Mental Simulation: Paths to Predictive Intelligence

Cody Canning

December 17, 2012

Abstract

Mental simulation, the tight coupling of perception, action and cognition, has been thrust upon researchers in embodied artificial intelligence as the key to intelligent interaction and behavior in robots. This paper outlines the default representational approach to mental simulation before contrasting it with the burgeoning theory of hierarchical predictive coding. The latter approach promises to deliver a robust solution to robots with practical intelligence.

What is at the heart of interaction?

Cognitive scientists have long posited theories of human intelligence. One of the major goals in cognitive science and artificial intelligence (AI) is to create agents that can interact naturally and effectively with humans and other agents. These robotic agents would, ideally, be capable of interacting in

natural language, picking up on social cues, participating in mixed-initiative team tasks, and reasoning about the world. One major step toward developing these capabilities in artificial agents is the identification of a general interactive framework. Consider *mental simulation* as being at the heart of interaction.

What is mental simulation? The concept intended here is not intuitive. Consider mental simulation as the unification of perception, action and decision-making with a central *predictive* component. Crucially, it allows an agent to explore the consequences of potential actions without actually enacting them in the real world. However, it is *not* strictly a simulation of the kind wherein a virtual agent acts on a virtual environment – like a computer simulation or video game. Today in cutting-edge robotics, both demand for and expectations of this kind of mental simulation are high (as demonstrated on page 3). The present paper argues that mental simulation, as a vehicle for embodied artificial intelligence, has been narrowly pursued by cognitive scientists. The reigning approach to mental simulation is a kind of virtualized, representational schematic, limited in its capacity to endow robots with practical intelligence. The present paper explores an alternative biologically-inspired approach that consists in the tight integration of perception, action and cognition. Before investigating these two different approaches to mental simulation, it is necessary to underscore the capabilities engendered by expectation and meta-expectation within an agent.

Why is prediction such an important component of mental simulation? Prediction enables an agent to form, act on, and reason over expectations of the consequences of its behavior in the world. Consider an agent with the following goal: *pick up the red cup on the table*. Accomplishing this goal at least requires identification of the table and the red cup, movement to the table, grasping the red cup, and lifting it off the table. The ability to pick up the cup in large part depends on predicting the effects of actuation, that is, knowing the right way to move the arm toward the goal. It also involves predicting the behavior of other "agents" in the world, like the table and the cup. In this example, knowing the types of behaviors that the table and the cup will produce, i.e., that they will remain stationary unless set in motion, is important to accomplishing the task. Explicitly computing these behaviors with complex simulators in order to form predictions is one approach to mental simulation.

Evidence from research in neuroscience, specifically the structure of the neocortex, has suggested the manner by which humans couple prediction with behavior. Hierarchical prediction mechanisms are built into the neocortex. Unconsciously, humans form and circulate predictions, and meta-predictions, at every moment. Consider a human faced with the same task described above. He or she will accomplish the task in part by expecting the arm to move in a certain way, the cup to exhibit certain tactile, "graspable" properties, the table *not* to move, etc. These taken-for-granted expectations come in the form of predictions about patterns of sensory experience and extend

to various modalities. For example, vision expects certain patterns and regularities to arrive at the visual cortex as the human moves toward the table and grasps the red cup. These patterns exist within the environment and are captured by perception. Relatedly, touch expects the cup to feel a certain way based on what vision says about the cup and how that relates to prior experiences with similar cups. Moreover, the auditory system has its own set of expectations linked to those established by vision and touch. The capacity to form expectations and, importantly, to recognize the violation of expectations, encompasses predictive intelligence.

Forming predictions about the world at various levels of detail might help alleviate what is known as "the frame problem" in robotics. The term, coined by Patrick Hayes and John McCarthy in 1969, refers to the difficulty of separating out what is relevant to the current moment or task from what is irrelevant. The deluge of irrelevant information constantly exacerbating a problem's computational complexity has stymied the development of general, goal-oriented behavior in robots. Learning multi-modal predictions and meta-predictions about what information is relevant to a certain kind of behavior would be helpful in this context. Concentrating principally on *violated* and *goal-oriented* expectations within a scenario might tunnel a robot's vision sufficiently to make the problem computationally tractable in real time. The undercurrent of expectations and prediction schemas is central to realistic interaction with the world.

