Why We Need a Physically Embodied Turing Test and What It Might Look Like

Charles L. Ortiz, Jr.

■ *The Turing test, as originally conceived, focused on language and reasoning; problems of perception and action were conspicuously absent. To serve as a benchmark for motivating and monitoring progress in AI research, this article proposes an extension to that original proposal that incorporates all four of these aspects of intelligence. Some initial suggestions are made regarding how best to structure such a test and how to measure progress. The proposed test also provides an opportunity to bring these four important areas of AI research back into sync after each has regrettably diverged into a fairly independent area of research of its own.*

For Alan Turing, the problem of creating an intelligent machine was to be reduced to the problem of creating a thinking machine (Turing 1950). He observed, however, that such a goal was somewhat ill-defined: how was one to conclude whether or not a machine was thinking (like a human)? So Turing replaced the question with an operational notion of what it meant to think through his now famous Turing test. The details are well known to all of us in AI. One feature of the test worth emphasizing, however, is its direct focus on language and its use: in its most well known form, the human interrogator can communicate but not see the computer and the human subject participating in the test. Hence, in a sense, it has always been tacitly assumed that physical embodiment plays no role in the Turing test. Hence, if the Turing test is to represent the de facto test for intelligence, having a body is not a prerequisite for demonstrating intelligent behavior.[1]

The general acceptance of the Turing test as a sensible measure of achievement in the quest to make computers intelligent has naturally led to an emphasis on equating intelligence with cogitation and communication. But, of course, in AI this has only been part of the story: disembodied thought alone will not get one very far in the world. The enterprise to achieve AI has always equally concerned itself with the problems of perception and action. In the physical world, this means that an agent needs to be able to perform physical actions and understand the physical actions of others.

Also of concern for the field of AI is the problem of how to quantify progress and how to support incremental development; it is, by now, pretty much agreed upon that the Turing test represents a rather weak tool for measuring the level of demonstrable intelligence or thinking associated with a particular subject, be it human or artificial. The passing of the test by a machine would certainly justify one in announcing the arrival of human-level AI, but along the way, it can only provide a rather crude measure. To address this deficiency, variants of the Turing test have been proposed and are being pursued; one notable example is the Winograd Schema Challenge[2] that supports incremental testing and development (Levesque, Davis, and Morgenstern 2012). The Winograd Schema Challenge does not, however, address the physical embodiment concerns that are the subject of this article. Nevertheless, any proposed alternative must bring with it a reasonable set of quantifiable measures of performance.

So, what is it about the Turing test that makes it unsuitable for gauging progress in intelligent perception and action?[3] From the perspective of action, the Turing test can only be used to judge descriptions of actions that one could argue were sufficiently detailed to be, in principle, executable. Consider some simple everyday ascriptions of action: "Little Johnny tied his shoelace," or "LeBron James just hit a layup." If perception is taken completely out of the picture, a purely linguistic description of these types of actions is rather problematic (read: a royal pain): one would have to write down a set of rules or axioms that correctly captured the appropriate class of movement actions and how they were stitched together to produce a particular spatiotemporally bounded high-level movement, in this case, bona fide instances of shoelace tying or basketball layups. A more sensible alternative might involve learning from many examples, along the lines demonstrated by Li and Li (2010). And for that, you need to be able to perceive. It's hard for me to describe a shoe-tying to you if you have never seen one or could never see one.[4]

However, consider now the problem of judging the feasibility of certain actions without perception, such as reported by the statement, "the key will not fit in the lock." Through a process of spatial reasoning, an agent can determine whether certain objects (such as a ball) might fit into certain other objects (such as a suitcase). However, this sort of commonsense reasoning could only help with our example during initial considerations: perhaps to conclude whether a particular key was a candidate for fitting into a particular lock given that it was of a particular type. After all, old antique keys, car keys, and house keys all look different. However, it would still be quite impossible to answer the question, "Will the key fit?" without being able to physically perceive the key and the keyhole, physically manipulating the key, trying to get it into the hole, and turning the key.[5] It's no surprise, then, that the challenges that these sorts of actions raise have received considerable attention in the robotics literature: Matt Mason at CMU categorizes them as paradigmatic examples of "funneling actions" in which other artifacts in the environment are used to guide an action during execution (Mason 2001). Note that from a purely linguistic standpoint, the details of such action types have never figured into the lexical semantics of the corresponding verb. From a commonsense reasoning perspective in AI, their formalization has not been attempted for the reasons already given.[6]

