al., 2012). Many of these arguments can even be thought of as virtual or desired object states (Jiang et al., 2012). For example, one could sample potential hand positions for grasping objects and model each one as a symbol for specifying the grasping action (Garrett et al., 2018; Bohg et al., 2014). To obtain valid grasp frames, the robot can learn a classifier to determine if sampled grasp frames will lead to successful
grasps (Saxena et al., 2008; Nguyen and Kemp, 2014; Herzog et al., 2014; Pinto and Gupta, 2016), or sample from a learned probability density over successful grasps or cached grasps from similar objects (Detry et al., 2011; Brook et al., 2011; Goldfeder et al., 2009; Choi et al., 2018). The process of learning the pre- and post-conditions with additional policy arguments is thus similar to learning the preconditions without the arguments.

Once the robot has learned the pre- and postconditions over the arguments, it can use these to select skill parameters for new situations. This process usually involves sampling different parameter values and evaluating them in terms of the pre- and postconditions. The robot could use sampling and optimization methods to select argument values with high likelihoods of successful skill executions. As these parameters are often lower dimensional than full skills, active learning and multi-armed bandit methods can be used to select suitable argument values (Montesano and Lopes, 2012; Mahler et al., 2016; Kroemer et al., 2010). When the positive distribution is learned directly, the robot can sample from the distribution and then evaluate if it is consistent with other predicates, e.g., not in collision (Ciocarlie et al., 2014). The arguments are usually selected in order to achieve certain predicates in the postconditions. One can therefore think of the argument selection process as predicate synthesis (Ames et al., 2018).

**Preconditions and Affordances:** Affordances are an important concept in manipulation and a considerable amount of research has explored learning affordances for robots (Sahin et al., 2007; Min et al., 2016; Jamone et al., 2018). The affordances of an object are the actions or manipulations that the object affords an agent (Gibson, 2014). An object that can be used to perform an action is thus said to afford that action to the agent. For example, a ball affords bouncing, grasping, and rolling. Affordances are thus closely related to the preconditions of skills. As affordances connect objects to skills (Montesano et al., 2008), affordance representations often include skill arguments that define how the skill would be executed (Ugur et al., 2011; Kroemer et al., 2012).

The exact usage of the term affordances tends to vary across research papers (Jamone et al., 2018). In most cases, affordances can be seen as a form of partial preconditions. The affordances often correspond to specific predicates that the robot learns in the same manner as the preconditions. The other components of the precondition are usually constant or at least all valid, such that the success or failure of the skill only depends on the component being learned. As partial preconditions, some affordances are more specific than others, e.g., balls can be rolled versus balls on a plane within reach of the robot can be rolled. Ultimately, the full preconditions need to be fulfilled to perform the skill. However, the partial preconditions of affordances provide modularity and can help the robot to search for suitable objects for performing tasks.

### 8. Learning Compositional and Hierarchical Task Structures

In the previous sections, we have focused on learning models of individual objects, or on learning to perform or characterize individual motor skills. However, manipulation tasks often have a substantial modular structure that can be exploited to improve performance across a family of tasks. Therefore, some research has attempted to decompose the solution to a manipulation task into component skills. Decomposing tasks this way has several advantages. Individual component skills can be learned more efficiently because each skill is
shorter-horizon, resulting in a substantially easier learning problem and aiding exploration. Each skill can use its own internal skill-specific abstraction (Diuk et al., 2009; Konidaris and Barto, 2009a; van Seijen et al., 2013; Cobo et al., 2014; Jiang et al., 2015) that allows it to focus on only relevant objects and state features, decomposing a problem that may be high-dimensional if treated monolithically into one that is a sequence of low-dimensional subtasks. A skill’s recurrence in different settings results in more opportunities to obtain relevant data, often offering the opportunity to generalize; conversely, reusing skills in multiple problem settings can avoid the need to relearn elements of the problem from scratch each time, resulting in faster per-task learning. Finally, these component skills create a hierarchical structure that offers the opportunity to solve manipulation tasks using higher-level states and actions—resulting in an easier learning problem—than those in which the task was originally defined.

