Opinion: Naive psychology depends on naive physics

Shari Liu (sliu199@jhu.edu) **Joseph Outa (jouta1@jhu.edu)**
Dept Psychological and Brain Sciences, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218 USA

Seda Karakose-Akbiyik (sea337@g.harvard.edu)
Dept Psychology, Harvard University
33 Kirkland Street, Cambridge, MA 02138 USA

Abstract

People are immaterial minds, housed in material bodies. How do we consider psychological and physical information to make sense of them? Research across the cognitive sciences, including cognitive neuroscience, developmental psychology, and clinical psychology, has accumulated evidence that naive psychology and naive physics are modular, non-interacting systems. We disagree. To the contrary, we use evidence from each of these disciplines, and research from computational cognitive science, to argue that naive psychology and physics constitute parallel and integrated systems in human minds and brains. We end by previewing a research program to investigate this integration.[1]

Keywords: cognitive development; domain-specificity; physical reasoning; social cognition

Interacting with people and objects is fundamental to our everyday life. From observing people's actions (e.g. a person shuffling over a frozen lake), we can gain insights into their goals and mental states (e.g. that they want to cross and feel apprehensive). Similarly, when observing objects (e.g. the opening break in a game of billiards) we can infer hidden physical properties (e.g. the smoothness of the table, the material of the billiard balls). How do our adult minds and brains accomplish these feats, and how do these abilities develop?

Evidence for independence

Across the cognitive sciences, research has shown that our intuitive understanding of the physical and psychological world constitutes two separate, independent, and even modular systems supported by their own distinct representations and computations. In brief: First, human infants have early-emerging and strikingly domain-specific expectations about agents and objects (Spelke, 2022; Carey, 2011), and these domain-specific computations persist into adulthood (Scholl & Tremoulet, 2000; Scholl, Pylyshyn, & Feldman, 2001). Second, separable neural architectures process social and physical information in people of all ages (Jack et al., 2013; Wilcox & Biondi, 2015; Pitcher & Ungerleider, 2021; Richardson et al., 2018). Third, these two abilities are doubly dissociable in Autism Spectrum Disorder and Williams Syndrome (Baron-Cohen, Victoria Scahill, & Lawson, 2001; Kamps et al., 2017). From these observations, scholars proposed that psychological and physical understanding are modular, non-interacting systems supported by their own distinct representations and computations (Spelke, 2022; Carey, 2011; Baron-Cohen, 1998). Some even propose that these two systems are competitive, fundamentally incompatible modes of thought (Bloom, 2005; Jack, 2014).

Evidence for integration

Here, we argue that naive psychology cannot perform its core functions (understanding other people's actions and minds) without access to either the outputs or the intermediate representations of naive physics. Therefore, the two systems cannot be independent; *computations about the mind require input from naive physics in order to return an answer*. Instead, we hypothesize that naive physics and psychology are early-emerging, domain-specific, and integrated systems of cognition.

In support of this proposal, infants have the ontological commitment that agents have the physical properties of objects (Saxe, Tzelnic, & Carey, 2006), flexibly reason about the effects of the physical environment on agents' actions and mental states (Gergely & Csibra, 2003; Luo & Baillargeon, 2007) and use inferred physical information (e.g. effort) to make inferences about agent's minds (Liu et al., 2017). Regions of the frontoparietal cortex that are engaged for physical reasoning also contain representations relevant for understanding the actions of agents: for example, event kinematics (Karakose-Akbiyik, Caramazza, & Wurm, 2023; Karakose-Akbiyik, Sussman, et al., 2023) and physical stability (Pramod et al., 2022). And while individuals with Autism Spectrum Disorder or Williams Syndrome display selective difficulty with social and physical reasoning, both populations can make judgments about intent and perceptual access, retaining some physical and social representations relevant for action understanding (Leslie & Thaiss, 1992; Hamilton, Brindley, & Frith, 2007). Lastly, despite ongoing debate about the computational basis of social cognition, researchers agree that in order to get social cognition off the ground, it is helpful to either build in or learn the sort of rudimentary physical knowledge available to young infants (Malik & Isik, 2023; Rabinowitz et al., 2018; Shu et al., 2021); It is not possible to reverse engineer human social intelligence without learning or building in physical knowledge.

