Combining Psychology with Artificial Intelligence: What could possibly go wrong?

Iris van Rooij[1,2,3] and Olivia Guest[1,2]
[1]Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
[2]Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands
[3]Department of Linguistics, Cognitive Science, and Semiotics & Interacting Minds Centre, Aarhus University, Denmark

The current AI hype cycle combined with Psychology's various crises make for a perfect storm. Psychology, on the one hand, has a history of weak theoretical foundations, a neglect for computational and formal skills, and a hyperempiricist privileging of experimental tasks and testing for effects. Artificial Intelligence, on the other hand, has a history of conflating artifacts for theories of cognition, or even minds themselves, and its engineering offspring likes to move fast and break things. Many of our contemporaries now want to combine the worst of these two worlds. What could possibly go wrong? Quite a lot. Does this mean that Psychology and Artificial Intelligence can best part ways? Not at all. There are very fruitful ways in which the two disciplines can interact and theoretically inform the interdisciplinary study of cognition. But to reap the fruits one needs to understand how to steer clear of potential traps.

Keywords: theoretical psychology, artificial intelligence, cognitive science, computationalism, epistemology

Psychology has been living through various crises that have left the field grappling with its scientific status. Crises can lead to positive change, for instance, when they stimulate a reflective re-imagining of theoretical foundations and epistemological practices. However, crises can also leave a field vulnerable to false prophets that promise illusory quick-fixes. Right when Psychology is vulnerable, society is going through an AI hype cycle. This AI hype cycle's impact seems even worse than the ones that came before. With a devastating ecological footprint and the exploitation of hidden labour, hyped-AI serves to amplify discrimination and other social, economic and environmental injustices. Psychological scientists predominantly prefer to ignore such real-world harms and instead ask: 'How can AI benefit us?'[1] This is understandable.[2] After all, hyped-AI promises to deliver candidate theories and statistical inferences through automated processes, also known as *machine learning*. This is music to the ears of Psychology, which wants good theories and good statistical practices, but whose scientific practitioners predominantly lack theoretical, computational, and statistical skills. Hyped-AI even promises that it can replace human participants with 'artificial minds' that are amenable to standard psychological experimental methods. It seems, thus, that Psychology can just keep doing what it has been doing—e.g., effects hunting (Cummins, 2000; van Rooij & Baggio, 2021)—and no deep re-imagining of our discipline's foundations and epistemology may appear to be needed. However, these appearances and promises are all false. In this brief review we explain how and why. We end with guidance on how a fruitful interface between Psychology and Artificial Intelligence is possible that does not fall in these traps.

1 Traps to avoid

1.1 AI systems are not minds

If one reads the news and advertisements from the technology industry (sometimes disguised as scientific papers), one could be led to believe that we are on the verge of creating genuine artificial minds. For decades the domain-generality of cognition was recognized as making cognition hard—and perhaps impossible—to explain, model, or replicate computationally (Fodor, 2000; Pylyshyn, 1987; Ryle & Tanney, 2009; van Rooij et al., 2019). But these days many people have come to believe that by training on massive

Iris van Rooij Correspondence concerning this article should be addressed to Iris van Rooij, Donders Centre for Cognition, Faculty of Social Science, Radboud University, Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands. E-mail: i.vanrooij@donders.ru.nl

[1]For notable exceptions, see Guest (2024), Birhane and Guest (2021), and Prather et al. (2022).

[2]We do not mean to say it is ethical, but merely that we understand the psychological and socio-scientific factors behind this desire.

Table 1

Typology of traps, how they can be avoided, and what goes wrong if not avoided. Note that all traps in a sense constitute category errors (Ryle & Tanney, 2009) and the success-to-truth inference (Guest & Martin, 2023) is an important driver in most, if not all, of the traps.

traps to avoid	*how to avoid*	*problems if not avoided*
AI SYSTEMS ARE NOT MINDS	realise AI systems are decoys	dehumanisation, shoddy science
AI SYSTEMS ARE NOT THEORIES	realise AI systems are decoys, prediction is not explanation, correlation does not imply cognition, tasks are not capacities, and AI systems cannot scale	theoretical deterioration, pseudo-explanation, fallacious metatheoretical calculi, overclaiming, obfuscation of human-in-the-loop
COGSCI CANNOT BE AUTOMATED	think deeply, slow science, take computationalism seriously, be non-makeist	theoretical *cul-de-sacs*, pseudoscience, deskilling, proliferation of decoys, creation of the other traps

amounts of human data it is possible to create AI systems that can think and act in a domain-*general* way, *just like humans*. The intuition seems to be that as long as one has enough human data to train one's AI systems, those systems will asymptote to human-level/-like behavior. Such envisioned human-level AI is also known as Artificial *General* Intelligence (abbreviated AGI).