These observations raise the question of whether verbal behavior and reasoning are the major indicators of intelligence, as Descartes and others believed. The lessons learned from AI over the last 50 years should suggest that they do not. Equally challenging and important are problems of perception and action. Perhaps these two problems have historically not received as much attention due to a rather firmly held belief that what separates human from beast is reasoning and language: all animals can see and act, after all: one surely should not ascribe intelligence to a person simply because he or she can, for example, open a door successfully. However, any agent that can perform only one action — opening a door — is certainly not a very interesting creature, as neither is one that can utter only one particular sentence. It is, rather, the ability to choose and compose actions for a very broad variety of situations that distinguishes humans. In fact, humans process a rather impressive repertoire of motor skills that distinguish them from lower primates: highly dexterous, enabling actions as diverse as driving, playing the piano, dancing, playing football, and others. And certainly, from the very inception of AI, problems of planning and acting appeared center stage (McCarthy and Hayes 1969).

Functional Individuation of Objects

The preceding illustrations served to emphasize the difficulty in reasoning and talking about many actions without the ability to perceive them. However, our faculty of visual perception by itself, without the benefit of being able to interact with an object or reason about its behavior, runs up against its own difficulties when it attempts to recognize correctly many classes of objects.

For example, recognizing something as simple as a hinge requires not only that one can perceive it as something that resembles those hinges seen in the past, but also that one can interact with it to conclude that it demonstrates the necessary physical behavior: that is, that it consist of two planes that can rotate around a common axis. Finally, one must also be able to reason about the object in situ. The latter requires that one can reason commonsensically to determine whether it is sufficiently rigid, can be attached to two other objects (such as a door and a

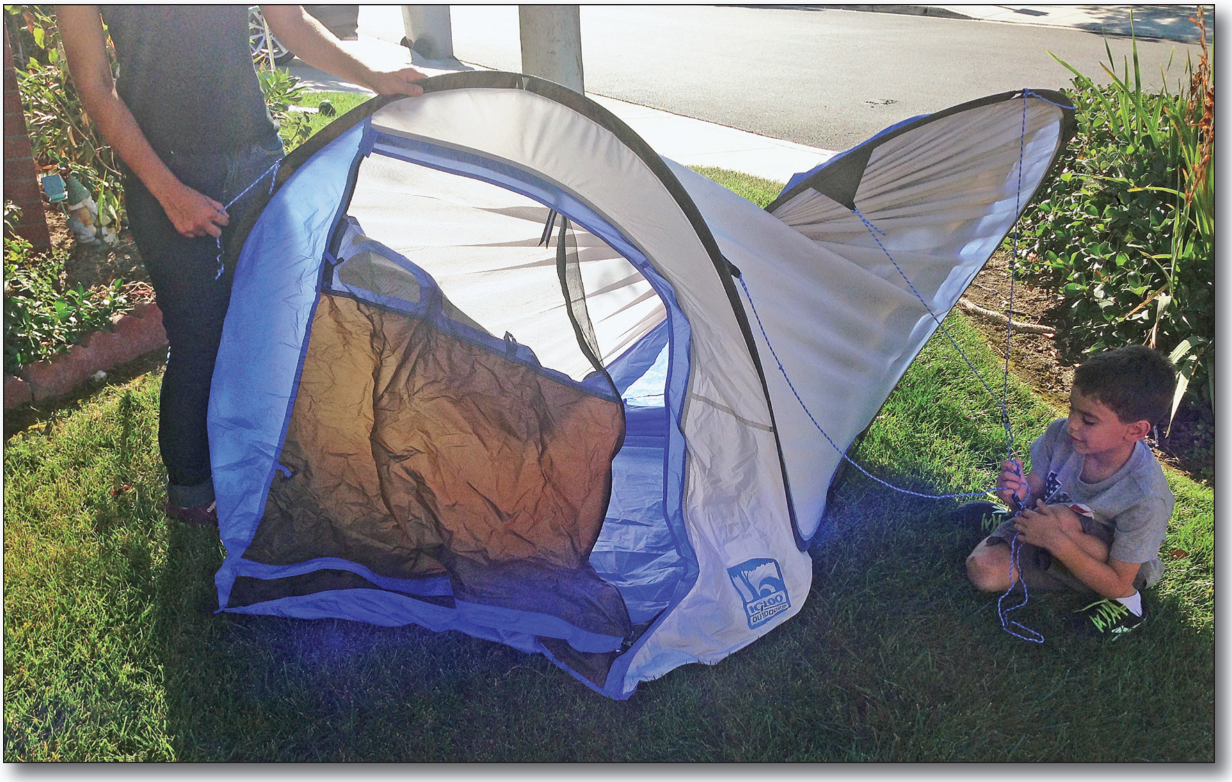

Figure 1. Collaboratively Setting Up a Tent.