8.1 The Form of a Motor Skill

The core of hierarchical structure in manipulation learning tasks is identifying the component skills from which a solution can likely be assembled.

Recall that skills are often modeled as options, each of which is described by a tuple $o = (I_o, \beta_o, \pi_o)$, where:

- $I_o : S \rightarrow \{0, 1\}$ is the initiation set, which corresponds to a precondition as discussed in Section 7.

- $\beta_o : S \rightarrow [0, 1]$, the termination condition, describes the probability that option $o$ ceases execution upon reaching state $s$. This corresponds to a goal as discussed in Section 7, but is distinct from an effect; the goal is the set of states in which the skill could (or perhaps should) terminate, whereas the effect describes where it actually terminates (typically either a subset of the goal or a distribution over states in the goal).

- $\pi_o$ is the option policy.

In many cases, option policies are defined indirectly using a reward function $R_o$, often consisting of a completion reward for reaching $\beta_o$ plus a background cost function. $\pi_o$ can then be obtained using any reinforcement learning algorithm, by treating $\beta_o$ as an absorbing goal set. The core question for finding component skills when solving a manipulation problem is therefore to define the relevant termination goal $\beta_o$—i.e., identify the target goal—from which $R_o$ can be constructed.

The robot may construct a skill library, consisting of a collection of multiple skills that can be frequently reused across tasks. This requires extracting a collection of skills, either from demonstrations, or from behaviors generated autonomously by the robot itself. The key question here is how to identify the skills, which is a difficult, and somewhat under-specified, challenge. There are two dominant approaches in the literature: segmenting task solution trajectories into individual component skills, or directly including skill specification as part of the overall problem when learning to solve tasks.
8.2 Segmenting Trajectories into Component Skills

One approach to identifying a skill library is to obtain solution trajectories for a collection of tasks, and segment those trajectories into a collection of skills that retroactively decompose the input trajectories. This is commonly done using demonstration trajectories, though it could also be performed on trajectories generated autonomously by the robot, typically after learning (Hart, 2008; Konidaris et al., 2011a; Riano and McGinnity, 2012). However they are generated, the resulting trajectories must be segmented into component skills. The literature contains a large array of methods for performing the segmentation, which we group into two broad categories.

**Segmentation Based on Skill Similarity:** The most direct approach is to segment demonstration trajectories into repeated subtasks, each of which may occur in many different contexts (Dang and Allen, 2010; Meier et al., 2011). Identifying such repeated tasks both reduces the size of the skill library (which in turn reduces the complexity of planning or learning which skill to use when) and maximizes the data available to learn each skill (Lioutikov et al., 2015). This requires a measure of skill similarity that expresses a distance metric between two candidate skill segments, or more directly models the probability with which they were generated by the same skill. Therefore, several approaches have used a measure of skill similarity to segment demonstration trajectories, often based on a variant of a Hidden Markov model, where the demonstrated behavior is modeled as the result of the execution of a sequence of latent skills; segmentation in this case amounts to inferring the most likely sequence of skills.

The most direct approach (Jenkins and Matarić, 2004; Chiappa and Peters, 2010; Grollman and Jenkins, 2010; Niekum et al., 2015b; Meier et al., 2011; Daniel et al., 2016b) is policy similarity, which measures skill similarity by fitting the data to a parameterized policy class and measuring distance in parameter space. Alternative approaches to policy similarity fit models to the underlying value function (Konidaris et al., 2012) or the unobserved reward function that the skill behavior is implicitly maximizing (Ranchod et al., 2015; Krishnan et al., 2019; Michini and How, 2012; Choi and Kim, 2012; Babes et al., 2011). Some approaches segment demonstration trajectories and then merge similar skills in a separate post-processing step (Konidaris et al., 2012), while the most principled probabilistic approaches infer shared skills across a collection of trajectories as part of the segmentation process (Niekum et al., 2015b; Ranchod et al., 2015). Rather than using a parametric model, a latent space can be discovered that efficiently encodes skills, which are either learned from experience in a reinforcement learning context, or via a latent-space segmentation of trajectories in an imitation learning setting (Hausman et al., 2017, 2018). Other recent approaches have relied on observations only, learning to segment videos of multi-step tasks into composable primitives (Goo and Niekum, 2019b; Yu et al., 2018; Huang et al., 2019; Xu et al., 2018a).