Recently van Rooij, Guest, et al. (2024) have shown that training AI systems to scale up to human-level cognition is intractable. This has two implications. First, creating AGI through machine learning inherently consumes astronomical amounts of resources—sooner will the sun die out than that we will create AGI, and in the mean time we will just be polluting our planet and exploiting un(der)paid labour. Second, any AI systems that *can* be created in the short-term are but *decoys*—these systems can trick us into thinking they are like human minds, but they are anything but (Guest & Martin, 2023). Even though these decoys can appear impressive and trick us, they are not hard to unmask through careful tests (Dentella et al., 2024) or even common sense probes.

It is thus worrisome that some researchers seem to believe that AI systems can replace human participants in experimental research (e.g. Dillion et al., 2023). Confusing AI systems for human minds is not only a category error (Ryle & Tanney, 2009) and dehumanising (Bender, 2024; Erscoi et al., 2023), it is also a recipe for shoddy science (Guest & Martin, 2023). After all, AI systems are decoys that cannot possibly approximate human cognition and behavior in any reliable way (van Rooij, Guest, et al., 2024). The methodological crisis in psychology has been bad enough (Flis, 2019). There is no benefit in making it worse by replacing the people whom we wish to study with decoys.

1.2 AI systems are not theories

So AI systems cannot be minds. But can these engineered systems be theories of how cognition works? It seems *prima facie* that training neural networks—or other cognitively inspired[3] computational architectures—on cognitive tasks and/or human data produces viable computational theories of how cognition works. After all, if such an AI system can mimic human behavior and predict[4] human performance on cognitive tasks, then human cognition must work more or less analogously to how the AI system works, right? Many people seem to believe that this implication holds. However, no matter how intuitively appealing, the belief is fundamentally mistaken, for several reasons.

Prediction is not explanation. Being able to predict human behavior does not imply being able to explain the *why* or

[3]For modern connectionism, appeals to 'neural plausibility' or describing systems as 'cognitively inspired' are often made without much basis (Guest & Martin, 2024), so not too much weight should be given to them. We merely mention them to acknowledge that these are common intuitive appeals made by proponents of AI systems as theories of cognition.

[4]As per the *Ingenia Theorem* (van Rooij, Guest, et al., 2024) we know AI systems fundamentally cannot predict human-level performance on cognitive tasks in a domain-general sense, even if such systems may predict or mimic human performance on well-circumscribed tasks.

how of that behavior. That prediction and explanation are dissociable is easily illustrated by considering the tides (Cummins, 2000; see also Blokpoel and van Rooij, 2021–2025, Chapter 2): We could predict the tides long before we could explain them (i.e., in terms of the gravitational pull of the moon). Even today, we use tide tables, i.e., large lookup tables that map dates and times of day to positive or negative levels of the tide. Using such tide tables for different locations on earth one can quite precisely *predict* the tides at any time of the day. Yet, no-one would claim that tide tables *explain* the tides. Similarly, a huge look-up table (or a compressed version that interpolates/guesstimates for some unknown/unseen input-entries, like a regression model, a neural network, or a large language model) that more or less predicts what behaviors people will display in different situations and conditions does not provide an explanation for why the behavior is as it is, nor of how cognition works (van Rooij, 2022).

Correlation does not imply cognition (Guest & Martin, 2023, p. 224). Some may object that even though prediction of outward behavior is insufficient for a model to be explanatory, surely when internal parameters of the model correlate with brain data that shows the model matches the mechanistic workings of brains/minds. Unfortunately, this inference is mistaken, too. As shown by Guest and Martin (2023), it is invalid to infer from correlations between a model and brain data that the model works like the brain. This follows from *multiple realisability*[5]: Just like both a digital clock and an analog clock can tell the time, and one will be able to correlate parts from one with the other, they operate in fundamentally different ways (Fig. 2 in Guest & Martin, 2023).

