Fodor and Pylyshyn's systematicity challenge still stands: A reply to Lake and Baroni (2023)

Michael Goodale and Salvador Mascarenhas
Institut Jean-Nicod
Department of Cognitive Studies
Ecole Normale Supérieure

December 2023

The recent successes of neural networks producing human-like language have captured the attention of the general public. They have also caused significant stir in cognitive science, with many researchers arguing that classical puzzles about human cognition and challenges to artificial intelligence are being solved by neural networks. An article recently published in Nature [1], covered by the journal's media department as a "breakthrough" in AI, argues that a particular machine-learning technique has succeeded where others failed: to match and perhaps explain the human ability to reverse engineer generative processes (rules) based on few examples. We demonstrate that these conclusions are premature. Among other results, we found that the model displays different rates of generalization success depending on what labels are attached to what meanings. This is in sharp contrast with the fact that there are no linguistic or broader cognitive benefits from calling a carbonated beverage "pop" or "soda," nor from calling the objects of study of dendrology "trees" or "Bäume." Crucially, our examples of failures lie squarely within the narrow task that the article focuses on, calling into question the ambitious conclusions and the bullish media coverage the article received.

1 Failures of systematic generalization

One stated central goal of Lake and Baroni's (L&B) is to address Fodor and Pylyshyn's (F&P) negative outlook on the prospects of neural networks as models of higher cognitive faculties [2]. The gist of this argument from *systematicity* is that many distinctively human faculties display biconditional interdependencies: a human has faculty X precisely to the extent that they have faculty Y. One has the ability to understand "Ann introduced Bill to Claire" precisely to the extent that one can understand "Bill introduced Claire to Anne," "Claire introduced Dan to Ed," and so forth. As one of the arguments goes, due to their high sensitivity to the training data, neural networks cannot guarantee systematicity in this sense, while symbolic systems would have to go out of their way to be unsystematic.

There are legitimate reasons to be critical of F&P's arguments, both regarding the problem with neural networks and the virtues of classical symbolic systems [3, 4]. Since L&B manifestly accept the challenge, we take it as fair to assess their results through the lens of classical systematicity, whatever its flaws. Moreover, some neural networks (curiously, trained in a non-traditional way) were *already* known to be capable of classical systematicity [5]. This would not have surprised F&P: their question was whether purely associative learning can *guarantee* systematicity, not whether *there exist* particular neural networks that are systematic. Consequently, the importance of L&B's results must be that their meta-learning technique indeed guarantees systematic learning.

With these considerations in mind, we took a close look at the "gold grammar" defined by L&B. The gold grammar is a blueprint for languages where the rules are all the same, but different labels and colors are used. We also evaluated similar language blueprints with small changes to the rules. We found failures of systematicity that we classify into three levels, in increasing (qualitative) abstraction. Level one failures are *within-grammar* failures, where the model succeeds for a string but fails for a minor modification of it. Level two considerations look *across instantiations of a grammar blueprint* at correlations between network performance and properties of grammars which oughtn't to matter, specifically what labels (words) are mapped to what colors (meanings). At level three we look *across grammar blueprints* at the learning profiles for different rules. Our observations here concern the `algebraic` model, which had the best generalization performance, but these problems are present with the other variants.

Below, we use c_i both to represent a specific color and the word for that color. c_1, c_2, c_3 are the colors that were used with function words in training examples and c_h is the held-out color used to test generalization. The three rules of the gold grammar are "after" (flip two strings), "thrice" (repeat a color three times) and "surround" (surround a color with two instances of another color). For example, "c_1 surround c_h" can be instantiated in a particular language as `dax kiki fep`. This string would have the schematic meaning $c_1 c_h c_1$.

1.1 Level one failures

The article reports several successfully generalized strings, but we found the model fails for strings of comparable complexity. Thus, the models do not display *compositionality*, a central feature of systematicity in natural language.

A striking example is the schema "c_1 surround c_h after c_h thrice," which is incorrectly predicted in many grammars (see Table 1 and the supplementary materials). Importantly, this schema is a very minor variation on one tested in the article and successfully learned within the same grammars that fail our variant: "c_h surround c_h after c_h thrice." The two schemata are different special cases of "x surround y after z thrice," showing that performance differs dramatically between equally valid combinations of x, y, z (c_h, c_h, c_h vs. c_1, c_h, c_h). This is akin to a human understanding the English sentence "Ann introduced Bill to Claire" while being stumped by "Dan introduced Claire to Anne."

1.2 Level two failures

The model is sensitive to precisely which labels (words) are mapped to which meanings (rules and colors), with different mappings producing different performances. Thus, the networks do not conform to the arbitrariness of linguistic signs, whereby no particular benefits or drawbacks are created by American English "sofa" vs. (Old) Canadian English "chesterfield."