A less direct approach is to measure skill-similarity based on pre- and post-condition similarity, where the skill policy or trajectory itself is not used for segmentation. Instead, trajectory segments that achieve the same goal can be clustered together, while those that do so from very different initial conditions could be split (Kroemer et al., 2014). Thus, reaching and pushing motions are different skills due to their distinct pre- and post- conditions, i.e., moving or not moving an object, even if the skill policies are similar or the same. This
approach is often used in developmental learning approaches to discover skills (Weng et al., 2001); the robot executes a skill policy in a variety of scenarios and then clusters together the post- and pre-conditions to create distinct skills (Xie et al., 2018; Niekum and Barto, 2011; Ugur et al., 2011; Ugur and Piater, 2015; Ugur and Piater, 2015).

**Segmentation Based on Specific Events**: Several approaches use pre-designated events to indicate skill boundaries. These can range from hand-specified task-specific events to more generally applicable principles.

One common approach is to segment by **salient sensory events** defined by haptic, tactile, audio, or visual cues (Juett and Kuipers, 2018; Su et al., 2016; Fitzpatrick et al., 2006; Aksoy et al., 2011). For example, insertion skills are easier to monitor if they result in a distinct *click* upon successful completion. Similarly, we can determine if a light switch was properly pushed if the light goes on or the button cannot be pushed any further. Human-made objects are often designed to provide salient event feedback to reduce errors while using the objects. Such skills have the advantage that their termination conditions are easy to detect and monitor.

Another important class of pre-defined segmentation events is **transitioning between modes** (Baisero et al., 2015; Kroemer et al., 2015). Switching between modes allows the robot to switch between the ability to interact with different objects. Hence, the robot must first transition to a suitable mode to manipulate an object. Skills for transitioning to specific modes allow the robot to decouple accessing a mode and using the mode to perform a manipulation task. Grasping, lifting, placing, and releasing are all examples of skills used to switch between modes in pick-and-place tasks. Mode transitions can also be verified by applying additional actions to determine if the skill execution was successful. However, multiple skills within a mode, e.g., squeezing, shaking, and tilting a held object, cannot be detected using only this type of decomposition.

### 8.3 Discovering Skills While Solving Tasks

An alternative approach is to discover component motor skills **during** the process of learning to solve one or more manipulation tasks. This has two crucial advantages over solution trajectory segmentation. First, learned skills could aid in solving the tasks to begin with; many complex manipulation tasks cannot be solved directly without decomposing them into simpler tasks, so it will be infeasible to require that a complete solution precede skill discovery. Second, imposing hierarchical structure **during** learning can inject bias that results in much more compact skill libraries. For example, there may be multiple ways to turn a switch; there is no reason to expect that a robot that independently learns to solve several tasks involving switches will find a similar policy each time. However, if the learned policy for turning a switch is identified and retained in the first task, then it will likely be reused in the second task.

Broadly speaking, there are substantially fewer successful approaches that learn skills while solving tasks than there are approaches that retroactively segment, in part because the problem is fundamentally harder than segmentation. For example, skill similarity approaches are difficult to apply here. However, some researchers have applied methods based on specific salient events (Hart, 2008) to trigger skill creation. These methods could be combined with skill chaining (Konidaris and Barto, 2009b; Konidaris et al., 2011a)—an ap-
A Review of Robot Learning for Manipulation

proach where skills are constructed to either reach a salient event or to reach another skill’s preconditions—to learn the motor skills online. An alternative is to structure the policy class with which the robot learns to include hierarchical components which can be extracted after learning and reused. This approach naturally fits recent research using deep neural networks for policy learning (Nachum et al., 2018; Vezhnevets et al., 2017; Levy et al., 2019), which are sometimes structured to be successively executable in a manner similar to skill chaining (Kumar et al., 2018), but have been used in other more compact representations (Daniel et al., 2013, 2016a). Another approach aims to incrementally discover hierarchical skills that progressively learn to control factorized aspects of the environment, such as objects (Chuck et al., 2020). These new results are exciting but have only just begun to scratch the surface of how component motor skills can be learned during the robot’s task learning process.

