Leveraging Developmental Psychology to Evaluate Artificial Intelligence

David S. Moore
Psychology Field Group
Pitzer College
Claremont, CA, USA
dmoore@pitzer.edu

Lisa M. Oakes
Department of Psychology
University of California, Davis
Davis, CA, USA
lmoakes@ucdavis.edu

Victoria L. Romero
National Security and Innovative Solutions
CACI International Inc.
Falls Church, VA, USA
victoria.romero@caci.com

Koleen C. McCrink
Department of Psychology
Barnard College-Columbia University
New York, NY, USA
kmccrink@barnard.edu

***Abstract*—Artificial intelligence (AI) systems do not exhibit human-like common sense. The principles and practices of experimental psychology – specifically, work on infant cognition – can be used to develop and test AIs, providing insight into the building blocks of common sense. Here, we describe how the evaluation team for DARPA's Machine Common Sense program is applying conceptual content, experimental design techniques, and analysis tools used in the field of infant cognitive development to the field of AI evaluation.**

Keywords—AI evaluation, machine common sense, infant cognitive development, experimental design, factorial design

I. Introduction

Over the first years of life, children exhibit 'common sense' reasoning, looking for objects where they have seen them hidden, constructing towers out of blocks carefully balanced one atop another, and using information provided by others to solve problems. As any user of an artificial intelligence (AI) system knows, these ubiquitous programs can also do impressive things, such as recognizing and transcribing some spoken words, identifying the contents of some visual images, and predicting the words a user is likely to type next into their text message. However, most users also know that these AI systems often fail to exhibit human-like common sense, rendering them untrustworthy in critical situations and frustrating in daily use. In 2018, the U.S. Defense Advanced Research Projects Agency (DARPA) funded the Machine Common Sense, or MCS, program (https://www.darpa.mil/program/machine-common-sense), initiated by David Gunning and subsequently led by Matthew Turek and Howard Shrobe, with the goal of establishing "the computing foundations needed to develop machine commonsense services to enable AI applications to understand new situations, monitor the reasonableness of their actions, communicate more effectively with people, and transfer learning to new domains" [1].

One branch of the MCS program uses as its benchmarks the findings from human infants and toddlers. This work is devoted to simulating early-developing, nonverbal common sense—the kind exhibited by infants and toddlers. The approach driving this developmental emphasis has at its roots a theoretical stance that is quite old, if unfulfilled. In his seminal 1950 article on computing machinery and intelligence, Alan Turing wrote "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain" [2, p. 456].

In the MCS program, teams of engineers and developmental scientists are working together to fulfill Turing's vision and create AIs with the nonlinguistic common sense of infants. Some of these teams—the so-called MCS performers—are building these AIs; other teams, including our own, are creating tools to evaluate how the AIs perform. Although the performances of AIs tested in our evaluations are of interest, this paper focuses not on the results of any evaluations or on the architectures or training regimens that produce those results, but on the designs and utility of the evaluation tools themselves.

Unfortunately, although good commonsense reasoning is required for the successful completion of a wide variety of tasks, AI researchers have made relatively little progress in this domain [3]. Davis and Marcus noted in 2015 that while some advances have been made in the areas of taxonomic reasoning, temporal reasoning, and qualitative reasoning, a number of obstacles have prevented improvement in general commonsense reasoning [3]. These obstacles include the fact that common sense frequently requires reasoning that is abductive rather than inductive or deductive, that many situations require a reasoner to make complex inferences about other individuals, and that even when a domain of competence is characterized by a relatively small collection of common examples, there is often a "long tail" of infrequent examples that are likely to confuse artificial—but not natural—agents [3]. As recently as 2019, Mitchell noted that several of these obstacles are still impeding progress on commonsense AI [4]. Finally, scholars have not reached agreement on what constitutes common sense. Nonetheless, one area of general agreement is that an AI system with common sense should be able to generalize previous competences to novel situations that rely on the same underlying ability, with little (if any) re-training.

