1 Elephants Don't Write Sonnets
The Grounded Turing Test for embodied AI.

Stefanie Tellex and David Watkins

1.1 Introduction

In 1950, Turing proposed the Imitation Game as a test for intelligence. He predicted that in 50 years' time, there would exist a computer that could outperform humans in the Imitation Game. The criteria for success is that a computer emitting written language can fool humans more than 70% of the time. In 2023, LLMs such as GPT-4 and Jurassic-2 have more or less met this criteria (AI21 2023). For example, ChatGPT can easily respond to one of Turing's questions and write a sonnet about the Forth Bridge (Figure 1.1). It is clear that the super-human ability to process the entire Internet into a deep neural network can create a computer program that can pass many aspects of the Turing test.

Turing explicitly scoped his test to be disembodied; in contrast, we focus on embodied systems, which we define as a computational agent embedded in space and time, with high-dimensional continuous sensor input, and high-dimensional continuous motor output. Turing first proposed disembodied language as a test for what it means to be intelligent because the power (and terror) of language is that it can connect to every aspect of an intelligent system, and thus he believed it was a sufficient condition to demonstrate intelligence.

Yet he was wrong: despite the ability of large language models like GPT to write sonnets more quickly and easily than any human, we claim, along with many others, that they are not truly intelligent (Srivastava et al. 2023). When defining intelligence, we seek something more integrative, something that spans the broad spectrum of capabilities exhibited by human intelligence. And it is now clear that language itself is one of those capabilities. Just as calculators can do math faster than a human, we now have "calculators" such as ChatGPT that can write sonnets faster. Turing's hypothesis that language was

Upon the Firth, where waves in rhythm glide,
A marvel stands, in grandeur unconfined.
Its steel embraces winds, a timeless guide,
A testament to human craft and mind.

With spans that stretch o'er waters deep and wide,
It spans the gap where the earth and sky align.
In arcs of strength, its form does proudly stride,
Defying age, a beacon that will shine.

In days of yore, when steam and sweat combined,
Bold engineers conceived its towering frame.
Their vision etched in every bolt, each line,
A monument to courage, not to fame.

So let us praise the Forth Bridge's might,
A symbol of mankind's relentless flight.

Figure 1.1
ChatGPT's response to Turing's prompt, "Please write me a sonnet on the subject of the Forth Bridge." In Turing's original dialog, the competitor refused to do this task because they "never could write poetry."

a sufficient test for an intelligent machine is false: we claim that that language on its own is neither necessary nor sufficient for intelligence.

Something is still missing from our programs. Brooks (1990) memorably pointed out that elephants do not play chess, and they do not write sonnets either. Many aspects of intelligence that are present in animals are not captured by the Turing Test. Specifically, we claim that what is missing is embodiment: our programs must be robots, agents embodied in space and time with high-dimensional sensor input and high-dimensional motor outputs. For example, Figure 1.2 shows a crow filling a cup of water with rocks so it can drink the water. We consider this behavior intelligent even though a crow cannot use human language.

Why, then, should we study language? Language is about the world; language is situated in space and time. Fundamentally, humans use language to talk about what they perceive, believe, and do. Human language is unique in the universe (so far as we know) for its ability to compose novel sentences that have never previously been uttered or thought by any person before. We can also change our topic of conversation from things directly in front of us, like

Figure 1.2
A crow drops rocks into a tube of water so that it can drink the water. Recognizing the displacement of water, that there is a boundary to itself when interacting with the physical world, and requiring temporal actions to accomplish something in the future are all attributes of intelligence not captured in the original Turing test. Image from the Audubon (Saha 2023).

"pick up the speck of dust off the floor" to abstract discussions of philosophy such this chapter. Because of the universality and power of language, Turing proposed it as a medium through which to conduct a test for intelligence.

While human language is not necessary for an intelligent agent (as established by the existence of elephants, rats, and crows), it is still a means of peering into the mind. Human language is a powerful way to penetrate the veil of the skull, building on the mental substrate that exists in humans to create a powerful and flexible framework that spans sounds, written words, gestures, posture, and simply being in a physical spot at a specific time. Language provides a unique window into human cognition because it can connect to all aspects of a cognitive agent.

More generally, language is a means of communication in the information-theoretical sense as defined by Shannon (1948). A bird calling to its flock, a horse moving around in its herd, and even RNA codons all form languages because they exist in space-time and have a cause-and-effect relationship with other entities. For example, horses establish and communicate a hierarchy within their herds through movement towards and away from other horses and resources (Ransom and Cade 2009). Harnad (1990) formalized this intuition as the symbol grounding problem, pointing out that words refer to things in the external world.

This chapter proposes to use language and communicative actions to formulate a new benchmark for intelligence, a grounded Turing Test. If we think of communicative acts and language as the debugging print statement for intelligent creatures, we can enumerate the different kinds of ways language is

used, e.g., interpreting instructions, statements about the world, and information about mental states, delivering information, asking for help, and answering questions. An embodied agent that can use language in all of these ways passes the test, and we claim this is a sufficient (but not necessary) condition for intelligence. In this chapter we review all of these ways of using language along with technical approaches that address these ways of using language. Our field's grand research challenge is to bring them together into one system, embedded in an agent with high-dimensional input and output in the physical world.

1.2 What Is Intelligence?

As roboticists, we frame the problem of intelligent behavior as an agent that interacts with the world in a goal-directed way with high-bandwidth input from sensors and high-bandwidth output through actuators. To act in a goal-based way, the agent needs 1) the ability to process high-framerate information from the environment such as vision, tactile information, audio etc. 2) the ability to respond through its actuators in ways that achieve its goals over time. This embedding in space-time with long-term goal-directed behavior is essential for what we mean by intelligence. For mammalian biological embodied agents, such as humans, the lowest level substrate for acting on the world is through muscle movements. Robotic systems have a completely different set of actuators that can affect the world, for example LEDs, motors, and speakers. Ultimately, they all necessitate high-framerate interaction with the world to produce goal-directed behavior in their changing environment. This forms a hierarchy of communication that allows embodied agents to communicate their internal state, as shown in Figure 1.3.

Brooks (1990) pointed out that elephants do not play chess; similarly, we consider elephants intelligent even though they do not use human language. The elephant makes plans; the elephant has goals; the elephant engages in goal-directed behavior in the physical world. Dennett (1989) pointed out that in this situation we can ascribe intentions and goals even to a fly, which may be doing very bounded and well-defined computation as it moves through the world. In approaching the question of what is intelligence, we are like the four blind men and the elephant. In the story, four blind men approach an elephant from different angles, and each one describes it differently, as a snake, a fan, a tree trunk, a spear. Our position is that intelligence is all of these things at once, and an intelligent agent must exhibit multiple behaviors embedded in space and time to pass our test.

Others have observed that intelligence does not require language and have proposed extensions to the Turing Test to account for this fact. Straightforward

extensions to the Turing test to embodied agents (Zador et al. 2023) lead to attempts to cross the uncanny valley, such as Gemanoid (Nishio, Ishiguro, and Hagita 2007; Leshchev 2021), again sidestepping what we mean by intelligence. Harnad (1991) proposed the Total Turing Test, and Schweizer (1998) extended it to the Truly Total Turing Test, focusing on both linguistic and physical behavior. Our approach falls within this family, but we specify the sorts of linguistic and robotic behavior that must be displayed by the system. Zador et al. (2023) define an embodied Turing test in terms of animal behavior; for example, an embodied Turing test for a beaver tests the beaver's ability to build a dam. They point out that animals 1) engage their environments, 2) behave flexibly, and 3) compute efficiently. Our framework is a form of an embodied Turing test because it is essential that an agent is grounded in the physical environment. However, the specific physical behavior, such as building a dam, is not the point; rather, we employ language grounding—the ability of an agent to connect words to perceptions and actions in the world—as a window into the agent's abilities to take action, leading to a test that is independent of a specific physical embodiment, even though it requires one. Srivastava et al. (2023), in contrast, proposes a text-based, non-embodied extension to the Turing test, called the Beyond the Imitation Game Benchmark (BIG-bench), designed for the age of large language models. BIG-bench consists of 204 text-based tasks from a diverse array of domains ranging from linguistics to biology to common-sense reasoning. Current LLMs perform poorly on these tasks in an absolute sense; although performance improves with model size, it is nowhere near the level of human raters (Mirzadeh et al. 2024). But even if (when!) models pass these benchmarks, we argue that because the resulting model will not be embedded in high-dimensional space-time with goal-based behavior, we should not consider it intelligent.