A major challenge is to coordinate and describe actions, such as "Hold the pole like this while I attach the rope."

wall), and is also constructed so that it can bear the weight of one or both of those objects. So this very simple example involving the functional individuation of an object requires, by necessity, the integration of perception, action, and commonsense reasoning. The challenge tasks described in the next section nicely highlight the need for such integrated processes.

The Challenge

This leads finally to the question of what would constitute a reasonable physically embodied Turing test that would satisfy the desiderata so far outlined: physical embodiment coupled with reasoning and communication, support for incremental development, and the existence of clear quantitative measures of progress.

In my original description of this particular challenge, I attempted to parallel the original Turing test as much as possible. I imagined a human tester communicating with a partially unseen robot and an unseen human; the human would have access to a physically equivalent but teleoperated pair of robot manipulators. The tester would not be able to see the body of either, only the mechanical arms and video sensors. Significant differences in the appearance of motion between the two could be reduced through stabilizing software to smooth any jerky movements.

The interrogator would interact with the human and robot subject through language, as in the Turing test, and would be able to ask questions or make commands that would lead to the appropriate physical actions. The tester would also be able to demonstrate actions.

However, some of the participants of the workshop at which this idea was first presented[7] observed that particular expertise involving tele-operation might render comparisons difficult. The participants of the workshop agreed that the focus should instead be on defining a set of progressively more challenging problem types. The remainder of this document follows that suggestion.

This challenge will consist of two tracks: The construction track and the exploration track.

The construction track's focus will be on building predefined structures (such as a tent or modular furniture) given a combination of verbal instructions and diagrams or pictures. A collaborative subtrack will extend this to multiple individuals, a human agent and a robotic agent.

The exploration track will be more improvisational in flavor and focus on experiments in building, modifying, and interacting with complex structures in terms of more abstract mental models, possibly acquired through experimentation itself. These struc-