II. The Role of Evaluation

The field of developmental psychology has the potential to inform AI research in several ways. For example, researchers can try to program developmentally inspired learning capabilities into their AI systems [5, 6]. Likewise,

Research supported by Machine Common Sense (MCS) HR001119S0005 from the Defense Advanced Research Project Agency.

developmental psychologists can work to identify the ideal training data required to instill common sense in AI systems [7] or address theoretical questions about the origins of human cognitive phenotypes [8, 9]. In our work as developmental scientists involved with the MCS project, we have brought a different kind of contribution to the table: evaluation.

Several AI theorists have explicitly discussed the importance of evaluation to the project of developing commonsense AI. When discussing ways to spur progress in the development of artificial common sense, Davis and Marcus noted "There may be no single perfect set of benchmark problems, but as yet there is essentially none at all, nor anything like an agreed-upon evaluation metric; benchmarks and evaluation marks would serve to move the field forward" [3, p. 102]. Hernández-Orallo devoted his 2017 book to the importance of developing ways to effectively evaluate natural and artificial intelligence [10], emphasizing that the field should focus on discrete and unconfounded abilities, and conceptualize with rigor the algorithmic information that dictates the difficulty of each task.

Recently, some researchers have turned their attention to producing evaluation tools that will be of use to developers in the AI community who are working on the types of common sense found in nonverbal organisms. In 2020, Crosby et al. published "The Animal-AI Testbed and Competition," an AI evaluation platform inspired by research on the commonsense cognition of nonhuman animals [11]. The following year, one of the MCS evaluation teams published the "Baby Intuitions Benchmark (BIB)," an AI evaluation tool motivated specifically by research on human infants [12]. In addition, Riochet et al. developed IntPhys 2019, a benchmark that evaluates whether AIs can reliably detect the plausibility of infant-psychology-inspired physics scenes generated from a game engine [13].

Developmental scientists who regularly design assessments for infants and young children are well positioned to contribute to the effort to create evaluations of AI systems, evaluations based on infant abilities. As such, we have been building the Psychometric Intelligent Agent Graphical Environment and Testbed (PIAGET) in an effort to combine and extend previous benchmarks and to provide a more comprehensive set of tools for evaluating nonlinguistic common sense in artificial systems. The fabrication of these tools has been informed by theoretical, methodological, and basic scientific insights gleaned from developmental psychology, and more specifically from the study of infant cognition.

III. The PIAGET Evaluation Tools

We have developed evaluation tools based on classic experiments in developmental psychology that were designed to assess infants' cognition and behavior. Several of our tests take advantage of the fact that many studies of infant cognitive development are non-motoric, requiring only the reliable measurement of the duration of visual fixations. Infants' understanding of how objects behave in the real world can be inferred from increased fixation times, which are taken to signal a violation of expectation (VOE); when infants see an event that appears to violate how the world normally works, they look longer at that event than at expected events [14]. After being presented with video clips showing plausible or implausible events, AI systems can be programmed to respond with a VOE signal that allows for the assessment of the AI's evaluation of the scene's plausibility. In this way, AI systems can be presented with a large number of simplified video representations of events and tested on their ability to reliably tag an event as plausible or implausible (a method developed after that used by Riochet et al. [13]).

However, our evaluations go beyond creating a set of plausible and implausible events. We have designed tasks that illuminate *which aspects* of a scene might render it challenging for an AI system to correctly identify the scene as plausible or implausible. Each task assesses distinct abilities that together give rise to the competence we are evaluating in that task. Accordingly, we have built a collection of tests that allows us to evaluate the presence or absence of specific abilities as drivers of overall performance on the task. Much like how developmental scientists manipulate specific variables when testing infants' understanding or perception of an event, we vary features of scenes we present to AI systems in order to uncover the kinds of information that make common sense reasoning more or less difficult for the AIs. For example, when testing AIs' competence at evaluating collisions, we use scenes in which two objects move in a single depth plane (and, therefore, can actually collide) or in two different, parallel depth planes (and therefore cannot collide). We also use scenes involving trained or untrained objects and scenes in which the collisions happen visibly or when occluded. These manipulations allow us to test for aspects of collision reasoning that interface, respectively, with low-level vision, generalization prowess, and enduring representations of aspects of the scene. This systematic design approach—imported from experimental developmental psychology—has rarely been used in the construction of AI benchmarks. Our approach yields factorial designs (represented as experimental "hypercubes") that capture the contribution of distinct independent variables in tests of particular abilities and allows us to look at how these variables interact with each other to affect performance.