1.3 The Grounded Turing Test

In this section we enumerate the ways language can be used by an intelligent robot, and then describe the research problems inherent to each. We define a test that is sufficient for establishing the agent is intelligent: To pass the grounded Turing test, a system must support all of these capabilities rather than any single one. In addition, we make no claims that passing our test is necessary for intelligence; like the elephant, it is not necessary to use language to be intelligent, as shown in Figure 1.3. We follow the breakdown in our survey paper (Tellex et al. 2020) and first consider an agent responding to information provided by a person, then an agent providing information to a person in response to their questions, and finally a fluid collaborative dialog as described by Clark (1996). Much existing work in the field follows this breakdown into

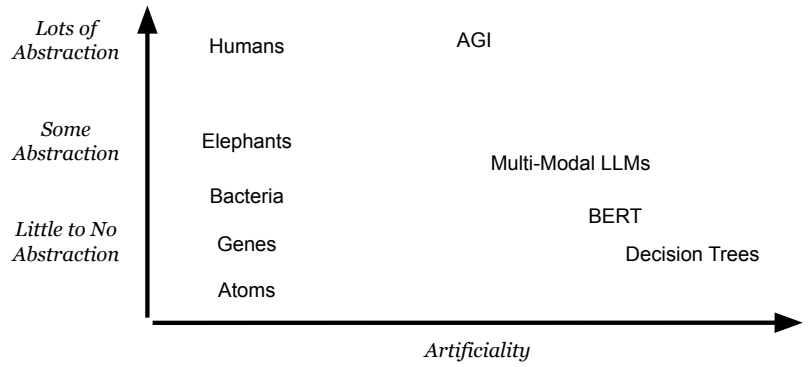


Figure 1.3
The x-axis describes the artificiality of the system and the y-axis describes the levels of abstraction the entity has. The capacity for abstraction happens at many layers of physicality, whether for an atom, a gene, an elephant, or a human. In parallel, disembodied abstractions show up in the form of decision trees (Devlin et al. 2019) or multi-modal LLMs, but their abstractive power is hindered by their disembodiment. To achieve what an embodied intelligence, physicality in space-time is required.

subproblems because it enables a focus on uni-directional communication and also supports evaluation: did the robot do the right thing in response to this one (often text-based) command? Evaluating the success of a collaborative dialog is more expensive because it requires a dialog partner.

Crucially, to pass the Grounded Turing Test, an agent must be embodied, acting in the world, processing high-dimensional, high-frequency input from its sensors, and producing high-dimensional, high-frequency output from its actuators. It is the robot's behavioral response to language input that defines whether it succeeds or fails. Intelligence can only take place in the context of a computational process enacted over time and space. This substrate is the stage on which goal-oriented intelligent behavior plays out. Table 1.1 summarizes each of the sorts of language in our test along with a nominal example from each category.

1.3.1 Human-to-Robot Communication

Human-to-robot tasks consist of the human providing some kind of input to the robot, either verbal, gestural, or behavioral, and then observing its response.

Interpreting Instructions Interpreting instructions means mapping between language and some action in the external world. The challenge is to map

Problem Setting	**Example**
Human to Robot Communication	
Interpreting Instructions	**H:** Pick up the red block. **R:** <Picks up the red block>
Interpreting Statements About the World	**H:** The red block is on the table. **R:** <Updates world model. Later gets the red block when asked.>
Interpreting Information About Mental State	**H:** I want the red block. **R:** <Gives the red block to the person.>
Robot to Human Communication	
Delivering Information	**R:** The red block is on the table next to the cup.
Asking for Help	**R:** Can you give me the red block?
Asking Questions	**R:** Where is the red block?
Answering Questions About Perceptions	**H:** Where is the red block? **R:** I see it on the table.
Answering Questions About Mental States	**H:** What are you trying to do? **R:** Get the red block.
Answering Questions About Beliefs	**H:** Where is the red block? **R:** You told me that it's on the table.
Answering Questions About Actions	**H:** Why did you drive to the table? **R:** You want me to pick up the red block, and you told me that the red block is on the table.
Human-Robot Dialog	Fluid dialog in different settings with all of the above elements.

Table 1.1
Classes of linguistic interaction and nominal examples from each. **H** indicates the human speaking; **R** indicates the robot. Ultimately we imagine fluid dialog where the speaker role does not matter; however we separate it here because the behavior of the robot is very different depending on the role it plays in each dialog turn. A robot that can flexibly engage in all of these types of language and behaves appropriately passes the Grounded Turing Test.

between words in language and objects, places, and actions in the external environment, enabling the robot to choose low-level, mid-level, and high-level actions to take in the world, to change the world state according to the goal expressed in language. Compared to other interfaces, language is unique because a person can give commands to the robot at a variety of levels of abstraction, from very low level commands (e.g., "move your arm up just a bit, okay there, stop!") to very high-level commands (e.g., "clean the kitchen") that might take hours to fully execute.

Robots can use large language models (LLMs) to understand language more flexibly and more capably than ever before. Our review (Cohen et al. 2024) situated the literature into a spectrum with two poles: 1) mapping between language and some manually defined formal representation of meaning, and 2) mapping between language and high-dimensional vector spaces that translate directly to low-level robot policy. Using a formal representation allows the meaning of the language to be precisely represented, limits the size of

the learning problem, and leads to a framework for interpretability and formal safety guarantees. Methods that embed language and perceptual data into high-dimensional spaces avoid this manually specified symbolic structure and thus have the potential to be more general when fed enough data, but require more data and computing to train.

For example, our recent paper, Lang2LTL (Liu et al. 2023) enables grounding natural language commands to task specifications in a widely used formal language, Linear Temporal Logic (LTL). Grounding navigational commands to LTL leverages its unambiguous semantics for reasoning about long-horizon tasks and verifying the satisfaction of safety constraints. Existing approaches require training data from the specific environment and landmarks that will be used in natural language to understand commands in those environments, but Lang2LTL leverages LLMs to ground temporal navigational commands to LTL specifications in environments with no prior language training data. Figure 1.4 shows example commands from our evaluation set.

Regardless of the approach taken for command understanding, it is critical to address the problem of hierarchical abstraction. Language can specify very low-level commands, such as "move your arm one centimeter" and very high-level abstract commands such as "clean the kitchen," so we must connect language to high-dimensional perception from the physical world and high-dimensional action output over timescales from seconds to minutes to hours.

Interpreting Statements about the World Language provides a symbolic way of providing spatial and temporal information about the external world to an embodied agent. A robot must be able to map statements about the world to predictions about future perceptual input and the effects of its future actions. A statement like "the block is on the table" is fundamentally a statement about where the robot can look to find the block in the physical world. Animal alarm calls are of this nature; for example, Templeton, Greene, and Davis (2005) showed that chickadee alarm calls encode information about predator risk, enabling them to communicate that there were smaller (and therefore more maneuverable and thus more dangerous) predators to other members of their flock and eliciting a more energetic mobbing response. We should be able to observe our intelligent robot changing its behavior in response to these sorts of communicative statements about the world, showing that it can adjust its behavior based on language from its human partner.

For example, Walter et al. (2013) showed a method for incorporating information from language into the robot's ability to make geometric maps of its environment. After a person informs the robot that "the kitchen is down

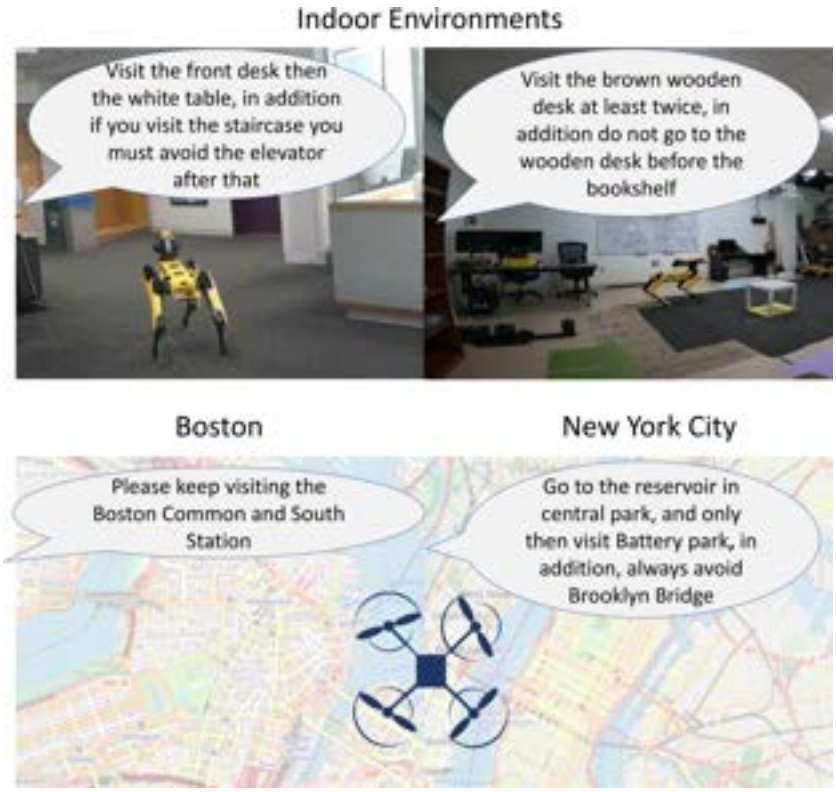


Figure 1.4
Lang2LTL (Liu et al. 2023) can ground complex navigational commands in household and city-scaled environments—without environment-specific training.

the hall," the robot incorprates this information into its mapping system to construct a more accurate geometric and semantic map, as shown in Figure 1.5.