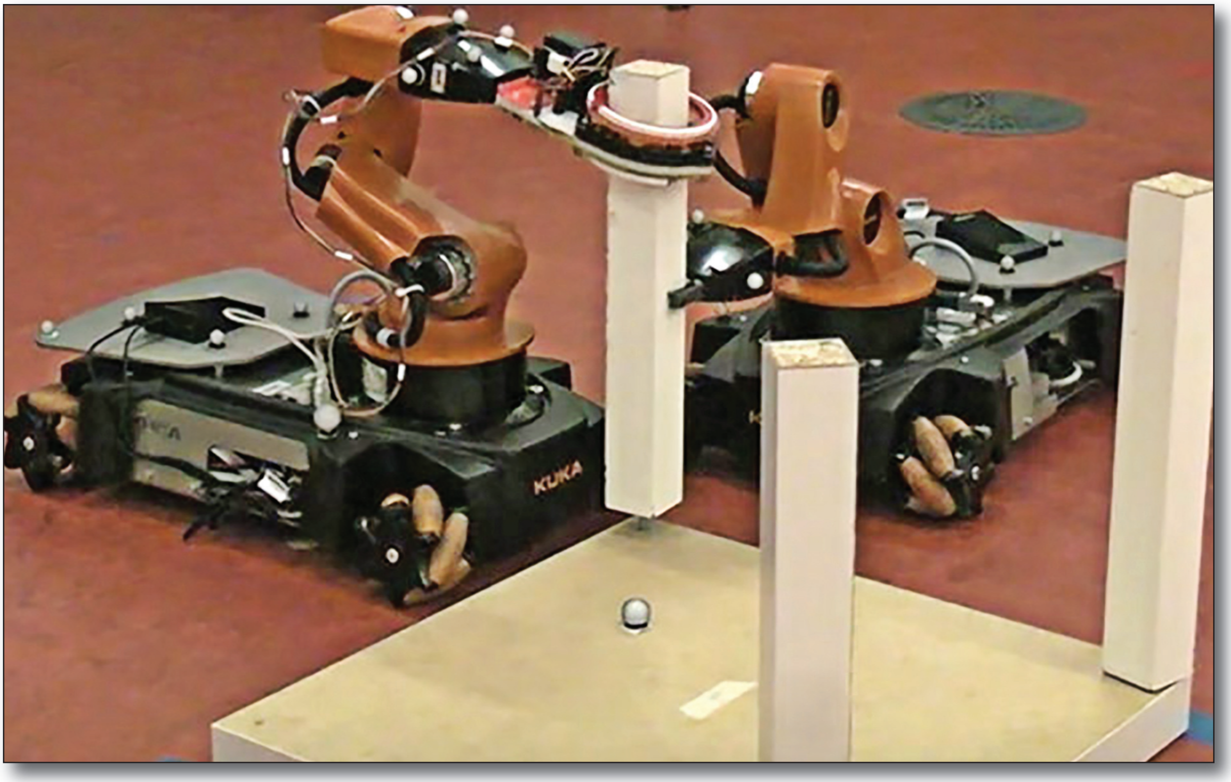

Figure 2. The IkeaBots Developed at the Massachusetts Institute of Technology Can Collaborate on the Construction of Modular Furniture.

tures can be static (for example, as in figure 3) or dynamic (as in figure 6).

Communication through natural language will be an integral part of each track. One of the principal goals of this challenge is to demonstrate grounding of language both during execution of a task and after completion. For example, for both the exploration and the construction tracks, the agents must be able to accept initial instructions, describe and explain what they are doing, accept critique or guidance, and consider hypothetical changes. [8]

The Construction Track

The allowable variability of target structures in the construction track is expected to be less than in the exploration track. The construction task will involve building predefined structures that would be specified through a combination of natural language and pictures. Examples might include an object such as a tent (figure 1) or putting together Ikea-like furniture (figure 2). Often, ancillary information in the form of diagrams or snapshots plays an important role in instructions (see, for example, figure 4). During the task challenge definition phase, the degree to which this complex problem can be limited (or perhaps included as part of another challenge) will be investigated. Crowdsourced sites that contain such instructions might be useful to consult in this respect.[9]

The collaboration task requires that the artificial and human agents exchange information before and during execution to guide the construction task. A teammate might ask for help through statements such as, “Hold the tent pole like this while I tighten the rope”; the system must reason commonsensically about the consequences of the planned action involving the rope-tightening to the requested task as well as how an utterance such as “Hold . . . like this . . .” should be linguistically interpreted and coordinated with the simultaneous visual interpretation.

Rigidity of materials, methods of attachment, and the structural function of elements (that is, that tent poles are meant to hold up the fabric of a tent) will be varied as well as the successful intended functionality of the finished product (for example, a tent should keep water out and also not fall apart when someone enters it). Eventually, time to completion could also be a metric; however, for now, these proposed tasks are of sufficient difficulty that the major concern should simply be success.

The description given here of the construction task places emphasis on robotic manipulation; however, there are nonmanipulation robotic tasks that could be incorporated into the challenge that also involve an integration of perception, reasoning, and action.

Stage	Abilities demonstrated
Construction by one agent	Basic physical, perceptual, and motor skills
Collaboration	Monitoring the activity (perceive progress, identify obstacles), contribute help
Communication	Reference ("hold like <this>"), offer help, explain, question answering ("why did you let go?"), narrate activity as necessary

Table 1. Some Possible Levels of Progression for the Construction Track Challenge Tasks.

Certain capabilities might best be first tested somewhat independently; for example, perception faculties might be tested for by having the agent watch a human perform the task and being able to narrate what it observes.

Examples include finding a set of keys, counting the number of chairs in a room, and delivering a message to some person carrying a suitcase.[10] An organization committee that will be selected for this challenge will investigate the proper mix of such tasks into the final challenge roadmap.

There are many robotic challenges involving manipulation and perception related to this challenge. However, a number of recent existence proofs provide some confidence that such a challenge can be initiated now. The final decisions on subchallenge definition will be made by the organizing committee. As the complexity of these tasks increases, one can imagine their real-world value in robot-assistance tasks as demanding as, say, repairing roads, housing construction, or setting up camp on Mars.

The IkeaBot system (figure 2) developed at MIT is one such existence proof: it demonstrates the collaboration of teams of robots in assembling Ikea furniture during which robots are able to ask automatically for help when needed (Knepper et al. 2013). Other work involving communication and human-robot collaboration coupled with sophisticated laboratory manipulation capabilities has been demonstrated at Carnegie Mellon University, and represents another good starting point (Strabala et al. 2012).

Research in computer vision has made impressive progress lately (Li and Li 2010, Le et al. 2012), enabling the learning and recognition of complex movements and feature-rich objects. It is hoped that this challenge would motivate extensions that would factor in functional considerations into any object-recognition process.