We can rigorously examine patterns in the data by comparing design cells that are distinguished by just one variable, statistically controlling for factors irrelevant to that analysis. In this way, our approach draws on the strengths of psychologists who not only design these types of experiments but have extensive experience with hypothesis testing and the data analysis tools needed to draw inferences about genuine vs. spurious differences. Thus, rather than providing a general benchmark displayed on a leaderboard that reflects overall performance, we evaluate AIs relative to *themselves*, by introducing control trials that can reveal whether a deficit is due to a variable of interest in the question at hand. For example, in some trials there is very little action occurring in the scene. This permits measurement of a particular AI's baseline rate of returning a "plausible" signal in the absence of significant activity, enabling comparisons with plausibility ratings on experimental trials and allowing us to statistically control for baseline plausibility responding.

Another important component of our approach is that we did not generate a multitude of training scenes for the MCS performers (i.e., the teams of engineers and psychologists developing the AI systems). Infants who arrive at developmental psychology laboratories have a good deal of experience with the

concepts being tested in those labs, but that experience has not been accrued in the testing environment. Instead, infants' "training" happens as they encounter and explore the world at large. Psychologists who employ VOE methods can be confident that their experiments reflect infants' generalization of their real-world experiences to the test. It is not currently possible to test AI systems in the kinds of novel environments in which infants are tested, so as a starting point we provided the performers with a scene generator they could use to give their AI systems experience in our simulated "environment;" we also gave the performers the code for a subset of plausible scene types. Although the teams could train their AIs using any data they chose, our scene generator allowed them to produce a limited number and type of scenes closely related to the scenes and environment used for evaluation. For example, the teams could only generate plausible scenes. In addition, they could not generate all types of object movement used at test or all object shapes and sizes. The scene generator therefore let the teams train their AIs on a *sub*sample of the types of plausible scenes presented later, during the test trials. Teams had access to information about how the evaluations were guided by the developmental psychology literature and about the types of abilities we would be evaluating—for example, different aspects of movement, or representing what may happen to an occluded object—but they were not able to train their AIs on everything they might encounter in the testing environment. Although this makes our evaluation quite challenging, we feel it is an important step that will encourage programmers to overcome the central issue of lack of generalizability in AI systems, which interferes with commonsense reasoning in many contexts.

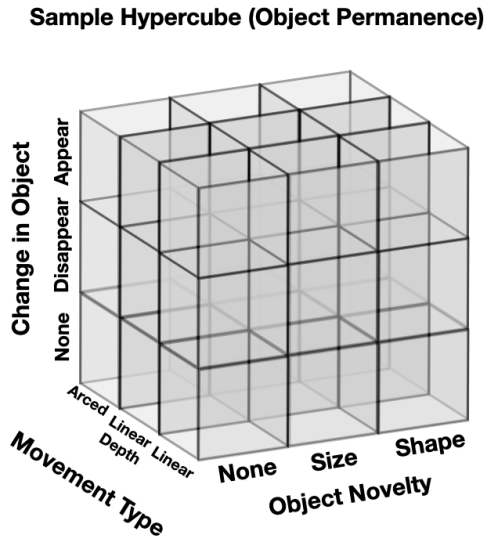

Fig. 1. An example of the experimental hypercube design for the Object Permanence evaluation task.