Similarly, our work on object search demonstrates that a robot can use language to more efficiently find objects, illustrated in Figure 1.5. Zheng et al. (2021) enabled a person to use spatial language to describe object locations and their relations to a robot. The robot could then use this information to more efficiently find an object. We considered spatial language a form of stochastic observation. To model ambiguous, context-dependent prepositions (e.g. the car in front of the building), we designed a convolutional neural network that predicts the language provider's latent frame of reference given the environment context. Search strategies are computed via an online POMDP planner.

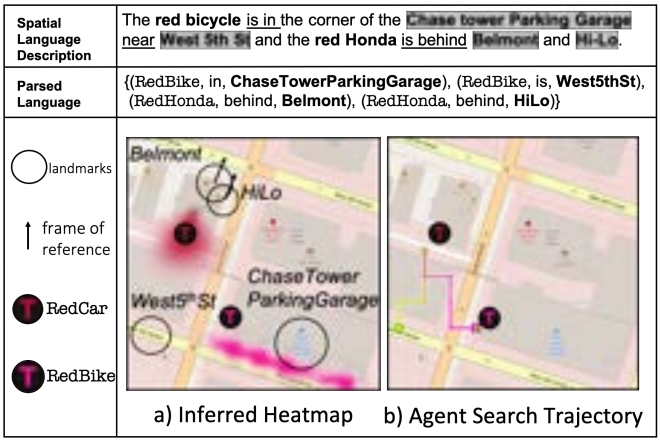


Figure 1.5
Top: A robot interprets information from language to improve its ability to make semantic maps of the environment (Walter et al. 2013). Bottom: A robot uses language information to better search for objects (Zheng et al. 2021).

Figure 1.5 shows the probability distribution that results from the language input, which then biases the search process by incorporating that information to find the object more quickly.

More generally, we need ways for humans to tell robots information about the world using language, both the locations of objects and advice on how to execute skills (e.g., "you can sweep better if you use the large broom"). This capability should lead to shared planning and problem-solving mediated by human-robot dialog. To address this problem, it is important to have shared

representations for meaning and action. Herb Clark referred to this as establishing "common ground" (Clark 1996). By establishing common ground between the person and the robot, shared states of knowledge—the first steps towards collaboration—are possible. This problem requires collaborative dialog to build up shared representations about word meaning over time.

Interpreting Information about Mental States A key aspect of the external world is other agents' mental states: their goals, desires and beliefs about the world. Based on the person's behavior, both verbal and non-verbal, the robot needs to make inferences about what the person wants. Grice (1975) defined the cooperative principles that people use when communicating with each other: quantity (say no more than what is necessary), quality (say things that are true), relation (say things relevant to the discussion), and manner (say things clearly, briefly, and orderly). Sometimes violating these maxims itself has communicative efficacy, as when a person uses an implicature. For example, when a person states, "this coffee is cold," at a lexical level is a factual statement about the world. However, by the maxim of relevance, this statement implies the mental state in the person that they want hot coffee, and a helpful robot or waiter should fetch it for them. Building on conversational implicature, one can imagine a robot making inferences about a person's beliefs, goals, and desires, interpreting a person's ironic or sarcastic statements, laughing at jokes, and engaging in collaborative planning dialog on complex tasks.

Vogel et al. (2013) showed a computational approach for the emergence of Gricean maxims from multi-robot decision theory. This approach showed that using a decentralized-POMDP framework (Dec-POMDP) enabled robots to learn or generate a language to communicate the state of the world to each other in a partially observed setting. This work showed that a ListenerBot that did not take into account the belief or actions of its partner did not perform as well as a DialogBot that maintains a model that fits the partner's beliefs. This work was limited in its ability to generalize to new tasks, was extremely computationally challenging, and required robots "collude" with each other at the beginning; the language they use to communicate with each other emerges computationally from the fact that they have access to a model of their internals at the beginning of the task, even if they can only communicate in limited ways during the task.

[**DW:** Chatter paragraph] In recent work, Kross (2021) looks at the effect that positive affirmations have on mental states of a person. It points to the generative nature of the internal monologue in the human mind. In Chatter, Kross argues the internal monologue shapes our collective experience. This becomes a test for understanding the internal human state and the debugging nature

of language. Cognitive Behavioral Therapy and Cognitive Processing Therapy exploit similar phenomena in order to repair the way humans think. The Rorschach test is a primitive evaluation that humans have employed to debug what humans are predicting in their perception of the world.

Open research questions in this space include how to interpret a person's goals from their physical and social actions, predicting how the robot's actions will affect the person's beliefs about them (e.g., Dragan, Lee, and Srinivasa (2013) described a framework for reasoning about predictable motion that aims to achieve a task as efficiently as possible vs. legible motion that aims to communicate with an observer about what task the robot is performing, and Tellex et al. (2014) which reasons about how to ask for help by modeling a person's mental state in order to uniquely and concisely specify an object).

1.3.2 Robot-to-Human Communication

Robot-to-human communication primarily focuses on question answering, although a robot could also choose to unilaterally initiate dialog with a person, for example to tell them about something dangerous. Answering questions from a human interlocutor is a fundamental capability of any robot that uses language. We divide question answering into subsections based on the type of question and the information provided.

Delivering Information A robot should have the means to convey information to other embodied agents. A recent study on elephants used machine learning to differentiate the sounds that they make to each other to determine that they have a naming scheme (Pardo et al. 2024). Under this paradigm, we should also be able to distinguish the gestures and communication that an embodied agent has. While baking English into such a system may give the illusion of delivering information to a user, we need to be able to indicate holistically that the behavior has intent instead of an LLM detached from the reality the agent finds itself in.

A robot could deliver information to a person for many reasons. It may be tasked with telling the person if the environment changes in some way; it may exist as an information delivery robot, for example, to give people a tour or welcome in a building (Lee, Kiesler, and Forlizzi 2010; Bohus, Saw, and Horvitz 2014). In this setting, it is key for the robot to have some model of what information is relevant to convey to the person based on their current knowledge and goals. Rendering this information to text is easily done by LLMs; the key problem is to model what the person wants to know and to find the relevant information in the physical world. Previous work has approached this problem by pre-programming the robot with information about the world,

Figure 1.6
A robot engaged in assembling an IKEA LACK table requests help using natural language. A vague request such as "Help me" is challenging for a person to understand. Instead, Knepper et al. (2015) presents an approach for generating targeted requests such as "Please hand me the black table leg that is on the white table."

as in Bohus, Saw, and Horvitz (2014), which placed a tour guide robot at the entrance of a lab area and endowed the robot with knowledge of the layout of the lab.

Asking For Help The ability to ask for help can enable robots to increase their effectiveness by eliciting aid from a person to bridge gaps in their current capabilities. Additionally, asking for help provides the human partner with information about the robot's goals, as well as availing itself of a key resource for getting things done: its human partner. Knepper et al. (2015) showed that a robot could improve task success rates by asking for help from a person when it got confused, trading off between making its request specific to the problem it needed without providing irrelevant information. Figure 1.6 shows an example where a robot can ask for help in vague terms, such as "help me" or alternately make a more targeted request like "please hand me the black table leg that is on the white table." The challenge for generating specific, useful requests is that the robot needs to model the person's mental state and how they will interpret the robot's words in order to generate a succinct yet specific request that is easy to understand.

Answering Questions About Perceptions Answering questions about the world is a key ability to probe a robot's perceptual system. The robot needs to describe its perception of the environment, not just from a single image but from a sequence of high-dimensional perceptual inputs gathered over time, in

Figure 1.7
A Kuri robot maps its environment and stores information about objects it has seen (Idrees et al. 2021).

a goal-directed way, using language to specify what to pay attention to and what to ignore.