Finally, the organization committee hopes to be able to leverage robotic resources under other activities such as the RoboCup Home Challenge,[11] as much as possible.

The Exploration Track

If you've ever watched a child play with toys such as Lego blocks, you know that the child does not start with a predefined structure in mind. There is a strong element of improvisation and experimentation during a child's interactions, exploring possible structures, adapting mental models (such as that of a house or car), experimenting with sequences of attachment, modifying structures, and so on. Toys help a child groom the mind-body connection, serving as a sort of laboratory for exploring commonsense notions of space, objects, and physics.

For the exploration track, I therefore propose focusing on the physical manipulation of children's toys, such as Lego blocks (figure 3). The main difference between the two tracks is that the exploration track supports experimentation involving the modification of component structures, adjusting designs according to resources available (number of blocks, for example), and exploring states of stability during execution. These are all possible because of the simple modular components that agents would work with. The exploration track would also allow for testing the ability of intelligent agents to build a dynamic system and describe its operation in commonsense terms.

Incremental progression of difficulty would be possible by choosing tasks to roughly reflect levels of child development.

Table 2 summarizes possible levels of progression. The idea is to create scenarios with a pool of physical resources that could support manipulation, commonsense reasoning, and abstraction of structures and objects, vision, and language (for description, explanation, hypothetical reasoning, and narrative).

Figure 3 illustrates a static complex structure while the object in figure 6 involves the interaction of many parts. In the latter case, success in the construction of the object also involves observing and demonstrating that the end functionality is the intended one. In the figure, there is a small crank at the bottom left that results in the turning of a long screw, which lifts metal balls up a column into another part of the assembly in which the balls fall down ramps turning various wheels and gates along

Development stage	Example
1. Simple manipulation	Create a row of blocks; then a wall
2. Construction and abstraction	Connect two walls; then build a "house"; size depends on number of blocks available
3. Modification	Add integrated structures such as a parking garage to house
4. Narrative generation	"This piece is like a hinge that needs to be placed before the wall around it; otherwise it won't fit later" (said while installing the door of a house structure)
5. Explanation	"The tower fell because the base was too narrow"
6. Hypothetical reasoning	"What will happen if you remove this?"

Table 2. A Sequence of Progressively More Sophisticated Skills to Guide the Definition of Subtask Challenges Within the Exploration Track.