8.4 Learning Decision-Making Abstractions

A collection of abstract motor skills provides procedural abstraction to the robot; it can abstract over the low-level details of what it must do to solve the task. However, these motor skills also provide an opportunity for abstracting over the input the robot uses to decide which skill to execute. Learning such abstract structures have two potential advantages: First, the new abstract input may make learning for new tasks easier (or even unnecessary by enabling generalized abstract policies that work across tasks) and support abstract, task-level planning. Second, the resulting abstract representations may be much easier for a non-expert to edit, update, and interpret.

Broadly speaking, the literature contains two types of learned abstractions for decision-making. The first type learns an abstract policy directly, while the second learns an abstract state space that serves as the basis for either planning or much faster learning in new tasks.

Learning Abstract Policy Representations: Here the goal is to learn a generalized abstract policy that solves a class of problems. Often this policy encodes a mapping from abstract quantities to learned motor skills. For example, the policy for opening a door might include first checking to see if it is locked and if so, executing the unlock skill; then grasping the handle, turning it, and opening the door. Here the motor skills subsume the low-level variations in task execution (e.g., turning differently shaped doorknobs) while the abstract quantities subsume differences in task logic (e.g., checking to see if the door is locked). Approaches here differ primarily by the representation of the policy itself, ranging from finite-state machines (Niekum et al., 2015b) to associative skill memories (Pastor et al., 2012) to hierarchical task networks (Hayes and Scassellati, 2016) to context-free grammars (Lioutikov et al., 2018). In some cases these result in a natural and intuitive form of incremental policy repair and generalization (Niekum et al., 2015b; Hayes and Scassellati, 2016).

Learning Abstract State Spaces: Alternatively, the robot could learn abstract representations of state, which when combined with abstract actions result in a new, but hopefully much simpler, and typically discrete, MDP. The robot can then use that simpler MDP to either learn faster, or to learn a task model and then plan. Several approaches have been applied here, for example using the status of the skills available to the robot as a state space (Brock et al., 2005; Hart, 2008), finding representations that minimize planning
loss (Kulick et al., 2013; Jetchev et al., 2013; Ugur and Piater, 2015; Ugur and Piater, 2015), using qualitative state abstractions (Mugan and Kuipers 2012) and constructing compact MDPs that are provably sound and complete for task-level planning (Konidaris, 2016; Konidaris et al., 2018; Ames et al., 2018). These approaches offer a natural means of abstracting away the low-level detail common to learning for manipulation tasks, and exploiting the structure common to task families to find minimal compact descriptions that support maximally efficient learning.

Taken together, hierarchical and compositional approaches have great promise for exploiting the structure in manipulation learning tasks, to reduce sample complexity and achieve generality, and have only begun to be carefully explored. This is a challenging area full of important questions, and where several breakthroughs still remain to be made.

9. Conclusion

This paper has presented an overview of key manipulation challenges and the types of robot learning algorithms that have been developed to address these challenges. We explained how robots can represent objects and learn features hierarchically, and how these features and object properties can be estimated using passive and interactive perception. We have discussed how the effects of manipulations can themselves be captured by learning continuous, discrete, and hybrid transition models. Different data collection strategies and model types determine how quickly the robot can learn the transition models and how well the models generalize to new scenarios.

Skill policies can often be learned quickly from human demonstrations using behavioural cloning or apprenticeship learning, while mastery of manipulation skills often requires additional experience and can be achieved using model-based or model-free reinforcement learning. Given a skill, the robot can learn its preconditions and effects to capture its utility in different scenarios. Learning to monitor manipulation skills by detecting errors and goals imbues them with an additional level of robustness for working in unstructured environments.

Finally, we have seen how a robot can exploit the modular nature of manipulation tasks, such that each learned skill can be incorporated into a larger hierarchy to perform more advanced tasks, promoting reusability and robustness of skills across different tasks.