Existing work on visual question answering (Manmadhan and Kovoor 2020) focuses on answering questions about a single image, usually taken manually by a person and framed in a camera viewfinder. Idrees et al. (2021) described an approach where a robot equipped with situational awareness can help humans efficiently find their lost objects. Their approach allows for a variety of query patterns, such as querying for objects with or without the following: 1) specific attributes, 2) spatial relationships with other objects, and 3) time slices. The challenge is that the robot may have partial views of the object and also multiple views of the same object which change over time. Furthermore, at mapping time, the robot does not know what object the person will be searching for later, at query time. Figure 1.7 shows the robot, the metric map of the environment, and spatial-temporal clusters of images which the robot has determined refer to the same object. This clustering enables the robot to efficiently search its memories at query time to find objects that match a natural language description.

Answering Questions about Mental States Researchers in cognitive science have long recognized the importance of mental states and language processing in human behavior (Boyd and Schwartz 2021). The concepts of theory of mind (ToM) and cognitive architectures have been explored in various fields, including linguistics (Villiers 2007; Hale and Tager-Flusberg 2003) and neuroscience (Huyck 2020). Recently, these ideas have begun to influence research in LLMs ("Theory of Mind Workshop at ICML 2024" 2024) and related studies on cognitive-inspired LLMs (Zhu and Wang 2020). In robotics, these concepts are being applied to Human-Robot Interaction (HRI),

particularly in development robotics, where robots need to understand human mental states and develop more sophisticated interaction mechanisms (Langley et al. 2022). Building on this foundation, the grounded Turing test aims to evaluate embodied agents' ability to model human mental states and engage in meaningful interactions.

Intelligence can be understood as an emergent property of complex neural systems arising from the interaction between sensory inputs and cognitive processes. Each transition to understanding what another human is saying, gesturing, or gesticulating can be seen as a transition between mental states, represented as hyperdimensional points in a high-dimensional space. This perspective allows us to describe all states of the mind as a set of N-dimensional points, with a transition function between these points representing the lowest energy representation averaged over time. The values of these points can be mapped onto the state of all synapses in the mind, and intelligence can be seen as the dynamic process of transitioning between these points through space-time. This framework suggests that speech impacts both the speaker and the listener by transitioning their mental states, creating a joint system for processing information.

This perspective on language and intelligence also reveals its debugging nature. Instead of a mind being forced to internalize its understanding of the world, it is driven to find low-energy representations of how to express its internal state to convey meaning to other beings. This idea can be seen as an alternative explanation for Chomsky's universal grammar, where the fundamental basis for language is not an innate property but rather an emergent property of complex cognitive systems interacting with their environment (Chomsky 1965). This emergent property perspective can be further understood through the lens of information theory, which provides a framework for analyzing the transmission of information. As Shannon's information theory (Shannon 1948) suggests, encoding and decoding processes play a crucial role in the transmission of information. Conveying detailed information about its internal state would require significant energy and resources from a physically embodied entity, whereas transmitting abstracted representations is more efficient. It is essential to note that evolution is not optimization (Hertzmann 2024), as it does not always lead to the most efficient solutions; rather, it finds satisfactory ones given the constraints. Language has developed within an organism to convey information as efficiently as it can, but there are limits to the complexity of the information that can be described by the abstraction the agent can make. To convey more information about itself, it needs higher-order representations of itself and the environment it is in. The need for these higher-order representations leads to the concept of Kolmogorov complexity (Kolmogorov

1965), which measures the complexity of an object or sequence in terms of the length of its shortest description, providing a framework for understanding the trade-offs between representation complexity and information transmission efficiency.

Answering Questions about Beliefs As our robot operates and observes the environment with its sensors, it must build a model of its beliefs about the world. In robotics, these problems are characterized as state estimation (Thrun 2002). For example, in the Simultaneous Localization and Mapping (SLAM) problem (Durrant-Whyte and Bailey 2006), the robot must build a model of its environment over time from sensors such as its camera, LIDAR, and motion sensors. The robot should be able to give insight into what it knows about the world by answering questions about its beliefs. Zhong et al. (2022) surveys work in video question answering, which is usually focused on answering questions about surveillance videos or movies and films curated by people. In robotics, the robot must be able to answer questions about its own experiences and inferences it has made about those experiences. For example, Idrees et al. (2021) showed a system that allows a robot to fuse its memories of the environment into an efficient data structure that allows it to answer questions about the locations of missing objects. Related, Das et al. (2018) defined a new AI task of embodied question answering, where an agent in a 3D is asked a question like "what color is the car?" To answer the question, the agent must explore the environment to gather visual information through an egocentric camera.

Answering Questions about Actions Interpretability and explainable AI are key concerns as robots become more autonomous and more capable. A robot needs to be able to explain its behavior (both successes and failures), and answer questions about why it did what it did (both successes and failures).

Many approaches have used symbolic methods to enable interpretable AI. For example, Li et al. (2019) created a framework for interpretable reinforcement learning for robotic planning using linear temporal logic. Raman et al. (2024) created a framework for learning grounded symbols using LLMs. The system creates a language-based, perceptually grounded symbol by prompting an LLM, making it easy for a planner to generate explanations of why the system did what it did. Das, Banerjee, and Chernova (2021) studied different approaches for generating interpretable explanations for failures, showing with a user study that untrained humans prefer explanations that include the context and history of the robot's actions, and that this additional information enables them to more accurately identify the causes for a robot's failure (Chen et al. 2024).

1.3.3 Human-Robot Dialog

When we put the pieces of human-to-robot communication and robot-to-human communication together, we create a collaborative (or adversarial!) dialog between the human and the robot. Clark (1996) describes dialog as a *joint activity* between two humans, analogous to dancing or playing a basketball game. Each participant is constantly observing the other for signals large and small, ranging from backchannel expressions like "uh-huh" to non-verbal cues like hand gestures and eye gaze to explicit speech utterances to physical actions in the world, like picking up an object.

1.4 Passing the Grounded Turing Test

Defining intelligence is challenging because it encompasses a collection of capabilities, not just one. While progress can be made by addressing individual subproblems, integration is crucial to pass the grounded Turing test. To achieve this, we need one system that demonstrates all the required capabilities simultaneously in a physical body with high-dimensional sensory input and high-dimensional motor output. Any single ability does not constitute AI; only by exhibiting flexibility, tenacity, connection with the world, and responsiveness can intelligence be demonstrated. Passing this test is not necessary for a robot to be intelligent, but rather a sufficient one. For instance, rats and crows are intelligent without being able to pass the Grounded Turing Test, highlighting that intelligence exists on a spectrum (see Figure 1.2 for an example of such behavior). The spectrum of intelligence is highlighted in Figure 1.3, showing the abstraction capabilities different organisms can develop and comparing them to existing AI systems.

We must move beyond traditional machine learning paradigms that rely on static, pre-collected datasets to advance towards true artificial general intelligence. Current approaches often involve training models offline, which limits their ability to adapt in dynamic environments. Instead, we must develop AI systems capable of continuous online learning, where agents can process and learn from high-dimensional sensory inputs in real-time. This approach is crucial for embodied agents, such as robots, that must interact with and adapt to their environment moment-to-moment. By enabling agents to learn from ongoing experiences, we can build systems that evolve and improve without the need for periodic retraining on curated datasets. Sutton (2019) suggests that the "bitter lesson" of machine learning highlights the importance of systems that can learn efficiently from their interactions with the environment, rather than relying on human-labeled data.

1.4.1 Sensorimotor Substrate

To build any framework for generalized intelligence, we need a sensorimotor substrate (see Chapter 13 of this book), which can be characterized as the things an elephant (or a rat) needs to move, navigate, and survive in the physical world. For example, Milford, Wyeth, and Prasser (2004) proposed RatSLAM, an implementation of a model of a rodent's hippocampus that combines properties of grid-based, topological, and landmark-based representations for localization and mapping. Similarly, the brain contains specialized submodules for visual perception, manipulation, tactile sensing, reflex responses, and more. These submodules enable their respective problems to be solved on a submanifold where the problem's structure can be exploited more efficiently. For example, for navigation in an Euclidean space, one can switch from breadth-first-search to A* and exploit a strong heuristic to increase inference speed dramatically. Similarly, for SLAM, the Rao-Blackwelized particle filter exploits the conditional independence of landmark locations relative to the robot's trajectory (Durrant-Whyte and Bailey 2006). Whether learned or designed by a person, these structures are essential for efficient inference in the physical world. The structures are critical for being able to model both the intended behavior and the behavior of those around the system.