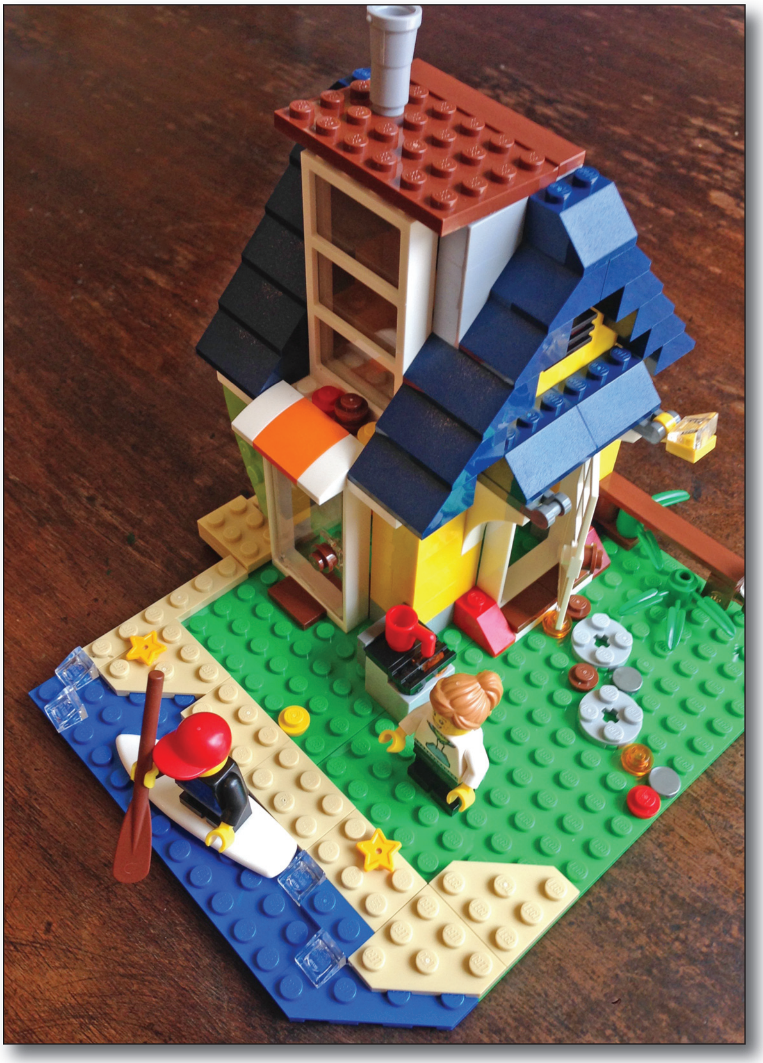

Figure 3. An Abstract Structure of a House Built Using Lego Blocks.

the way. A description along the lines of the last sentence is an example of the sort of explanation that a robot should be able to provide, in which the abstract objects are functionally individuated, in the manner described earlier.

Figure 5 shows another assembly that demonstrates the creation of new objects (such as balls from clay, a continuous substance), operating a machine that creates small balls, fitting clay into syringes, and making lollipop shapes with swirls made from multiple color clays.[12] Tasks involving explaining the operation of such a device, demonstrating its operation, having a particular behavior replicated, and answering questions about processes involved are all beyond the abilities of current AI systems.

Manipulation of Lego blocks and other small toy structures would require robotic manipulators capable of rather fine movements. Such technology exists in robotic surgical systems as well as in less costly components under development by a number of organizations.

Relation to Research in Commonsense Reasoning

The more ambitious exploration track emphasizes the development of systems that can experiment on their own, intervening into the physical operation of a system and modifying the elements and connections of the system to observe the consequences and, in the process, augment their own commonsense knowledge. Rather than having a teacher produce many examples, such self-motivated exploring agents would be able to create alternative scenarios and learn from them on their own. Currently this is all done by hand; for example, if one wants to encode the small bit of knowledge that captures the fact that

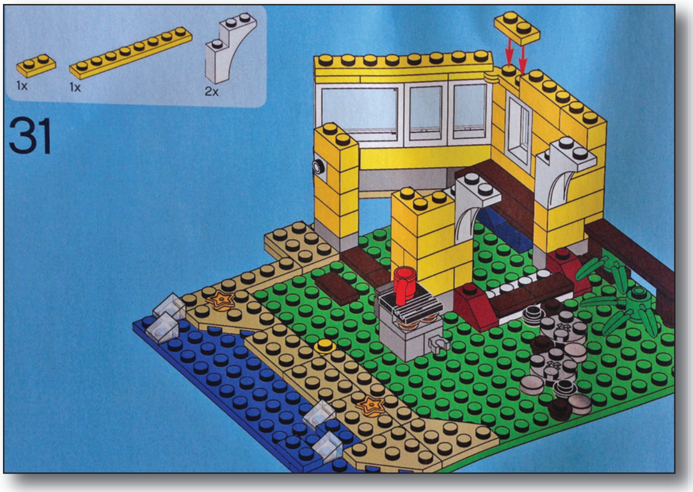

Figure 4. Instructions Often Require Pictures or Diagrams.

The step-by-step instructions are for a Lego-like toy. Notice that certain pieces such as the window or wheels are unrecognizable as such unless they are placed in the correct context of the overall structure.

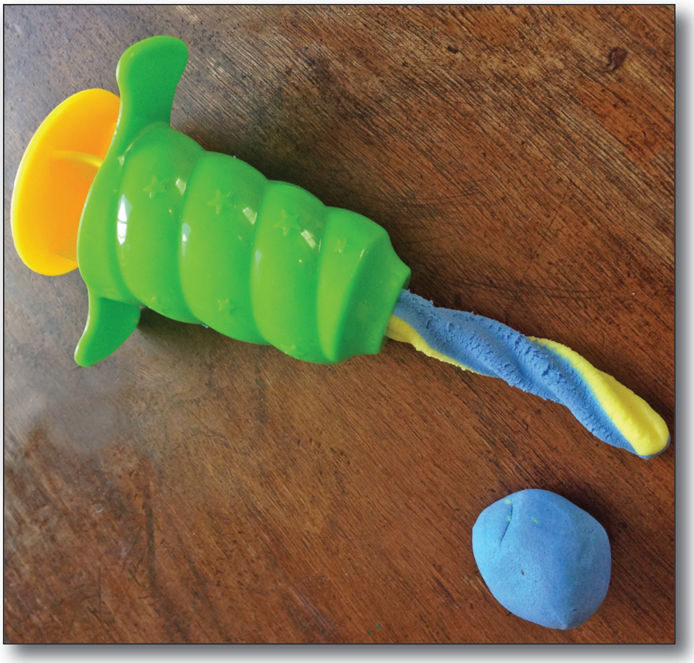

Figure 5. Manipulation and Object Formation with Nonrigid Materials.

not tightening the cap on a soda bottle will cause it to loose its carbonation, one would write down a suitable set of axioms. The problem, of course, is that there is so much of this sort of knowledge.