An example of the kind of hypercube design we use to evaluate AI systems—specifically, the hypercube design we use to evaluate how AI systems respond to object permanence violations—is pictured in Fig. 1. In this hypercube, we systematically manipulate three variables expected to impact object permanence reasoning. Along the x-axis, the novelty of the object in the scene is manipulated. One third of the test scenes include an object that was available for the AIs to train on in our scene generator, one third of the test scenes include an object of a novel (untrained) size but familiar (trained) shape, and one third of the test scenes include an object of a new shape. Along the y-axis, the nature of the object permanence violation is manipulated. Objects either undergo no violation of object permanence, disappear from an expected location, or appear in an unexpected location. Along the z-axis, the type of movement in the scene is systematically manipulated. An object is either tossed into the scene from offscreen, slid across the scene in a single depth plane, or slid across the scene while moving in depth closer to or farther from the point where the AI is viewing the scene.

This hypercube design can disclose single cells where performance is especially strong or weak, and enables experimental control of variables that are irrelevant to the question at hand. Further, it allows predictions as to where performance should be especially strong or weak, based on findings in both developmental psychology and the state of the art in AI engineering. For example, infants can apply their physical reasoning readily to objects they have never seen, along a variety of movement paths, but both of these variables—generalization to unseen images and depth calculations—strain AI systems. Furthermore, adults are more likely to detect the spontaneous appearance of an object than they are to detect the spontaneous disappearance of an object [15], whereas infants are more likely to detect an object's disappearance than appearance [16]. Including the appearance/disappearance variable in this design facilitates exploration of whether developmentally inspired AI systems produce signals that mimic distinct developmental stages of cognition.

Using designs like this, we generate a series of video clips that constitute a single test set that manipulates only the variables of interest; everything else in the test set is held constant, such as the colors of the walls, the texture of the floor, and the size of any occluders present in the room (all of which we consider to be "surface features" of the scenes). Fifty different test sets are generated according to the specifications of the hypercube design, producing many different scenes of the same conceptual type (but differing surface features) that can collectively be used to evaluate how AI systems perform under the conditions specified by the independent variables. Such an approach provides unparalleled experimental control and mitigates concerns that any one non-critical variable (e.g., the texture of the floor) will drive performance.

To date, we have developed eight evaluation tasks that follow the VOE format, each of which was designed to evaluate a particular ability related to a commonsense concept that has been demonstrated in infants. A list of the commonsense concepts to be evaluated in the MCS program can be found in Table 1 in the Appendix. By the time the program ends in 2024, we expect the evaluation arm of MCS to have produced approximately 40 distinct tasks using a variety of evaluation measures. These tasks are organized into three commonsense domains – reasoning about objects, agents, and places. Several of the tasks we have already designed evaluate concepts that bridge multiple domains. The evaluation measure described here (generating a VOE signal) is used primarily in the domain of object understanding; as detailed in the Future Directions section below, we are now moving toward an interactive reward

learning method that continues to emphasize designs that assess componential abilities drawn from the infant cognition literature. Interested individuals can access additional information about the PIAGET evaluations, example video clips, and all of the PIAGET evaluation tools at https://www.machinecommonsense.com. These tools, which include all of the test scenes we have generated to date, are available to the public, and can be used by anyone in the AI community.

IV. Case Study: Spatiotemporal Continuity

In object commonsense reasoning tasks, we draw from classic work in infant psychology that illustrates infants' understanding of how objects exist in the world and interact with other objects in the environment [14, 17, 18, 19]. By the age of 12 months (and in some cases, even earlier), infants are sensitive to features of events such as the facts that objects move in depth, only change motion when contacted, do not occupy the same space as each other, persist when occluded, are subject to the forces of gravity, and have trajectories that can be anticipated along a spatiotemporal continuum [20, 21].