The move toward mental simulation

Approaches to intelligent behavior in robotics and AI spanned many theories developed over the latter half of the 20th and the beginning of the 21st century. Good Old-Fasioned Artificial Intelligence (GOFAI – a term coined by John Haugeland) was the original, dominant paradigm of research in AI. GOFAI leveraged the symbolic encoding of facts in order to generate informed behavior. Euclidean logic operated over these units of axiomized world knowledge, creating a systematized, rigid, computational approach to action and perception. "Shakey", a robot developed at the Stanford Research Institute in the late 1960s, is one of the principal archetypes of the GOFAI approach. Using computational formalisms Shakey was able to reason about its own actions and break down commands into step-by-step instructions. Through heuristic search and deduction Shakey navigated and performed actions within a limited environment of blocks and pyramids (Nilsson, 1984).

While Shakey was a monumental achievement in AI, it cannot be said that it was capable of true general-purpose, adaptive intelligence. Shakey did, however, demonstrate the incredible difficulty of the problem. More contemporary approaches to general-purpose intelligence have settled on mental simulation as a potential means. Robots that are capable of very realistically simulating their actions before they perform them are in high demand. This demand follows from the tantalizing idea that performing actions in simulation before executing them will drastically reduce error, as the robot

will have a reasonable idea of what the consequences of its actions will be. In other words, the robot can fail in simulation rather than the real world, reducing the number of risks taken. Such capabilities are requested in a very recent call for proposals from the Office of Naval Research (ONR), one of the major drivers of modern research in robotics. An exerpt from the document follows:

> **Thrust 3: Mental Simulation as a Unifying Framework for Perception, Cognition and Control in Autonomous Systems and Dexterous Robots**
>
> *The goal of this effort is to produce a computational treatment of perception, high-level cognition and motor control for autonomous systems inspired by recent findings across the cognitive and neural sciences. The core idea defining the approach is the notion of a mental simulation. Simulations allow us to consider hypothetical, counterfactual, past, and future situations: all of which differ from reality in important ways. It has been suggested that mental simulation lies at the heart of perception, cognition and action. Even the simplest motor movements are thought to be accompanied by motor simulations that predict expected consequences and compare those predictions against actual outcomes. This simulate-and-predict scheme has also been implicated in the ongoing construction and organization of episodic knowledge and in planning.*

Clearly, by this account mental simulation is absolutely crucial; it is the nexus of cognition, perception and action. How, then, should cognitive scientists engineer mental simulation? A common approach is to replicate in complete detail the geometry and physics of the agent and environment – what we commonly think of when we hear the word, *simulation*. We can consider this

approach in the context of the example from before: a robot with the goal to pick up the red cup on the table. The robot first computes the potential paths its arm could take to the cup and simulates them. It chooses a path according to heuristics provided by the programmer: probably one that looks smooth, moves the arm in a non-roundabout manner, and graps the cup.

This type of mental simulation will be termed "virtualization". While virtualizing an agent and its environment calls to mind three-dimensional simulations with computer graphics, it must also include the distillation of that idea into the complex interaction of a physics engine, hard-coded geometrics, and various algorithms for planning, motion-control, object-detection, etc. Virtualization in this case should not be thought to require graphical rendering; robots programmed to do mental simulation via virtualization compute actions based on explicit data structures and representations jointly provided by the programmer and built from perception.

The strength of virtualization is clear: robots can compute a set of possible action-sequences and choose the one that, in simulation, produces the best results. There are various types of heuristic search that govern this decision. For example, one algorithm "thinks" about the possible approaches until it reaches a time threshold at which point it returns the best answer found so far. While this approach seems promising, it still requires a systematic method for setting these thresholds in order to avoid the problem of the robot indefintiely scratching its head as it ponders potential solutions. Moreover,

manually programming the physics engine and the precise dimensions and properties of the robot is tedious and error-prone. And while physics engines continue to improve, they are not perfect. Also consider that virtualizing the rest of the environment, object by object, in real time, requires vast computational power and very accurate sensors. The hard-coding of the robot's chassis and effectors, in particular, must also be adaptive: the approach must be robust to damage to the robot's chassis or effectors and not require manual reprogramming.