Capacities are not tasks. Even if multiple realisability cannot be ruled out, it may seem that AI systems that can (be made to) perform like humans on cognitive tasks provide at least *possible* theories of how cognition *could* work. However, this inference is not licensed either. Computational models of tasks, and task performance, are not yet theories of cognition (Guest & Martin, 2021; Morrison & Morgan, 1999). Theories of cognition, minimally, should provide *possible* explanations of one or more *substantive* human capacities, such as vision, decision-making, reasoning, or communication (Cummins, 2000; Egan, 2018; van Rooij & Baggio, 2021). While it is true that in (hyperempiricist) Psychology these capacities are typically studied by having people perform various tasks, computationally (or more often, statistically) modeling task performance does not yield explanatory theories of capacities. This is so, not only because tasks do not map one-to-one, or in any other straightforward way, to substantive cognitive capacities, but even if they could, the models would not be able to scale up to situations of real world complexity.

AI systems cannot scale. In order for computational models of cognition to be able to scale from toy scenarios (such as studied in the psychological labs or such as form the bases of training AI systems through machine learning) these models should minimally be computationally tractable (van Rooij, 2008). At present, no computationally tractable account exists for substantive and domain-general cognitive capacities, such as reasoning, communication, decision-making, planning, analogizing, categorisation and concept formation (van Rooij et al., 2019) nor will any such tractable accounts be forthcoming via machine learning (van Rooij, Guest, et al., 2024) or otherwise (Rich et al., 2021). Hence, if someone claims their AI system is a 'theory' of cognition, they are overstating the scope and capacities of the system (van Rooij et al., 2019) and obfuscating the system's limits, including the human-in-the-loop needed to make such systems 'work' (Guest & Martin, 2025).

1.3 Cognitive science cannot be automated

Psychology has undergone some important cultural changes due to the so-called replication crisis.[6] Instead of improving Psychology's theoretical foundations and epistemological practices by adopting conceptual, computational, and formal tools from Computational Cognitive Science (Guest, 2024; Guest & Martin, 2021; Navarro, 2019, 2021; van Rooij & Baggio, 2020, 2021; van Rooij, Devezer, et al., 2024), the overwhelming response in the mainstream has been to push for statistical methodological reform that centers a rigid proceduralisation of empirical research (see also Devezer et al., 2021; Szollosi et al., 2020, for critiques). This move has further cemented hyperempiricism[7] into Psychology and it has left the field vulnerable to a view that science can be proceduralised and maybe, to a large extent, even automated.

Right at the time that Psychology is vulnerable, hyped-AI enters the scene and makes false and misleading promises that both theory-generation and scientific inference can be automated using *machine learning*. For Psychology such promises are very attractive, especially since its practitioners often lack theory building and advanced statistics skills. However, a common expression applies here: 'if something seems too good to be true, then it probably is'. Cognitive science *cannot* be automated, because theory generation is

[5]The principle of *multiple realisability* is foundational to *computationalism* (i.e., the idea that cognition is, or can be understood as, a form of computation). It is ironic that especially those who believe AI systems can be minds or theories of how cognition works ignore the fundamental principle of multiple realisability in their *metatheoretical calculus* (see also Guest & Martin, 2025; Guest et al., 2025).

[6]See, for instance, Nosek et al., 2022. But see also Devezer et al., 2021; Irvine, 2021; Szollosi et al., 2020 for critiques of how this crisis has been conceived and addressed by the mainstream.

[7]Here, by *hyperempiricism* we mean the idea that only empirical observation can be useful to understanding cognition, and that any other source of evidence is either of lesser import or irrelevant.

provably intractable (Rich et al., 2021; van Rooij, Guest, et al., 2024) and scientific inference cannot be reduced to statistical inference (Guest & Martin, 2023; Navarro, 2019).

Hyped-AI promises are not harmless (Guest, 2024). While automation may give the false impression of rigor and efficiency, it leads to conceptual and scientific deskilling, deteriorates reflexive theorizing, and can make us blind for important scientific paths we'd need to go down (Rich et al., 2021). Also, since building theories of substantive (human-level) cognitive capacities is computationally intractable, any efficient proceduralized way of generating theories can only produce decoys, leading to the other traps (van Rooij, Guest, et al., 2024). Last but not least, automated scientific inferences can cause deep scientific inconsistencies and theoretical confusions (Guest & Martin, 2023; Guest et al., 2025) and can give false credibility to harmful pseudoscientific ideas and practices (Birhane & Guest, 2021; Spanton & Guest, 2022)