1.4.2 The Human Robot Collaborative POMDP

Next, we need a unified framework for reasoning and acting in the world. Over the past fifteen years, we have built towards a new model for human-robot interaction that enables language interpretation and question-asking using a hierarchical representation for planning under uncertainty, called the Human-Robot Collaborative POMDP (Partially Observable Markov Decision Process). It builds on the POMDP framework (Kaelbling, Littman, and Cassandra 1998) framework because POMDPs enable the robot to reason about its beliefs about the world over space and time. POMDPs are the simplest model we are aware of that captures everything that a robot is: states that evolve, with noisy observations of the external world, goals, and the ability to take actions that achieve those goals. However, finding a policy for the robot based on its observations of human communication and actions is undecidable (Madani, Hanks, and Condon 1999).

To reason efficiently about high-dimensional perceptual inputs, our agent needs to acquire structured state representations and associated factored inference algorithms for the POMDP that lead to efficient reasoning and inference. We note that these representations and algorithms can be either designed by humans or learned by algorithms. This factored, hierarchical structure leads to information-gathering behavior in a framework that supports the efficient

incorporation of multimodal observations from low-level sensors, human language, and gestures, as well as background knowledge from large foundation models.

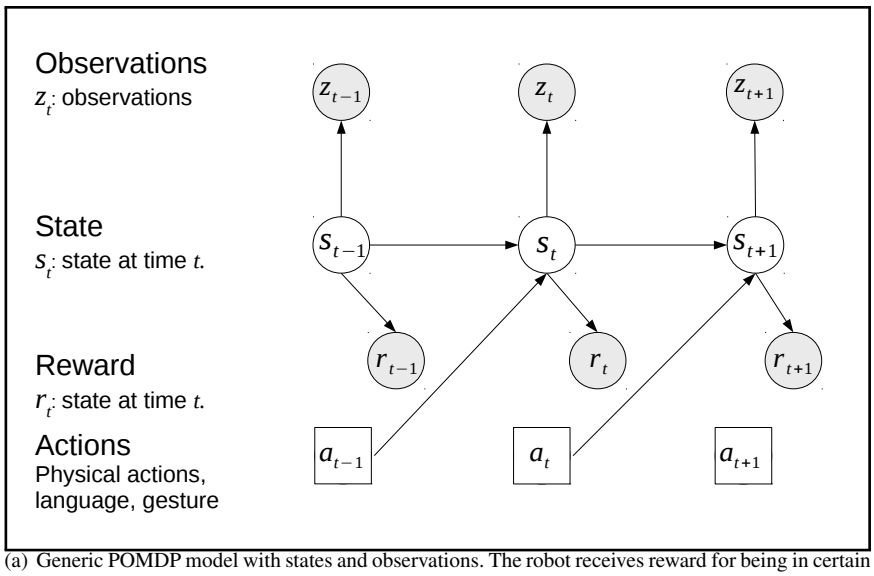
We define the Human-Robot Collaborative POMDP as a tuple, $\langle S, A, O, R, T, \Omega \rangle$, where:

- S is the set of states. Each state, s, is a tuple, $\langle O, h \rangle$ where h is the mental state of the human and robot, and O is the physical state of the world, factored into objects.
- $a \in \mathcal{A}$ is any action that the robot could take.
- z_t is the set of observations from the robot. It consists of a tuple of multimodal information streams, $\langle z_t, l_t, g_t \rangle$, corresponding to physical sensor information, language, and gesture.
- r_t is the reward at time t, which we define as achieving the person's aims, encoded in their mental state h_t.
- T: $p(s_{t+1} \mid s_t, a_t)$ is the transition function, which models how the world changes after the robot's action. This transition functions in terms of physical state and mental state.
- $\Omega(s', a^R, a^{H'}) = p(z_t, l_t, g_t \mid O_t, h_t)$, which factors into terms for physical and mental state.

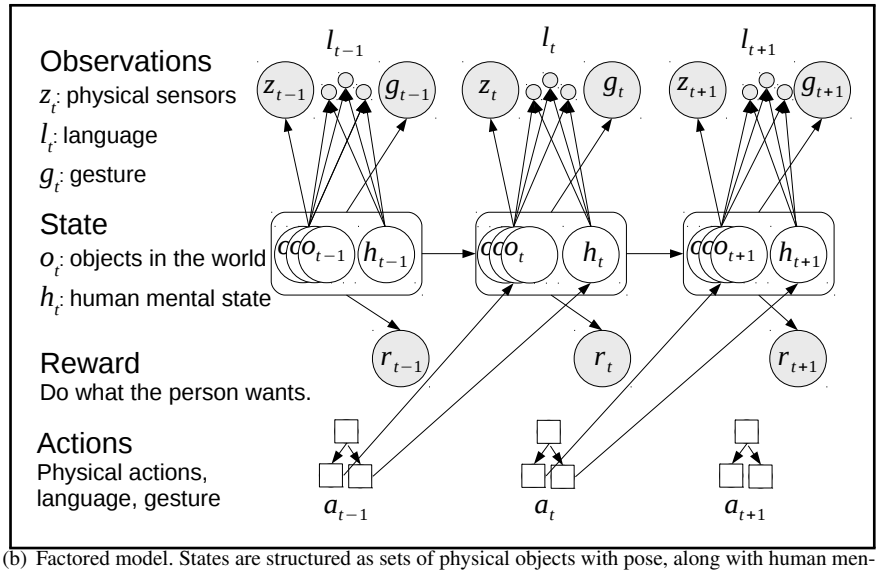
This structure is summarized in Figure 1.8. The Human-Robot Collaborative POMDP unifies the conditional independence assumptions and state representations. This unification is more than the sum of its parts: it enables the robot to interpret a person's requests, ask questions, and ask for help in a decision-theoretic framework by modeling its uncertainty about what a person wants and its own ability to interpret requests. The challenge in a unified framework is efficient inference when handling large observation spaces and commands at different levels of abstraction. The following sections propose new techniques for enabling efficient inference in this framework.

1.4.3 Compositional Learning

Intelligent agents need to infer abstract concepts in terms of causal relationships from high-dimensional perceptual input and make plans for high-dimensional motor output. Synchronicity is defined by Carl Jung as finding meaningful causal relationships in co-occurances of physical phenomena (Jung 1960). Pearl (2009) built on this idea by showing that the language of probabilistic graphical models can be used to enable inference of causal relationships from data. Inferring correct causal relationships and corresponding conditional independence then enables more efficient learning.



(a) Generic POMDP model with states and observations. The robot receives reward for being in certain states. Its goal is to pick actions to maximize its expected reward. States are Markov and change based on the robot's action and the current state. The robot cannot directly observe the state; it only observes noisy observations (shaded), obtained from its sensors.



(b) Factored model. States are structured as sets of physical objects with pose, along with human mental state. The robot plans efficiently using a hierarchy of physical actions and communication actions. Observations are factored into physical observations and human language and gesture. This structure enables efficient learning; our claim is that this structure itself must be learned, along with model parameters.

Figure 1.8
Human-Robot Collaborative POMDP.

Compositionality arises from conditional independence assumptions and enables robots to apply learned behaviors in various contexts, combine behaviors in new ways, and use language, perception, and action skills to accomplish complex tasks. This ability needs to be built on the sensorimotor substrate, described in the following section, and then built up to create layers of abstract concepts that connect to language.

1.4.4 Data-Driven Robot Learning

Leveraging the stream of high-dimensional, high-throughput data in an offline setting exists in existing methods for training robot policies. Existing policy models, such as diffusion policies (Chi, Xu, Feng, et al. 2024), ACT (Zhao et al. 2023), or Mamba policy (Cao et al. 2024) are all candidates for training specific behaviors or skills for robots. What is important about any method is the ease of data acquisition, the support for multiple sensors, and a method that can target multiple embodiments. Existing works, such as UMI (Chi, Xu, Pan, et al. 2024) or ALOHA (Zhao et al. 2023), accomplish this by building handheld portable devices or inexpensive teleoperated robotic platforms, respectively. We can leverage simulation, virtual-reality teleoperation, or hand-held human collected demonstrations to seed robot policies, and then further train them using reinforcement learning methods.

The primary objective of optimizing is scaling and multi-modality. We hypothesize that we can significantly reduce the data required for robot policy learning by increasing the number of modalities the robot can access. We can then further observe positive transfer between those skills by collecting a massive sweep of skills encompassing all behavior we can imagine the robotic platform is capable of. This data collection enables further multi-task and representation learning research to enhance robotic execution. We will also start accomplishing different aspects of the Grounded Turing Test by transferring human knowledge and the ability to use tools to these robots. Our joint research is actively testing this hypothesis at the RAI Institute.