In one sense, virtualized mental simulation appears to be GOFAI in disguise. Good old-fashioned artificial intelligence was dominated by heuristic searches through large sets of encoded facts. Systems of symbolic logic operated over these bits of syntax in order to derive semantics. Is not virtualized mental simulation just a relic of the GOFAI approach? It uses heuristic searches like "rapidly exploring random trees" (RRTs) and similar path-planning algorithms in order to derive goal-behaviors from complex encodings. These encodings amount to knowledge about, e.g., the degrees of freedom, dimensions, and weight of a robotic arm. Only by searching through high-dimensional space does the robot arrive at a "reasonable" solution to the problem; that is, a path to the goal state. Segmenting virtualization in this way indeed supports the analogy to the GOFAI approach.

What about other approaches to mental simulation? Is it possible to consider hypothetical, counterfactual, past, and future situations without explicitly

virtualizing the agent and the environment, without integrating advanced physics engines, without re-presenting the world to the robot? Approaches exist that derive inspiration from human and animal biology, specifically their neural structures. These frameworks have been developed out of research into the mammalian neocortex, evolution, and probability theory. Let us consider these kinds of approaches in order to understand whether they fit into the same picture of mental simulation as virtualization.

Action-oriented predicitve processing

In a recent paper, Andy Clark, captures all the essential aspects of another attempt at predicitve, general intelligence: hierarchical predictive processing. He generalizes this model to include action and dubs it action-oriented predicitve processing. On the hierarchical predictive processing approach the brain can be seen as a device capable of making rapid predictions over all domains relevant to human intelligence. It does so through a hierarchy of generative models that account for incoming sensory signals by matching them to top-down predictions. That is, there are high level "guesses" concerning the content of the information coming from the eyes, ears, etc. How well the content of the prediction matches the content of the sensory signal determines the severity of the error signal that propagates up the hierarchy. These guesses can also abstract over multiple modalities as hierarchies with increasingly stable representations of sensory information.

Why is this approach attractive? It offers a "deeply unified account of perception, cognition, and action", according to Clark. It seems that action-oriented predictive processing fits the bill as far as crucial aspects of embodied intelligence are concerned – recall that the ONR call for proposals, quoted above, specified mental simulation as "at the heart of perception, cognition and action" (Clark, 2012). The following paragraphs explore action-oriented predictive processing and investigate how it encapsulates mental simulation.

Clark references four recent works of major importance to this theory: Friston, Daunizeau et al. (2009), Friston (2010), Brown et al. (2011), and Hawkins and Blakeslee (2004). Their works together drive action-oriented predicitive processing. A concoction of their accounts depicts perception as a process of "canceling out" incoming sensory signals. It is a process of matching a cascade of top-down predictions (potentially Bayesian posteriors) to the incoming sensory signal. These predictions are formed from prior experience and pulled from a storage container of higher-level patterns. The difference between the top-down prediction and the bottom-up signal is called the error. Intelligent systems like these reduce the error signal as much as possible so that their expectations better and better match reality. One's predictions, then, precede sensation – in fact, they *determine* sensation. As a prediction unfolds over time error signals propagate up the hierarchy and drive behavior toward fulfilling the prediction. Thus, thinking, predicting, and acting are all inherently part of unfolding sequences working their ways down the hierarchy.

Since prediction and behavior are working hand in hand, it follows that the agent acts to minimize sensory prediction errors. That is, the agent literally moves its sensors in selective ways that reduce the error signal (Friston (2009), Friston, Daunizeau et al. (2011)). This piece of the action-oriented predictive processing perspective is called "active perception". Alva Noė argues a similar theory, which he calls "enactive perception". As an agent moves through the surrounding environment it somehow recognizes objects in its visual field as having temporal and spatial continuity. Navigation takes for granted one's implicit knowledge of the way the world changes as a result of one's own movement. Noė's posits that an agent derives semantics from the environment by using information from, e.g., its visual and tactile sensors in combination with knowledge of its own motor commands. Humans are not presented with the visual world all at once; we rely on movement of the head and eyes. The implicit knowledge of how the content of one's vision will change as a result of these movements is considered *sensorimotor* knowledge and is the basis of perceptual content.

On Noė's account, visual potentials specify the appearance of an object from unique perspectives. Consider an opaque, geometric cube. Cubes have six sides, twelve edges, and eight vertices, and one can never view them all simultaneously. In fact, the most one can see from a single vantage point is three of the six sides. How do we acknowledge the other sides as being part and parcel to the identity of the cube? While moving around in the environment and observing how an object's presence in the visual field

changes, it becomes clear that predictable, learnable patterns exist within these changes. He writes, "To experience the figure as a cube, on the basis of how it looks, is to understand how its look changes as you move" (Noė, 2005). He characterizes the mental consolidation of learned visual potentials as the sensorimotor profile of an object. Despite the complexity of spatial-temporal relations, human visual perception is sophisticated enough to make sense of them. Enactive perception is driven by invariant representations, or stable patterns, that exist in the environment and encodes them at various levels of the perception-action-cognition hierachy.