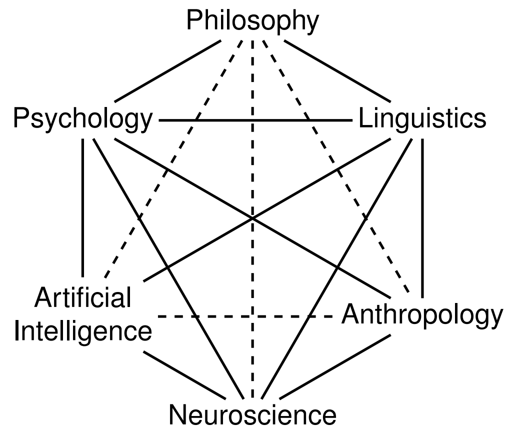

Figure 1

A visual depiction of the connections between the Cognitive Sciences. Solid lines denote stronger interdisciplinary ties; and dashed lines denote weaker ones. This figure is derived from the original put forth by the Sloan Foundation in 1978 and reproduced from Figure 4 in Pléh and Gurova (2013). Different versions of it over time have used 'Artificial intelligence' (as above) instead of 'Computer Science' and vice versa (cf. Miller, 2003).

2 A possible path forward

In this brief review we focused on what all can go wrong when combining Psychology with Artificial Intelligence in thoughtless ways. We realize the reader may appreciate guidance for traveling the winding, branching, and open-ended road that is cognitive science without falling into said traps. Step 1 in avoiding the traps is to be aware of them and to be able to recognize them 'in the wild' (e.g, in the literature or in scientific practices). To assist with this we provide in Table 1 an overview of the nature of each of the traps, what can be done to avoid them, and the problems that arise when one doesn't. Step 2 is to develop a research approach that removes the root causes of the traps and prevents them from arising in the first place. We (and others) have proposed that such a cognitive science is possible even within a computationalist framework, if we reconceptualise Artificial Intelligence (or Computer Science more broadly) as "a provider of computational tools (frameworks, concepts, formalisms, models, proofs, simulations, etc.) that support theory building in cognitive science" (van Rooij, Guest, et al., 2024, p. 616), but without confusing the *theoretical* possibility of explaining human cognition computationally with the *practical* feasibility of (re)making human cognition in factual computational systems (a.k.a. *makeism*, see Box 2 in van Rooij, Guest, et al., 2024). We coined this alternative computationalist approach *non-makeism*.

Non-makeist AI takes computationalism more seriously than makeist AI (Guest et al., 2025), as it bites all the bullets implied by computationalist axioms, such as multiple realisability of computation and fundamental limits imposed by computational intractability. "Cognitive science is itself a cognitive activity" (Rich et al., 2021, p. 3034). It, thus, follows from computationalism that cognitive scientists' explanations, inferences, and theory building are all limited by computational constraints as well. Lacking any efficient reliable procedure for generating explanatory theories, all we can do is postulate (often blatantly wrong) theories and rigorously analyze their explanatory scope and limits. By using whatever insights we may draw from such analyses we advance our scientific understanding (of our lack of understanding) one small step at a time. Good science is slow (Stengers, 2018) and if cognitive science wants to take AI as theoretical psychology back on board, then it needs to take computationalism seriously.

3 Conclusion

Psychology and Artificial Intelligence (or, more broadly, Computer Science) are two of the six traditional disciplines constituting the interdisciplinary study of cognition, called Cognitive Science (the other four being Philosophy, Linguistics, Neuroscience, and Anthropology; see Fig. 1). Over the last three decades Psychology came to dominate Cognitive Science with its hyperempiricist tendencies (Gentner, 2010, 2019), while Artificial Intelligence retracted from the field (Forbus, 2010) taking most computational theory building tools with it. Currently, we are witnessing a rapprochement of the two disciplines. While theoretical strengthening of Cognitive Science is welcome, great caution is needed to prevent a new status quo that is worse than the old one. Computational concepts remain valuable for carefully crafting theories in Cognitive Science (Guest & Martin, 2021; van Rooij

& Baggio, 2021), but they can only flourish if we *a*) do not confuse AI systems for minds or theories, *b*) do not confuse machine learning for the scientific method, and *c*) understand that our computational models can only track the scope and limits of our understanding.

4 Recommended readings

Guest, O. (2024). What makes a good theory, and how do we make a theory good? *Computational Brain & Behavior*, *7*(4), 508–522.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*(2), 213–227.

Guest, O., Scharfenberg, N., & van Rooij, I. (2025). Modern alchemy: Neurocognitive reverse engineering. *PhilSci*. https://philsci-archive.pitt.edu/id/eprint/25289

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697.