We need to build a generative model that is capable of predicting future states from the ongoing stream of high-dimensional data. This is going to take in visual, tactile, force, audio, and more data in a GPT-style model architecture and autoregressively predict future states based on this data. Because of this model, we can use it to determine whether the current observations we are seeing now fall within our distribution of past observed states, by comparing against what we predicted with what we observed. This is following similar models to how humans are able to process language by predicting what people are about to say to process concepts and achieve alignment during a conversation [citation]. This generative model is another way that we can train robots to

learn language naturally rather than having language as a first order input into the model architecture itself.

1.4.5 Updating the Weights and Structure Simultaneously

Many of the existing problems in robotics involve non-differentiable or non-smooth functions. For contact dynamics, trajectory optimization, control barrier functions, or observation models, each of these features has non-smooth or non-differentiable underlying functions (Qadri and Kaess 2023; Vemula and Bagnell 2020). Existing approaches often struggle to adapt and learn arbitrary functions due to their fixed internal structure when faced with complex, long-horizon tasks. Ideally, a neural network should be able to reorganize itself to find new neural pathways that can effectively tackle a given task. The big problem with the state-of-the-art learning paradigms is that they cannot update the network structure during training. The Chinese Restaurant Process (Blei and Ramachandran 2006) describes a situation where a restaurant has infinite tables, and customers can sit at any table. The first customer sits at the first table, and the second customer has a choice of whether to sit at the second table or at the first. This puzzle describes the same process we, as researchers, have to do with data passing through a neural network - we are not sure what the correct partitioning of the data is. While we pick a fixed set of parameters to update for our current neural networks, we are not building algorithms that can update the structure with those changes. This is crucial for language, where new concepts are developed through learned experience in the real world.

Many different ways have been proposed to update the structure of neural networks during training. Evolutionary algorithms have existed for decades. However, they have not yet been successfully used for large-scale training. What we are missing in developing language is a loop that allows a system to observe the consequences of its own actions. This is a system that leverages the ongoing estimation of the world, but has enough parameterization to model itself simultaneously. This is how we can build symbols that are generated automatically by the experience of the embodied agent. We can leverage the ever-expanding neurons to handle a wider variety of problems. We want language to develop organically from the fact that the robot learns rather than as a first-order input into the system.

Taking these ideas further, we have not seen any instances of successful online learning of robots from the realtime stream of data coming in from multiple sensors. Methods such as Sun, Cetin, and Tang (2025) and Behrouz, Zhong, and Mirrokni (2025) offer suggestions for implementing test-time learning strategies. This further suggests we can then schedule learning around learning within a demonstration, between demonstrations, and treat all data

that comes into models as a way of improving performance. This also provides an alternative form of memory for robot policies where the state of the system can now be embedded in the weight space rather than methods that require us to store memory in more recurrent fashions, such as Perceiver IO (Jaegle et al. 2022). We suggest that data-driven methods will be a good direction to take robotics research, in parallel with other methods we have suggested. They provide a logical structure for the human-robot collaborative POMDP, where part of the collaboration is happening in a data collection step.

One such implementation of a self-updating system is gradient-free reinforcement learning, such as those in neuroevolution (Such et al. 2017) and forward-forward (Hinton 2022). Gradient-free methods can simultaneously learn the structure and activation of neural networks (White et al. 2023). Newer works, such as NoProp (Li, Teh, and Pascanu 2025), suggest that we can optimize intermediate layers locally using denoising methods instead. Researchers have demonstrated using fixed-structure networks that they can generalize to the unseen hierarchical organization of tasks but struggle beyond simple pick and place tasks (Vijayaraghavan et al. 2024).

Forward-Forward (Hinton 2022) also implies alternative methods for updating neural network weights by leveraging the embodied nature of systems rather than relying on backpropagation-based approaches. This method involves a reward mechanism that promotes "good" examples within the domain and penalizes "bad" out-of-domain examples. However, effectively training an embodied agent requires a reliable mechanism for generating adversarial examples to define domain boundaries, which is currently lacking. The ultimate goal is to mimic human behavior, as humans represent our collective gold standard for intelligence. Notably, backpropagation lacks biological plausibility (Hunsberger 2017), and while deep learning systems may have some biologically-inspired components, their weight update mechanisms do not mirror human cognitive processes. To make progress, we should shift our focus from the convenient but potentially misleading realm of deep learning and instead explore gradient-free learning methods that prioritize human plausibility and design, avoiding the common pitfall of searching for solutions only where it is easy to look.

One promising approach is to draw inspiration from Neural Architecture Search (NAS) algorithms. NAS involves automatically searching for the optimal architecture for a given task, rather than relying on manual design. By leveraging reinforcement learning mechanisms, NAS methods can update their internal structure to better suit the task at hand (White et al. 2023). If we can reduce the complexity required to update the network architecture, we

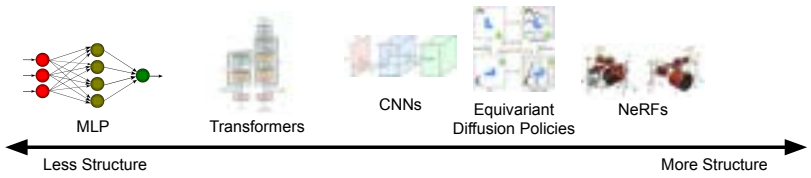


Figure 1.9
Network structure falls on a spectrum. From multi-layer-perceptrons to NeRFs, the amount of structure changes the kind of learning the system can perform and its brittleness in response to out of distribution effects (Rumelhart, Hinton, and Williams 1986; Mildenhall et al. 2020; Wang et al. 2024; Krizhevsky, Sutskever, and Hinton 2012; Vaswani et al. 2017).

may be able to more effectively handle larger-scale tasks. Moreover, considering hyperparameters and neural architectures as analogous to an organism's DNA highlights the potential for NAS to bridge the gap between software and hardware design, enabling us to express instructions that can build complex systems.

Silver et al. (2021) lays out a framework for defining a series of primitive rewards that, over time and collected experience, begin to develop into complex rewards. The reward system evolves into a larger organization via a mirror neuron system when these agents interact. Reward engineering for these primitive rewards is a critical field of research we need to begin to explore in the context of robotics.

By dynamically updating its structure, the robot can learn hierarchical and object abstractions in a way that would be impossible with a fixed network architecture. This is particularly relevant to complex systems like horse herding behavior, which we discussed earlier. A static network would require manual tuning of hyperparameters to capture the nuances of herd dynamics, such as dominance hierarchies and social interactions. By contrast, a dynamic architecture can adapt and refine its representations in real-time by building upon previously learned knowledge to improve performance. For instance, a model could seamlessly adjust to herd size or composition changes, learning to recognize and respond to new social structures without extensive retraining or computational resources. This flexibility enables the model to learn and adapt throughout its lifespan rather than being limited by its initial design.

The structure of the network is critical to getting this right. There is a systems engineering approach where one can train a general MLP on a series

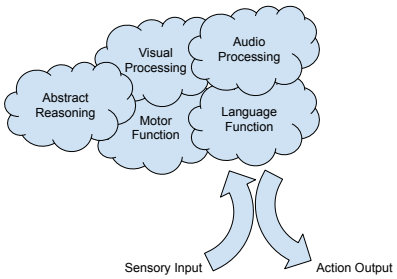


Figure 1.10
By biasing different portions of a proposed gradient-free learning system, we can have emergent structures that inherently lend themselves to different strategies. While these structures are chosen to emulate the brain, in practice, they will likely be more ambiguous in their relative function. It is also possible that the abstract reasoning and language sections will be the same.

of domains and then use a mixture of expert models to produce productive output. However, this approach will likely bias the model too heavily in intermediate results that we, as humans, believe should be relevant for the model to perform well. Instead, one could construct a model with different portions aligned with architectures such as CNNs, Transformers, Diffusion, and RNNs that are more amenable to particular kinds of data than others. The system can then use energy-based optimization to reduce the overall complexity of the representation, thereby enforcing the abstraction of the high-dimensional input data. A critical component is that the sensory data is provided simultaneously in a combined format, such as cross-attended modalities. This forces the potential utilization of sensory information and allows the network to optimize which portions of the data it views. For example, this might allow a robot to learn a closed-loop multi-modal sensorimotor policy for inserting a key into the door, considering vision, the sound of the lock turning, and the changing forces on the hand. Humans have co-opted portions of their brains to serve reasoning skills. Still, that expressive capacity is likely too limited to understand a concrete mathematical formulation for why certain concepts feel "correct."