The PIAGET Spatiotemporal Continuity (STC) task examines whether the AI systems developed by the MCS performers can exhibit several of these abilities (e.g., anticipate trajectories and recognize the plausibility of movement in depth and the plausibility of persistence when occluded). This task evaluates if AI systems can detect violations of STC. Three types of movement are presented, in which an object (a) moves linearly across the screen in a single depth plane, or (b) moves linearly in a way that brings it closer or farther from the AI's point of view as it proceeds across the screen, or (c) is tossed into the frame, tracing an arc prior to landing and then moving off the far edge of the screen. In half of the test trials (the plausible trials), the object moves without any violations of STC. In the other half of the test trials (the implausible trials), the object spontaneously disappears briefly mid-trajectory, reappearing a moment later where it would ordinarily be at that time point if spatiotemporal continuity was not violated. In addition, in half of the trials, these events transpire in a room containing no occluders, so the objects are always visible (with the exception of spontaneous disappearances during implausible trials). In the other half of the test trials, two occluders are present for the duration of the scene. In these trials, the occluders are initially lifted to reveal that nothing is behind them. After the occluders are lowered, an object moves across the screen. In plausible trials, the object moves behind the first occluder, reappears (as normal) between the two occluders, moves behind the second occluder, and finally emerges from behind the second occluder to move off screen. Implausible trials with occluders are identical to plausible trials with occluders, except that the object does not appear *between* the two occluders as it moves across the screen (see Fig. 2). Finally, in half of the trials, the moving object is one that the AI system has never encountered before, permitting evaluation of the extent to which an AI system can generalize the STC concept to untrained objects. Note that this particular design involves 4 independent variables, so it cannot be represented as a single cube in 3-dimensional space; hypercubes like this must be depicted using multiple cubes, as seen in Fig. 3, on the following page.

The STC task uses an experimental 3 (Movement type) x 2 (Plausible vs. Implausible) x 2 (Occluded vs. Unoccluded) x 2 (Trained vs. Untrained Object) factorial design. Data analyses from the most recent evaluation of the MCS performers' AIs indicate that on average, performance in plausible STC trials is better when objects are never occluded than when they are occasionally occluded (see Fig. 3). Likewise, on average, these AIs tend to exhibit better performance when objects move in a linear fashion than when they are tossed into the scene and move in an arc. This suggests that these AIs more readily processed action along the horizontal axis of a scene than movement along the vertical axis, in line with work on adult human visuospatial cognition [22]. Some of the evaluated AI systems generalize to novel objects better than other AI systems do. Although how and why the different AI systems produce better or worse performance is beyond the scope of this paper, queries about how the performers build and train their systems will likely yield inferences about what sorts of architectures or training regimens encourage flexible abstraction of concepts that can be generalized. Finally, all of the AIs we evaluated exhibited an implausibility bias; they were more likely to judge a plausible scene as implausible than an implausible scene as plausible. This is a pervasive finding in all of the evaluations of commonsense object understanding that we have done so far, indicating that current AI systems have a propensity to judge object behavior as implausible, which should be accounted for when developing and training these systems. Thus, the PIAGET evaluation tools have yielded data that (a) illuminate conditions in which individual AI systems perform well versus conditions in which these systems find the tasks especially challenging, and (b) reveal practical and theoretical findings that will be of interest to scientists who are studying cognition and development in both AIs and living organisms.

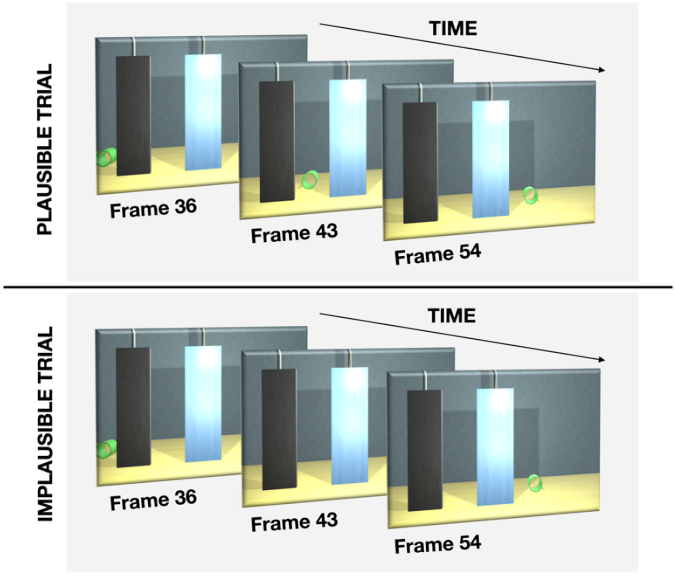

Fig. 2. Sequences of screenshots illustrating plausible and implausible test trials designed to evaluate competence on the PIAGET Spatiotemporal Continuity task. This example shows an object (a hollow green cylinder) moving linearly across the screen in a single depth plane, in a room containing two occluders.