Research in cognitive science suggests the possibility of the existence of bodies of core commonsense knowledge (Tenenbaum 2015). The exploration track provides a setting for exploring these possibilities. Perhaps within such a laboratory paradigm, the role of traditional commonsense reasoning research would shift to developing general principles, such as models of causation or collaboration. AI systems would then instantiate such principles during self-directed experimentation.

The proposed tests will provide an opportunity to bring four important areas of AI research (language, reasoning, perception, and action) back into sync after each has regrettably diverged into a fairly independent area of research.

Summary

This article was not about the blocks world and it has not argued for the elimination of reasoning from intelligent systems in favor of a stronger perceptual component. This article argued that the Turing test was too weak an instrument for testing all aspects of intelligence and, inspired by the Turing test, proposed an alternative that was argued to be more suitable for motivating and monitoring progress in settings that demand an integrated deployment of perceptual, action, commonsense reasoning, and language faculties. The challenge described in this document differs from other robotic challenges in terms of its integrative aspects. Also unique here is the per-

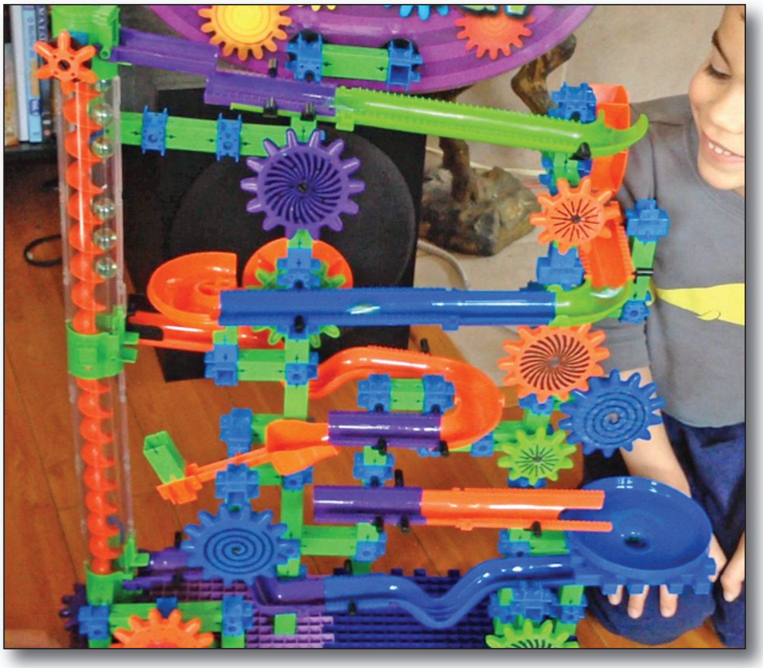

Figure 6. The Exploration Track Will Also Involve Dynamic Toys with Moving Parts and Some Interesting Aggregate Physical Behavior.

The modularity afforded by toys makes this much easier than working with large expensive systems. This picture is a good illustration of the need for functional understanding of elements of a structure. In the picture, the child can turn a crank at the bottom left — a piece that has functional significance — that turns a large red vertical screw that then lifts metal balls up a shaft after which they fall through a series of ramps turning various gears along the way.

spective on agent embodiment as leading to an agent-initiated form of experimentation (the world as a physical laboratory) that can trigger commonsense learning.

The considerable span of time that has elapsed since Turing proposed his famous test should be sufficient for the field of AI to devise more comprehensive tests that stress the abilities of physically embodied intelligent systems to think as well as do.

Notes

[1] One should resist the temptation here of equating intelligence with being smart in the human sense, as in having a high IQ. That has rarely been the case in AI where we have usually been quite happy to try to replicate everyday human behavior. In the remainder of this article, I will use the term *intelligence* in this more restrictive, technical sense.

[2] Winograd Challenge, 2015, commonsensereasoning.org/winograd.html.