While action-oriented predictive processing and similar theories are certainly promising in their theoretical underpinnings, they suffer from potential backlash of the hypothetical form: "Our principal goal is not to understand the human brain, it is to create robots with practical intelligence." The account presented in the present paper hopes to convince the reader that this approach is absolutely useful for creating robots with practical intelligence. These theories are also stymied by the fact that they are very opaque: the structures of the these hierarchical predictive networks are non-intuitive in most cases and quite complex at first blush. Furthermore, the approach relies on training embodied AIs. That is, these AIs do not come pre-supplied with much about the world; they have to learn their expectations through interaction, whereas the virtualization approach does provide the agent with a pre-existing body of knowledge.

Action-oriented predictive processing fits into the model of mental simulation as a strategy for predictive, robust behavior across tasks. While it is not "simulation" in the sense of virtualization – there is no agent representation acting on a representation of the world – it does consider past and future situations as they relate to reality. Mental simulation packaged in this way is not at the heart of perception, cognition and action, it *is* perception, cognition, and action. It encompasses the predictive mechanisms that drive behavior.

The way forward

Are the two approaches to mental simulation presented in this paper incompatible? The approach I have termed "virtualization" differs from action-oriented predictive processing in that, crucially, the former involves encoded representations of both the agent and its environment that operate within the agent's logic board. These representations fundamentally under-represent the real agent and environment. As of now, at least, these simulations are in no way comprehensive. Robots indeed run the risk of succeeding in simulation and failing in the real world due to the representations' error in matching the real world. Even dynamic simulators that adapt their representations based on sensory signals from the robot are burdened by the "drift" of their encodings from reality. Moreover, they are limited to the agent's sensory scope: anything that the agent does not perceive in the environment is not

represented, and therefore does not exist in simulation.

There is potential for approaches to straddle the boundary between virtualization and action-oriented predictive processing. These systems might encode information from representational sensory simulators in a multi-modal hierarchy of abstractions. While this kind of system would still suffer from perceptual drift, it would bootstrap these simulators to achieve behavioral reasoning over different modal domains. It could form stable expectations about the output of these simulators over time. Bayesian networks might be considered an example of this approach as they reason probabilistically, at a high level, leaving the implementation-level open. This sort of hybrid approach is certainly a step in the direction of integrating prediction into frameworks for embodied AI.

In conclusion, research in integrated robot systems would benefit from the move toward action-oriented predictive processing. Multimodal hierarchies that are temporally sensitive and non-representational (in a sense) leverage the world as its own best representation and promise to deliver more general task-agnostic intelligence. These abstraction-machines capture the invariant properties of the world and use them to drive behavior. With behavior unconstrained by representation at the perceptual level, artificial agents can form realistic, reliable expectations about the world and their role within it.

References

[1] H. Brown, K. Friston, and S. Bestmann. Active inference, attention, and motor preparation. *Frontiers in Psychology*, 2, 2011.

[2] A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci*, 2012.

[3] D. Dennett. Expecting ourselves to expect, 2012.

[4] K. Fitzpatrick, P. Carlson, M. A. Brewer, M. D. Wooldridge, and S-P. Miaou. Design speed, operating speed, and posted speed practices. Technical report, Texas Transportation Institute, College Station, TX, 2003.

[5] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[6] K. J. Friston, J. Daunizeau, and S. J. Kiebel. Reinforcement learning or active inference? *PLoS One*, 4(7), 2009.

[7] J. Hawkins and S. Blakeslee. *On intelligence*. St. Martin's Griffin, 2005.

[8] J. P. Masinick and H. H. Teng. An analysis on the impact of rubbernecking on urban freeway traffic. Technical report, University of Virginia, 2004.

[9] N. J. Nilsson. Shakey the robot. Technical report, DTIC Document, 1984.

[10] A. Noė. *Action in perception.* MIT Press, 2005.

[11] Office of Naval Research. *Special Notice 13-SN-0005*, 2013.

[12] S. C. Tignor and D. Warren. *Driver speed behavior on U.S. streets and highways*, 1991.