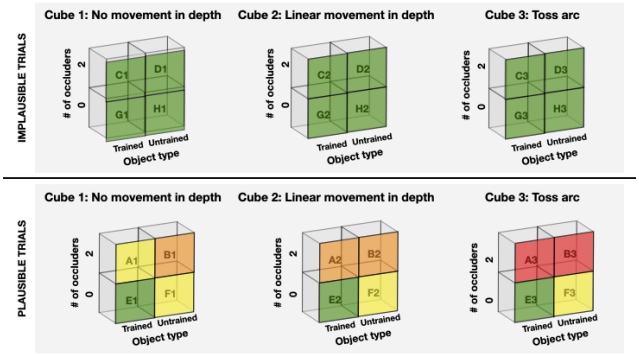

Fig. 3. The Spatiotemporal Continuity hypercube design, colored to show a sample data pattern that helps to visualize an AI's performance. Green shading indicates the AI correctly tagged the scene as plausible or implausible in 75% or more of the trials. Yellow shading indicates 50-75% correct tagging. Orange shading indicates 25-50% correct tagging. Red shading indicates 0-25% correct tagging. Note that strengths and weaknesses are immediately visible in these visualizations and can be followed up with statistical analyses to determine whether any apparent differences are statistically significant. In this example, the AI exhibits (a) excellent performance on implausible scenes (top row of cubes), and poorer performance on plausible scenes (bottom row of cubes), (b) worse performance for non-linear object movement (i.e., the red-shaded cells in the cube at the lower right), (c) better performance for trained (vs. untrained) objects that are not violating spatiotemporal continuity (i.e., more green on the left than right halves of the cubes in the bottom row), and (d) better performance on unoccluded scenes relative to scenes containing two occluders when the objects are not violating spatiotemporal continuity (i.e., more green on the bottom than top halves of the cubes in the bottom row).

V. Future Directions

The PIAGET evaluation tools described in this paper generate data patterns that allow the MCS performers to home in on the reasons their AI systems succeed or fail on specific commonsense tasks. These tools emphasize the need for AIs to possess broad and flexible concepts that can be applied in a variety of contexts. Furthermore, these tools enable testing of AI systems in a way that isolates the AIs' strengths and weaknesses on individual components of the task. By requiring these systems to respond to visual displays in ways that are analogous to how human infants respond to such displays, these tests encourage programmers to develop AI systems that behave in a manner consistent with early developing human common sense.

During the remainder of the MCS program, we will be designing additional tasks that do not rely on the generation of VOE signals and instead require AI systems to more actively engage with their environment. To this end, we have worked with the MCS evaluation engineering team to develop a scene generator called the Interactive Learning Environment (ILE), which allows AI systems to be trained in an environment that aligns with the one they will encounter in subsequent test trials. Rather than programming AI systems to simply generate VOE signals, PIAGET's interactive tasks will require these systems to move around in a simulated environment to obtain a reward, a technique in line with the Animal AI Olympics evaluation developed by Crosby et al. [11]. The performers in the MCS program can train their models in any way they choose, and in some cases they provide their AIs with a great deal of background training. However, we will not be giving them training data; a single template scene will be provided that contains the key engineering elements for each task. The ILE allows performers to situate and train their AIs in our evaluation world by building scenes—"environments"—in which they can implement reward learning to obtain a target and learn about the experimental setting as they do so. Importantly, it also allows evaluators to hold out different types of objects, agents, or places, and use them in evaluation scenes to test whether the AIs have learned a generalized and adaptable concept (or a restricted and inflexible one). Like the other PIAGET evaluation tools, the ILE and relevant information about it is available to the public at https://www.machinecommonsense.com.