Given the multi-faceted nature of manipulation, researchers have found great utility in being able to draw from methods that span the breadth of machine learning. However, despite access to these excellent generic machine learning methods, the challenges of learning robust and versatile manipulation skills are still far from being resolved. Some—but by no means all—of these pressing challenges are:

- Integrating learning into complete control systems
- Using learned components as they are being learned (in situ learning)
- Safe learning, and learning with guarantees
- Exploiting and integrating multiple sensory modalities, including human cues
- Better exploration strategies, possibly based on explicit hypotheses or causal reasoning
• Exploiting common-sense physical knowledge
• Better algorithms for transfer across substantially different families of tasks
• Drastically improving the sample complexity of policy learning algorithms, while avoiding having to empirically tune hyper-parameters

As the community tackles these and other challenges, we expect that the core themes that have emerged repeatedly in manipulation learning research—modularity and hierarchy, generalization across objects, and the need for autonomous discovery—will continue playing a key role in designing effective solutions.

Acknowledgments

The authors would like to thank Animesh Garg for all of his help and insightful discussions, and Ben Abbatematteo, Daniel Brown, Caleb Chuck, Matt Corsaro, Yuchen Cui, Eren Derman, Wonjoon Goo, Prasoon Goyal, Reymando Gutierrez, Ajinkya Jain, Timothy Lee, Qiao (Jacky) Liang, Rudolf Liutikov, Matt Mason, Akanksha Saran, Mohit Sharma, and Garrett Warnell for their feedback and comments on the draft.

This work was supported in part by the National Science Foundation (IIS-1724157, IIS-1638107, IIS-1617639, IIS-1749204, IIS-1717569), the Office of Naval Research (N00014-18-2243, N00014-18-1-2775) and the Sony Corporation.

References


A Review of Robot Learning for Manipulation


59


Alexandros Paraschos, Christian Daniel, Jan R Peters, and Gerhard Neumann. Probabilistic
movement primitives. In Advances in Neural Information Processing Systems 26, pages

Daehyung Park, Zackory Erickson, Tapomayukh Bhattacharjee, and Charles C. Kemp.
Multimodal execution monitoring for anomaly detection during robot manipulation. In
Proceedings of the 2016 IEEE International Conference on Robotics and Automation,

Peter Pastor, Heiko Hoffmann, Tamim Asfour, and Stefan Schaal. Learning and general-
ization of motor skills by learning from demonstration. In Proceedings of the 2009 IEEE

Peter Pastor, Mrinal Kalakrishnan, Sachin Chitta, Evangelos Theodorou, and Stefan Schaal.
Skill learning and task outcome prediction for manipulation. In Proceedings of the 2011

Peter Pastor, Ludovic Righetti, Mrinal Kalakrishnan, and Stefan Schaal. Online movement
adaptation based on previous sensor experiences. In Proceedings of the 2011 IEEE/RSJ

Peter Pastor, Mrinal Kalakrishnan, Ludovic Righetti, and Stefan Schaal. Towards associat-
tive skill memories. In Proceedings of the 2012 IEEE-RAS International Conference on

Deepak Pathak, Yide Shentu, Dian Chen, Pulkit Agrawal, Trevor Darrell, Sergey Levine,
and Jitendra Malik. Learning instance segmentation by interaction. In CVPR 2018

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via dis-
agreement. In Proceedings of the 36th International Conference on Machine Learning,
2019.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real
transfer of robotic control with dynamics randomization. In Proceedings of the 2018 IEEE

Jan Peters and Stefan Schaal. Natural actor-critic. Neurocomputing, 71(7-9):1180–1190,
2008.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In
Proceedings of the 24th AAAI Conference on Artificial Intelligence, pages 1607–1612,
2010.

A. Petrovskaya and O. Khatib. Global localization of objects via touch. IEEE Transac-

Pedro Piacenza, Weipeng Dang, Emily Hannigan, Jeremy Espinal, Ikram Hussain, Ioannis
Kymissis, and Matei T. Ciocarlie. Accurate contact localization and indentation depth


Kroemer, Niekum, and Konidaris


A Review of Robot Learning for Manipulation


79