The missing portion of this is the mirror system. A critical component of human development of intelligence is looking for instances of our own mental framework in things all around us (Cullen et al. 2014; Wan and Chen 2021). We should look to see this as an auxiliary reward signal for training a system better than it would be if we solely relied on external factors. The complex question is, what do we model? Do we build a mirror neuron system for humans? For other robots? For both? Should we model and seek faces as well, to match human language? How important is conveying emotion through

the face? These aspects of ourselves as humans are so critical for our development that it behooves us as researchers to begin teasing out what effects these seemingly purely social attributes have on the quality of the motor systems we have as people.

The human brain's impressive 100-500 trillion synapses (Drachman 2005) allocate only a fraction, roughly 3-5 trillion, for visual processing (Colonnier and O'Kusky 1981). However, as evolution is not optimization (Hertzmann 2024), we must reevaluate our approach to modeling visual processing. We can roughly equate one synapse to one parameter in a neural network (Millidge 2022) within an order of magnitude. The success of models like ResNet in processing single images suggests that even with 25.6 million parameters (He et al. 2016), such biological structures in brains are likely overfitting on sensory input. Zhang et al. (2020) has shown that overparameterization can result in the memorization of inputs. This may lead to issues like hallucinations and erroneous visual beliefs.

By contrast, an online learning agent might require only 50–100 million parameters to process single images effectively. Since images are dependent variables, they may not necessitate extensive video processing systems like CVF or continuous visual flow. Extending this idea to the motor cortex, we can estimate that many tasks do not require intense motor processing, implying a limited need for large-scale models for every subproblem.

This limitation is not due to computational constraints but rather our inability to integrate these models effectively. Despite neural networks being universal functional approximators, we lack the necessary training data and objective functions to evaluate end-to-end systems. Our goal in this chapter is to establish a foundation for addressing these challenges and developing more efficient and effective models for robot systems.

1.5 Closing Thoughts

"A picture is worth a thousand words." What is beautiful about language is not the words themselves but the circumstances that created those words, which are high-dimensional sensory inputs collected over time through active interaction with the physical world. Modern-day AI has surpassed the original Turing Test and failed to inspire the spark of intelligence we see in living organisms. It behooves us to invent another metric that captures the relationship between intelligent thought and existence in the world.

In this work, we have defined a behavioral test for intelligence. But what about consciousness? In the philosophical tradition of the Turing Test, we aim to sidestep this question and ask only if our robot supports these behaviors, leaving the question of whether it is conscious or not to the philosophers. It is

possible, even likely, that our Grounded Turing Test is not complete, and we need to add even more behaviors to our list before people agree that our robot is intelligent. Regardless, the series of software and hardware systems we need to build a physical AGI is now coming into focus. We can conceive the system we need to design, the impact that it will have, and what kinds of resources we need to allocate to study its behavior. With this technology within reach, we imagine a world not where robots replace man, but instead where novel systems of augmenting people make us more productive as a species.

Acknowledgments

We would like to thank Howie Choset, Gustavo Goretkin, Dogan Yirmibesoglu, Jessica Hodgins, Reena Leone, and George Konidaris for their comments on drafts of this paper.

References

AI21. 2023. *AI21 Labs concludes largest Turing Test experiment to date.* https://www.ai21.com/blog/human-or-not-results. Accessed: 2024-05-01.

Behrouz, A., P. Zhong, and V. Mirrokni. 2025. "Titans: Learning to Memorize at Test Time." *arXiv:2501.00663.*

Blei, D., and B. Ramachandran. 2006. "The Chinese Restaurant Process and Bayesian Nonparametric Inference." *Journal of Machine Learning Research* 7:1219–1244.

Bohus, D., C. W. Saw, and E. Horvitz. 2014. "Directions Robot: In-the-Wild Experiences and Lessons Learned." In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems,* 637–644.

Boyd, R. L., and H. A. Schwartz. 2021. "Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field." *Journal of Language and Social Psychology* 40 (1): 21–41.

Brooks, R. A. 1990. "Elephants Don't Play Chess." *Robotics and Autonomous Systems* 6 (1-2): 3–15.

Cao, J., Q. Zhang, J. Sun, J. Wang, H. Cheng, Y. Li, J. Ma, Y. Shao, W. Zhao, G. Han, et al. 2024. "Mamba Policy: Towards Efficient 3D Diffusion Policy with Hybrid Selective State Models." *arXiv:2409.07163.*

Chen, L., O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton. 2024. "Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving." In *Proceedings of the 2024 IEEE International Conference on Robotics and Automation,* 14093–14100.

Chi, C., Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. 2024. "Diffusion policy: Visuomotor policy learning via action diffusion." *The International Journal of Robotics Research.*

Chi, C., Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. 2024. "Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots." In *Robotics: Science and Systems XX.*

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: The MIT Press.

Clark, H. H. 1996. *Using Language.* Cambridge University Press.

Cohen, V., J. X. Liu, R. Mooney, S. Tellex, and D. Watkins. 2024. "A Survey of Robotic Language Grounding: Tradeoffs between Symbols and Embeddings." In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence,* 7999–8009.

Colonnier, M., and J. O'Kusky. 1981. "Number of neurons and synapses in the visual cortex of different species." *Revue Canadienne de Biologie* 40 (1): 91–99.

Cullen, H., R. Kanai, B. Bahrami, and G. Rees. 2014. "Individual differences in anthropomorphic attributions and human brain structure." *Social Cognitive and Affective Neuroscience* 9 (9): 1276–1280.

Das, A., S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. 2018. "Embodied Question Answering." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1–10.

Das, D., S. Banerjee, and S. Chernova. 2021. "Explainable AI for Robot Failures: Generating Explanations that Improve User Assistance in Fault Recovery." In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction,* 351–360.

Dennett, D. C. 1989. *The Intentional Stance.* MIT press.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 4171–4186.

Drachman, D. A. 2005. "Do we have brain to spare?" *Neurology* 64 (12): 2004–2005.

Dragan, A. D., K. C. Lee, and S. S. Srinivasa. 2013. "Legibility and Predictability of Robot Motion." In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction,* 301–308.

Durrant-Whyte, H., and T. Bailey. 2006. "Simultaneous Localization and Mapping: Part I." *IEEE Robotics & Automation Magazine* 13 (2): 99–110.

Grice, H. P. 1975. "Logic and Conversation." In *Syntax and Semantics 3,* edited by P. Cole and J. Morgan, 43–58. New York: Academic Press.

Hale, C. M., and H. Tager-Flusberg. 2003. "The Influence of Language on Theory of Mind: A Training Study." *Developmental Science* 6 (3): 346–359.

Harnad, S. 1990. "The symbol grounding problem." *Physica D: Nonlinear Phenomena* 42 (1-3): 335–346.

Harnad, S. 1991. "Other bodies, other minds: A machine incarnation of an old philosophical problem." *Minds and Machines* 1:43–54.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition,* 770–778.

Hertzmann, A. 2024. *Why Evolution Isn't Just Optimization.* Accessed: 2025-04-18. https://aaronhertzmann.com/2024/06/19/why-evolution-isnt-optimization.html.

Hinton, G. 2022. "The Forward-Forward Algorithm: Some Preliminary Investigations." *arXiv:2212.13345.*

Hunsberger, E. 2017. "Spiking Deep Neural Networks: Engineered and Biological Approaches to Object Recognition." PhD diss., University of Waterloo.

Huyck, C. R. 2020. "A neural cognitive architecture." *Cognitive Systems Research* 59:171–178.

Idrees, I., Z. Hasan, S. P. Reiss, and S. Tellex. 2021. "Where were my keys?–Aggregating Spatial-Temporal Instances of Objects for Efficient Retrieval over Long Periods of Time." In *The AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction.*

Jaegle, A., S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, et al. 2022. "Perceiver IO: A General Architecture for Structured Inputs & Outputs." In *The Tenth International Conference on Learning Representations.*

Jung, C. 1960. *Synchronicity: An Acausal Connecting Principle.* Translated by R. Hull. Princeton University Press.

Kaelbling, L. P., M. L. Littman, and A. R. Cassandra. 1998. "Planning and acting in partially observable stochastic domains." *Artificial Intelligence* 101 (1-2): 99–134.

Knepper, R. A., S. Tellex, A. Li, N. Roy, and D. Rus. 2015. "Recovering from Failure by Asking for Help." *Autonomous Robots* 39:347–362.