[1] An expanded version of this paper is available at https://osf.io/preprints/psyarxiv/u6xdz.

Theory of Mind: A joint theory of psychology and physics

We submit that starting in infancy and throughout our lives, our knowledge of other people's minds and actions constitutes an intuitive theory (Gopnik & Meltzoff, 1997) encompassing the psychological and physical domains, including information transfer across domains. One class of computational models that expresses these theories are Bayesian probabilistic generative models of Theory of Mind, or BToM (Baker et al., 2017). The most recent implementation of such models (Shu et al., 2021) combines a rational planner (Jara-Ettinger et al., 2016) with a physics engine (Ullman et al., 2017) in order to make sense of the dynamics and causal relations between agents' actions, mental states, and world states. Thus, BToM models express the Full Integration Hypothesis (Fig. 1D): naive physics and naive psychology are intimately connected, with access to both the outputs and intermediate representations of the other system, while still maintaining domain-specific computations for objects and agents. Such models also support the ability to plan interventions on other people's mental states (Ho, Saxe, & Cushman, 2022), which could account for infants' capacities to interact with other minds by requesting and sharing information (Begus & Southgate, 2012), and to help and comfort other people in need (Warneken & Tomasello, 2006).

Empirical Predictions

The first aim of this paper was to articulate the proposal that naive psychology and naive physics cannot be independent of each other (Fig. 1A), and instead are likely integrated domains in the human mind and brain (Fig. 1B-D). Now we (briefly) turn to the second aim of the paper, which is to establish an interdisciplinary research program to investigate the nature of this integration.

The alternative proposals in Fig. 1 make distinct predictions about (i) which physical and social capacities are related to each other, (ii) when a given physical or social reasoning ability should emerge in development, (iii) ease of transfer learning across physical and social domains, (iv) patterns of social and physical difficulty across neurodiversity, and (v) neural computation and organization. We preview two of these predictions here.

Hypotheses granting information change between domains predict that individual differences in people's physical reasoning should selectively predict their performance in psychological judgments requiring physical inputs, like judgments of physical effort and perceptual access, and not other judgments, like identifying an emotional state from a facial expression. The Full Integration Hypothesis, embodied by BToM models, also makes predictions about the timing of neural computations: judgments requiring integration across domains should happen quickly, and as quickly as purely physical or purely psychological computations, because each system can access the information it requires from the other at any point during processing.

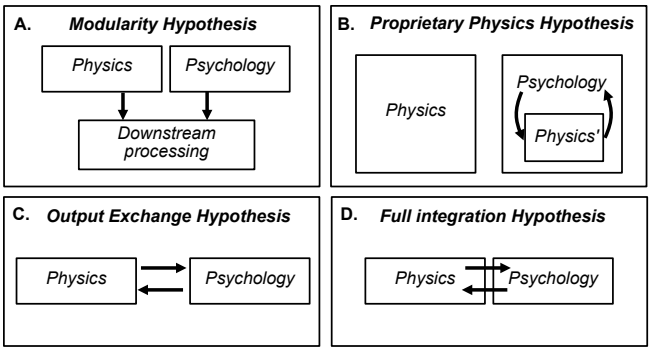

Figure 1: Hypotheses about information flow between naive physics and naive psychology. We argue against the (A) *Modularity Hypothesis*, in which the two systems do not directly interact, producing outputs that are integrated downstream. (B-D) Alternative hypotheses that grant information exchange. (B) *Proprietary Physics Hypothesis*: Naive psychology has its own proprietary physical representations ("physics prime"). (C) *Output Integration Hypothesis*: Both systems can take as input the outputs of the other system. (D) *Full Integration Hypothesis*: Both systems can call each other directly, with access to outputs and intermediate representations.

By contrast, hypotheses that draw a strict boundary between computations from these domains would not predict that particular abilities are connected. Furthermore, such proposals would predict that judgments falling squarely within each domain would occur quickly in domain-specific cortical regions, but judgments that require integration across domains would occur more slowly, since integration can only happen after domain-specific processing. In sum, the proposal that naive psychology depends on naive physics makes testable and falsifiable empirical predictions across the cognitive sciences.