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, *7*, 616–636.

References

Bender, E. M. (2024). Resisting dehumanization in the age of "AI". *Current Directions in Psychological Science*, *33*(2), 114–120.

Birhane, A., & Guest, O. (2021). Towards decolonising computational sciences. *Kvinder, Køn & Forskning*, *29*(2), 60–73.

Blokpoel, M., & van Rooij, I. (2021–2025). *Theoretical modeling for cognitive science and psychology*. https://computationalcognitivescience.github.io/lovelace/

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In *Explanation and cognition* (pp. 117–144). MIT Press.

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1), 28083.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society open science*, *8*(3), 200805.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*.

Egan, F. (2018). Function-theoretic explanation and the search for neural mechanisms. In *Explanation and integration in mind and brain science* (pp. 145–163). Oxford University Press.

Erscoi, L., Kleinherenbrink, A., & Guest, O. (2023). Pygmalion displacement: When humanising AI dehumanises women. *SocArXiv*. https://doi.org/10.31235/osf.io/jqxb6

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*(2), 158–181.

Fodor, J. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. MIT press.

Forbus, K. D. (2010). AI and cognitive science: The past and next 30 years. *Topics in Cognitive Science*, *2*(3), 345–356.

Gentner, D. (2010). Psychology in cognitive science: 1978–2038. *Topics in Cognitive Science*, *2*(3), 328–344.

Gentner, D. (2019). Cognitive science is and should be pluralistic. *Topics in Cognitive Science*, *11*(4), 884–891.

Guest, O. (2024). What makes a good theory, and how do we make a theory good? *Computational Brain & Behavior*, *7*(4), 508–522.

Guest, O., & Martin, A. (2025). Are neurocognitive representations 'small cakes'? https://philsci-archive.pitt.edu/24834/

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Guest, O., & Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*(2), 213–227.

Guest, O., & Martin, A. E. (2024). A metatheory of classical and modern connectionism. *PsyArXiv*. https://osf.io/preprints/psyarxiv/eaf2z_v1

Guest, O., Scharfenberg, N., & van Rooij, I. (2025). Modern alchemy: Neurocognitive reverse engineering. *PhilSci*. https://philsci-archive.pitt.edu/id/eprint/25289

Irvine, E. (2021). The role of replication studies in theory building. *Perspectives on Psychological Science*, *16*(4), 844–853.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in cognitive sciences*, *7*(3), 141–144.

Morrison, M., & Morgan, M. (1999). Models as mediators. *Perspectives on Natural and Social Science. Cambridge University Press*.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, *2*(1), 28–34.

Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, *16*(4), 707–716.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual review of psychology*, *73*(1), 719–748.

Pléh, C., & Gurova, L. (2013). Existing and would-be accounts of the history of cognitive science: An introduction. *Pléh Csaba–Gurova, Lilia–Ropolyi, László (2013 ed.) New Perspectives on the History of Cognitive Science. Akadémiai Kiadó, Budapest*.

Prather, R. W., Benitez, V. L., Brooks, L. K., Dancy, C. L., Dilworth-Bart, J., Dutra, N. B., Faison, M. O., Figueroa, M., Holden, L. R., Johnson, C., et al. (2022). What can cognitive science do for people? *Cognitive Science*, *46*(6), e13167.

Pylyshyn, Z. W. (1987). The robot's dilemma: The frame problem in artificial intelligence.

Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science? *Proceedings of the annual meeting of the cognitive science society*, *43*(43).

Ryle, G., & Tanney, J. (2009). *The concept of mind*. Routledge.

Spanton, R. W., & Guest, O. (2022). Measuring trustworthiness or automating physiognomy? a comment on safra, chevallier, gr\ezes, and baumard (2020). *arXiv preprint arXiv:2202.08674*.

Stengers, I. (2018). *Another science is possible: A manifesto for slow science*. John Wiley & Sons.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in cognitive sciences*, *24*(2), 94–95.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984.

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, *1*(3), 127–128.

van Rooij, I., & Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry*, *31*(4), 321–325.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697.

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and Intractability: A Guide to Classical and Parameterized Complexity Analysis*. Cambridge University Press.

van Rooij, I., Devezer, B., Skewes, J., Varma, S., & Wareham, T. (2024). What makes a good theory? interdisciplinary perspectives. *Computational Brain & Behavior*, 1–5.

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, *7*, 616–636.