Kolmogorov, A. N. 1965. "Three Approaches to the Quantitative Definition of Information." *Problems of Information Transmission* 1 (1): 3–11.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25,* 1097–1105.

Kross, E. 2021. *Chatter: The Voice in Our Head, Why It Matters, and How to Harness It.* Crown Publishing Group. ISBN: 9780525575238.

Langley, C., B. I. Cirstea, F. Cuzzolin, and B. J. Sahakian. 2022. "Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI: A Review." *Frontiers in Artificial Intelligence* 5.

Lee, M. K., S. Kiesler, and J. Forlizzi. 2010. "Receptionist or information kiosk: how do people talk with a robot?" In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work,* 31–40.

Leshchev, S. V. 2021. "Cross-modal Turing test and embodied cognition: Agency, computing." *Procedia Computer Science* 190 (C): 527–531.

Li, Q., Y. W. Teh, and R. Pascanu. 2025. *NoProp: Training Neural Networks without Back-propagation or Forward-propagation.* ArXiv:2503.24322.

Li, X., Z. Serlin, G. Yang, and C. Belta. 2019. "A formal methods approach to interpretable reinforcement learning for robotic planning." *Science Robotics* 4 (37): eaay6276.

Liu, J. X., Z. Yang, I. Idrees, S. Liang, B. Schornstein, S. Tellex, and A. Shah. 2023. "Grounding Complex Natural Language Commands for Temporal Tasks in Unseen Environments." In *7th Annual Conference on Robot Learning.*

Madani, O., S. Hanks, and A. Condon. 1999. "On the Undecidability of Probabilistic Planning and Infinite-Horizon Partially Observable Markov Decision Problems." In *Proceedings of the 16th AAAI Conference on Artificial Intelligence,* 541–548.

Manmadhan, S., and B. C. Kovoor. 2020. "Visual question answering: a state-of-the-art review." *Artificial Intelligence Review* 53 (8): 5705–5745.

Mildenhall, B., P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. 2020. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." In *Proceedings of the 2020 European Conference on Computer Vision,* 405–421.

Milford, M., G. Wyeth, and D. Prasser. 2004. "RatSLAM: A hippocampal model for simultaneous localization and mapping." In *Proceedings of the 2014 IEEE International Conference on Robotics and Automation,* 403–408.

Millidge, B. 2022. *The Scale of the Brain vs Machine Learning.* Accessed: 2025-04-19. https://www.beren.io/2022-08-06-The-scale-of-the-brain-vs-machine-learning/.

Mirzadeh, I., K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. 2024. *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models.* arXiv: 2410.05229.

Nishio, S., H. Ishiguro, and N. Hagita. 2007. "Geminoid: Teleoperated Android of an Existing Person." Chap. 20 in *Humanoid Robots: New Developments,* edited by A. C. de Pina Filho, 343–352.

Pardo, M., K. Fristrup, D. Lolchuragi, J. Poole, P. Granli, C. Moss, I. Douglas-Hamilton, and G. Wittemyer. 2024. "African elephants address one another with individually specific calls." *Nature Ecology & Evolution* 8:1353–1364.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference.* 2nd ed. Cambridge, UK: Cambridge University Press.

Qadri, M., and M. Kaess. 2023. "Learning Observation Models with Incremental Non-Differentiable Graph Optimizers in the Loop for Robotics State Estimation." In *ICML 2023 Workshop on Differentiable Almost Everything.*

Raman, S. S., Z. Yang, B. Hedegaard, S. Tellex, D. Paulius, and N. Shah. 2024. "Skill-Wrapper: Skill Abstraction in the Era of Foundation Models." In *2024 CoRL Workshop on Learning Effective Abstractions for Planning.*

Ransom, J. I., and B. S. Cade. 2009. *Quantifying Equid Behavior—A Research Ethogram for Free-Roaming Feral Horses.* Reston, VA: US Department of the Interior, US Geological Survey.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. "Learning Internal Representations by Error Propagation." Chap. 8 in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations,* 318–362. MIT Press.

Saha, P. 2023. *Crows Understand Caw-se and Effect.* Accessed: 2025-04-19. https://www.audubon.org/news/crows-understand-caw-se-and-effect.

Schweizer, P. 1998. "The truly total Turing test." *Minds and Machines* 8:263–272.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423.

Silver, D., S. Singh, D. Precup, and R. S. Sutton. 2021. "Reward is enough." *Artificial Intelligence* 299:103535.

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. 2023. "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models." *Transactions on Machine Learning Research* (May).

Such, F. P., V. Madhavan, E. Conti, J. Lehman, K. O. Stanley, and J. Clune. 2017. "Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning." *arXiv:1712.06567.*

Sun, Q., E. Cetin, and Y. Tang. 2025. "Transformer-Squared: Self-adaptive LLMs." In *Proceedings of the Thirteenth International Conference on Learning Representations.*

Sutton, R. S. 2019. "The Bitter Lesson." *Incomplete Ideas (Personal Blog),* http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Tellex, S., N. Gopalan, H. Kress-Gazit, and C. Matuszek. 2020. "Robots That Use Language." *Annual Review of Control, Robotics, and Autonomous Systems* 3:25–55.

Tellex, S., R. Knepper, A. Li, D. Rus, and N. Roy. 2014. "Asking for Help Using Inverse Semantics." In *Robotics: Science and Systems X.*

Templeton, C. N., E. Greene, and K. Davis. 2005. "Allometry of alarm calls: black-capped chickadees encode information about predator size." *Science* 308 (5730): 1934–1937.

"Theory of Mind Workshop at ICML 2024." 2024. In *Proceedings of the 41st International Conference on Machine Learning: Workshop on Theory of Mind.* Accessed: 2024-12-24. Vienna, Austria. https://tomworkshop.github.io/.

Thrun, S. 2002. "Probabilistic robotics." *Communications of the ACM* 45 (3): 52–57.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30,* 5998–6008.

Vemula, A., and J. A. Bagnell. 2020. "TRON: A Fast Solver for Trajectory Optimization with Non-Smooth Cost Functions." In *Proceedings of the 59th IEEE Conference on Decision and Control,* 4157–4163.

Vijayaraghavan, P., J. F. Queisser, S. V. Flores, and J. Tani. 2024. "Development of Compositionality and Generalization through Interactive Learning of Language and Action of Robots." *arXiv:2403.19995.*

Villiers, J. de. 2007. "The Interface of Language and Theory of Mind." *Lingua* 117 (11): 1858–1878.

Vogel, A., M. Bodoia, C. Potts, and D. Jurafsky. 2013. "Emergence of Gricean Maxims from Multi-Agent Decision Theory." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1072–1081.

Walter, M. R., S. Hemachandra, B. Homberg, S. Tellex, and S. Teller. 2013. "Learning Semantic Maps from Natural Language Descriptions." In *Robotics: Science and Systems IX.*

Wan, E. W., and R. P. Chen. 2021. "Anthropomorphism and object attachment." *Current Opinion in Psychology* 39:88–93.

Wang, D., S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, and R. Platt. 2024. "Equivariant Diffusion Policy." In *8th Annual Conference on Robot Learning.*

White, C., M. Safari, R. Sukthanker, B. Ru, T. Elsken, A. Zela, D. Dey, and F. Hutter. 2023. "Neural Architecture Search: Insights from 1000 Papers." *arXiv:2301.08727.*

Zador, A., S. Escola, B. Richards, B. Ölveczky, Y. Bengio, K. Boahen, M. Botvinick, D. Chklovskii, A. Churchland, C. Clopath, et al. 2023. "Catalyzing next-generation Artificial Intelligence through NeuroAI." *Nature Communications* 14:1597.

Zhang, C., S. Bengio, M. Hardt, M. C. Mozer, and Y. Singer. 2020. "Identity Crisis: Memorization and Generalization under Extreme Overparameterization." In *The Eighth International Conference on Learning Representations.*

Zhao, T. Z., V. Kumar, S. Levine, and C. Finn. 2023. "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware." In *Robotics: Science and Systems XIX.*

Zheng, K., D. Bayazit, R. Mathew, E. Pavlick, and S. Tellex. 2021. "Spatial Language Understanding for Object Search in Partially Observed Cityscale Environments." In *Proceedings of the 2021 IEEE International Conference on Robot and Human Interactive Communication,* 315–322.

Zhong, Y., J. Xiao, W. Ji, Y. Li, W. Deng, and T.-S. Chua. 2022. "Video Question Answering: Datasets, Algorithms and Challenges." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 6439–6455.

Zhu, N., and Z. Wang. 2020. "The paradox of sarcasm: Theory of mind and sarcasm use in adults." *Personality and Individual Differences* 163:110035.