[3] I certainly would not deny that a program that passed the Turing test was intelligent. What I am suggesting is that it would not be intelligent in a broad enough set of areas for the many problems of interest to the field of AI. The Turing test was never meant as a necessary test of intelligence, only a sufficient one. The arguments that I am presenting, then, suggest that the Turing test also does not represent a sufficient condition for intelligence, only evidence for intelligence (Shieber 2004).

[4] I take this point to be fairly uncontroversial in AI: a manual with a picture describing some action (such as setting up a tent) is often fairly useless without the pictures.

[5] A similar observation was made in the context of the spatial manipulation of buttons (Davis 2011).

[6] Put most simply, the best that the Turing test could test for is whether a subject would answer correctly to something like, "Suppose I had a key that looked like . . . and a lock that looked like . . . Would it fit?" How on earth is one to find something substantive to substitute (that is, to say) for the ellipses here that would have any relevant consequence for the desired conclusion in the actual physical case?

[7] Beyond the Turing Test: AAAI-15 Workshop WS06. January 25, 2015, Austin, Texas.

[8] One might be concerned that the inclusion of language is overly ambitious. However, without it one would be left with a set of challenge problems that could just as easily be sponsored by the robotics or computer vision communities alone. The inclusion of language makes this proposed challenge more appropriately part of the concerns of general AI.

[9] See, for example, www.wikihow.com/Assemble-a-Tent.

[10] I am grateful to an anonymous reviewer for bringing up this point.

[11] www.robocupathome.org.

[12] See www.youtube.com/watch?v=Cac7Nkki_X0.

References

Davis, E. 2011. Qualitative Spatial Reasoning in Interpreting Narrative. Keynote talk presented at the 2011 Conference on Spatial Information Theory, September 14, Belfast, Maine.

Knepper, R. A.; Layton, T.; Romanishin, J.; and Rus, D. 2013. IkeaBot: An Autonomous multiRobot Coordinated Furniture Assembly System. In *Proceedings of the 2013 IEEE International Conference on Robotics and Automation* (ICRA). Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/ICRA.2013.6630673

Le, Q. V.; Ranzato, M.; Monga, R.; Devin, M.; Chen, K.; Corrado, G. S.; Dean J.; and Ng, A. Y. 2012. Building High-Level Features Using Large Scale Unsupervised Learning. In *Proceedings of the 29th International Conference on Machine Learning*. Madison, WI: Omnipress.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference* (KR2012), 552–561. Palo Alto: AAAI Press.

Li, F. F., and Li, L.-J. 2010. What, Where, and Who? 2010. Telling the Story of an Image by Activity Classification, Scene Recognition, and Object Categorization. In *Computer Vision: Detection, Recognition, and Reconstruction,* Studies in Computational Intelligence Volume 285. Berlin: Springer.

Mason, M. T. 2001. *Mechanics of Robotic Manipulation*. Cambridge, MA: The MIT Press..

McCarthy, J., and Hayes, P. J. Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence 4,* 463–502. Edinburgh, UK: Edinburgh University Press.

Shieber, S., ed. 2004. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. Cambridge, MA: The MIT Press.

Strabala, K.; Lee, M. K.; Dragan, A.; Forlizzi, J.; and Srinivasa, S. 2012. Learning the Communication of Intent Prior to Physical Collaboration. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*. Piscataway, NJ: Institute of Electrical and Electronics Engineers. dx.doi.org/10.1109/roman.2012.6343875

Tennenbaum, Josh. Cognitive Foundations for Commons-Sense Knowledge Representation. Invited talk presented at the AAAI 2015 Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches. Alexandria, VA, 23–25 March.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–460. dx.doi.org/10.1093/mind/LIX.236.433

Charles Ortiz is director of the Laboratory for Artificial Intelligence and Natural Language at the Nuance Communications. Prior to joining Nuance, he was the director of research in collaborative multiagent systems at the AI Center at SRI International. His research interests and contributions are in multiagent systems (collaborative dialog-structured assistants and logic-based BDI theories), knowledge representation and reasoning (causation, counterfactuals, and commonsense reasoning), and robotics (cognitive and team robotics). He is also involved in the organization of the Winograd Schema Challenge with Leora Morgenstern and others. He holds an S.B. in physics from the Massachusetts Institute of Technolgoy and a Ph.D. in computer and information science from the University of Pennsylvania. He was a postdoctoral research fellow at Harvard University and has taught courses at Harvard and the University of California, Berkeley (as an adjunct professor) and has also presented tutorials at many technical conferences such as IJCAI, AAAI, and AAMAS.