Some of these interactive tasks will be two- or three-option forced-choice tasks; others will permit unconstrained movement on the part of the AI systems. In some cases, interactive tasks will be invented that assess the same competences assessed in the passive VOE tasks. For example, whereas our passive object permanence task requires AI systems to *recognize a violation* when an object seen moving behind an occluder is discovered to have disappeared, our interactive object permanence task requires AI systems to *move* to the side of a room that contains a reward that they have seen becoming occluded. In this interactive task, both sides of the room contain identical occluders, and the reward object is not visible when the AI system has the opportunity to start its movement. To succeed at this task, the system must (a) recognize that objects continue to exist when they are occluded, (b) remember where the object became occluded, and (c) move to the side of the room with the hidden object. Evaluating a given competence in both passive and interactive ways will encourage the development of AI systems that are useful in a wide variety of potential applications. This technique of employing two different measures that tap into the same commonsense concept also allows researchers to compare AI cognition to infant cognition with theoretical rigor. For some abilities, infants and toddlers show disparate patterns of performance for conceptually comparable passive and interactive tasks, with strong passive task performance suggesting mastery of a concept but weak motor-based task performance indicating a lack of understanding. For example, when asked to reach to the anticipated location of a falling object, two-year-olds fail to anticipate the effects of an obstacle that impedes the object's movement, so they reach instead for the object in a location at the end of an unobstructed trajectory [23]. In contrast, VOE studies with infants suggest a reliable detection of violations of solidity as falling objects interact with impeding obstacles [24, 25]. Using two different measures of the same commonsense competence will reveal whether the same developmental discordance occurs when AIs are programmed with developmental principles at their core.

Ultimately, we believe it is of great value for AI researchers to work with developmental scientists to collaboratively build systems with commonsense reasoning capabilities. In addition to helping AI researchers understand how human beings develop their common sense as they mature from infancy into adulthood, developmental scientists are proficient at experimental design, stimulus creation, statistical analysis, and data interpretation, enabling the production of evaluation tools appropriate for assessing common sense behavior at two distinct levels. On one hand, these tools can allow for the examination of the component parts that contribute to a higher-order ability; on the other, they can facilitate the study of the broader abilities themselves. We sincerely hope, and expect, that the evaluation

tools we are creating in the MCS program will ultimately be useful to the AI community's efforts to make AI systems with genuine, flexible, and abstract common sense.

ACKNOWLEDGMENTS

The authors thank the engineering team at CACI (in alphabetical order: Rachel Artiss, Jacob Audick, Clark Dorman, Kyle Drumm, Brian Pippin, Thomas Schellenberg, and Dean Weatherby) for their tireless implementation of the evaluation.

REFERENCES

[1] Gunning, D. (2018, October 18). Machine common sense [PowerPoint slides]. Email from Defense Advanced Research Projects Agency (DARPA).

[2] Turing, A. M. (1950). Computing machinery and intelligence. *Mind: A quarterly review of psychology and philosophy, LIX* (236), p. 433–460.

[3] Davis, E., & Marcus, G. (2015, August). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM 58*(9), 92–103. doi:10.1145/2701413

[4] Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans.* New York: Farrar, Straus, and Giroux.

[5] Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences, 21*, 649–665. doi:10.1016/j.tics.2017.05.012

[6] Weng, J. (2020). Conscious Intelligence Requires Developmental Autonomous Programming For General Purposes. *Joint IEEE international conference on development and learning and epigenetic robotics (ICDL-EpiRob)*.

[7] Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology, 8*, 2124. doi: 10.3389/fpsyg.2017.02124

[8] Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness*. Cambridge, MA: MIT Press.

[9] Cangelosi, A., & Schlesinger, M. (2015). *Developmental robotics: From babies to robots.* Cambridge, MA: MIT Press.