Conclusion

Our mental lives are occupied by other people's minds (their invisible desires, hypotheses, and superstitions). At the same time, our mental states, and the mental states of other people, are often about the physical world: worlds we live in now, and worlds we can imagine. How do our minds and brains connect these domains of information? Here, we grant that humans have domain-specific intuitions about minds and objects, but propose that starting in infancy and throughout our lives, we model other agents as physical bodies with minds. A joint naive theory of psychology and physics could account for our ability to learn about the physical world from other people, and about other people from their actions in a physical world.

Acknowledgments

NIH F32HD103363, Johns Hopkins Waldrop Graduate Fellowship, Harvard GSAS Dissertation Completion Fellowship.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.*, *1*(4), 1–10.

Baron-Cohen, S. (1998). Does the study of autism justify minimalist innate modularity? *Learn. Individ. Differ.*, *10*(3), 179–191.

Baron-Cohen, S., Victoria Scahill, S. W., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? a test with children with asperger syndrome. *J. Dev. Learn. Disord.*, *5*, 47–78.

Begus, K., & Southgate, V. (2012). Infant pointing serves an interrogative function. *Dev. Sci.*, *15*(5), 611–617.

Bloom, P. (2005). *Descartes' baby: How the science of child development explains what makes us human*. Random House.

Carey, S. (2011). *The origin of concepts*. New York, NY: Oxford University Press.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naıve theory of rational action. *Trends Cogn. Sci.*, *7*(7), 287–292.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. MIT Press.

Hamilton, A. F. d. C., Brindley, R. M., & Frith, U. (2007). Imitation and action understanding in autistic spectrum disorders: how valid is the hypothesis of a deficit in the mirror neuron system? *Neuropsychologia*, *45*(8), 1859–1868.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends Cogn. Sci.*, *26*(11), 959–971.

Jack, A. I. (2014). A scientific case for conceptual dualism: The problem of consciousness and the opposing domains hypothesis. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy: Volume 1.* London, England: Oxford University Press.

Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccia, A. H., & Snyder, A. Z. (2013). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *Neuroimage*, *66*(C), 385–401.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends Cogn. Sci.*, *20*(8), 589–604.

Kamps, F. S., Julian, J. B., Battaglia, P., Landau, B., Kanwisher, N., & Dilks, D. D. (2017). Dissociating intuitive physics from intuitive psychology: Evidence from williams syndrome. *Cognition*, *168*, 146–153.

Karakose-Akbiyik, S., Caramazza, A., & Wurm, M. F. (2023). A shared neural code for the physics of actions and object events. *Nat. Commun.*, *14*(1), 3316.

Karakose-Akbiyik, S., Sussman, O., Wurm, M. F., & Caramazza, A. (2023). The role of agentive and physical forces in the neural representation of motion events. *The Journal of Neuroscience*, *44*(2), e1363232023.

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: neuropsychological evidence from autism. *Cognition*, *43*(3), 225–251.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, *105*(3), 489–512.

Malik, M., & Isik, L. (2023). Relational visual representations underlie human social interaction recognition. *Nat. Commun.*, *14*(1), 7317.

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends Cogn. Sci.*, *25*(2), 100–110.

Pramod, R. T., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. (2022). Invariant representation of physical stability in the human brain. *Elife*, *11*.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine theory of mind. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 4218–4227). PMLR.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nat. Commun.*, *9*(1), 1027.

Saxe, R., Tzelnic, T., & Carey, S. (2006). Five-month-old infants know humans are solid, like inanimate objects. *Cognition*, *101*(1), B1–8.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? evidence from target merging in multiple object tracking. *Cognition*, *80*(1-2), 159–177.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends Cogn. Sci.*, *4*(8), 299–309.

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., ... Ullman, T. (2021). AGENT: A benchmark for core psychological reasoning. In *Proceedings of the 38th international conference on machine learning* (pp. 9614–9625).

Spelke, E. S. (2022). *What babies know: Core knowledge and composition volume 1*. Oxford University Press.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends Cogn. Sci.*, *21*(9), 649–665.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Wilcox, T., & Biondi, M. (2015). Object processing in the infant: lessons from neuroscience. *Trends Cogn. Sci.*, *19*(7), 406–413.