[10] Hernandez-Orallo, J. (2017). *The measure of all minds: Evaluating natural and artificial intelligence.* Cambridge, U.K.: Cambridge University Press.

[11] Crosby, M., Beyret, B., Shanahan, M., Hernández-Orallo, J., Cheke, L., & Halina, M. (2020). The Animal-AI Testbed and Competition. *Proceedings of machine learning research.*

[12] Gandhi, K., Stojnic, G., Lake, B. M. and Dillon, M. R. (2021). Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. *Advances in Neural Information Processing Systems (NeurIPS)*, 34 .

[13] Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2020). IntPhys: A framework and benchmark for visual intuitive physics reasoning. arXiv. https://arxiv.org/abs/1803.07616

[14] Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in 5-month-old infants. *Cognition, 20*, 191–208. doi: 10.1016/0010-0277(85)90008-3

[15] Cole, G., Kentridge, R. W., & Heywood, C. A. (2004). Visual salience in the change detection paradigm: The special role of object onset. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 464–477. doi: 10.1037/0096-1523.30.3.464

[16] Wynn, K., & Chiang, W.-C. (1998). Limits to infants' knowledge of objects: The case of magical appearance. *Psychological Science*, *9*, 448–455. doi:10.1111/1467-9280.00084

[17] Needham, A., & Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition, 47*, 121–148.

[18] Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development, 5*, 193–207. doi: 10.1016/0885-2014(90)90026-P.

[19] Slater, A., & Morison, V. (1985). Shape constancy and slant perception at birth. *Perception, 14*, 337–344.

[20] Spelke, E. S. (1990). Principles of object perception. *Cognitive science*, *14*(1), 29–56.

[21] Baillargeon, R. (2002). The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell handbook of childhood cognitive development, 1*, 46–83.

[22] Zwergal, A., Schöberl, F., Xiong, G., Pradhan, C., Covic, A., Werner, P., Trapp, C., Bartenstein, P., la Fougère, C., Jahn, K., Dieterich, M., & Brandt, T. (2016). Anisotropy of human horizontal and vertical navigation in real space: Behavioral and PET correlates. *Cerebral Cortex*, *26*, 4392–4404, https://doi.org/10.1093/cercor/bhv213

[23] Hood, B., Carey, S., & Prasada, S. (2000). Predicting the outcomes of physical events: Two-year-olds fail to reveal knowledge of solidity and support. *Child Development, 71,* 1540–1554.

[24] Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99*(4), 605–632. https://doi.org/10.1037/0033-295X.99.4.605

[25] Baillargeon, R. (1995). Physical reasoning in infancy. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 181–204). The MIT Press.

APPENDIX

TABLE I. LIST OF COMMONSENSE CONCEPTS EVALUATED

Core Domain	*Commonsense Concept*
Objects	Objects have depth >2D & move in 2.5D or 3D space
	Inanimate objects change motion when contacted and only when contacted
	Solid objects do not occupy the same space
	Objects persist, even when occluded
	Object functions can be predicted by their forms (e.g. certain affordances can be gleaned from the shape of an object)
	Solid objects are subject to the forces of gravity
	Sets of objects can contain more, or less, than other sets of objects
	Objects have trajectories that can be anticipated
Agents	Agents have preferences for object-based goals*
	Agents act efficiently*
	Agents affiliate with others who perform prosocial actions*
	Agents are preferred when they act prosocially, recognizing that agents' actions reflect intentions and beliefs informed by what they have observed*
	Agents share a set of common characteristics
	Agents only know what they have seen / experienced
	Agents can provide solutions to problems and convey knowledge
Places	One must continuously update one's own location in relation to features in the environment
	Landmarks can be used to navigate effectively
	One can navigate by encoding the geometry of the environment (distances/directions of stable surfaces)
	Objects can be located in space by a logical process of elimination
	Objects can be tracked over spatial displacement
	One can use the physics of an environment to obtain reward
	One should avoid places in an environment that are dangerous

** Indicates a concept tested by other members of the MCS evaluation team*