For example, the same schema discussed above ("c_1 surround c_h after c_h thrice") shows different learning successes *across* grammars, being correctly generalized for some label-color pairs and not for others (Table 1). Additionally, we found that the model has different success rates on the schema, depending on precisely which color is held out (Table 2).

1.3 Level three failures

We found that the network is sensitive to what it saw at the meta-learning phase. The meta-learning protocol only exposed the networks to rules whose outputs had lengths between 2 and 8 words, and we found that they could not learn a rule that was a trivial extension, violating this 2–8 constraint. Specifically, while the model can learn a function that takes

Output	Count across label-color pairs Static order	 Shuffled order
$c_h\ c_h\ c_h\ c_1\ c_h\ c_1$	1993	2785
$c_1\ c_h\ c_h\ c_h\ c_h\ c_1$	1045	934
$c_h\ c_h\ c_h\ c_h\ c_1\ c_1$	1282	598
$c_h\ c_h\ c_h\ c_h\ c_1\ c_h$	0	3

Table 1: Four very different outputs for the same schema "c_1 surround c_h after c_h thrice," with counts across label-color pairings. The first is the normatively correct output. The second is a possible output if one parses the string as "x surround (y after z thrice)," though technically this was unintended with the gold grammar, as "surround" examples only ever take two primitives (color-denoting labels) as arguments. The third and fourth aren't acceptable outputs under any parse. *Static order* vs. *shuffled order* concerns the order in which the study items were presented to the network. With static order of presentation the network's accuracy on this schema is at most 70%, with shuffled order at most 86%. The different results depending on order of presentation are instructive, as they indicate their own kind of non-systematic fragility.

	Constant presentation order		Shuffled presentation order	
Held out color	% correct	SEM	% correct	SEM
Blue	56.11	1.8506	59.72	1.8291
Green	62.08	1.8094	73.47	1.6464
Pink	38.47	1.8144	67.36	1.7487
Purple	50.28	1.8647	69.03	1.7244
Red	53.75	1.8594	75.83	1.5965
Yellow	16.11	1.3710	41.39	1.8368

Table 2: Accuracy on held-out-label-color combinations for the string "c_1 surround c_h after c_h thrice."

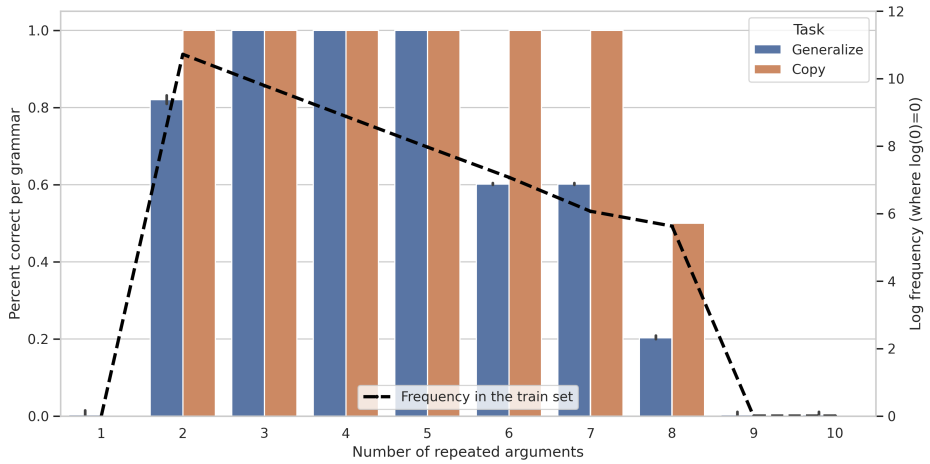

Figure 1: Learning functions that take an argument x and return n repetitions of x. Generalize strings are the original test sequences where we replace thrice by n. Copy shows the model's ability to repeat sequences that are provided as study examples (i.e. lookup). The figure shows confidence intervals but the distribution is has so little variation that they are hard to see.

one argument and repeats it n times for $2 \leq n \leq 8$, it couldn't learn an analogous rule that repeats its argument, say, 9 times. This shows that the model isn't a *systematic learner*, instead, it is entirely beholden to what precisely appeared in the meta-learning phase. Additionally, we found that performance within the 2–8 length bounds drops as the output gets longer, mirroring the frequency with which functions were encountered at meta-learning (Figure 1, note that performance drops even for the simple copying task where the query string was one that occurred in the study set).

2 Discussion and conclusions

We found three different levels of failures of systematic generalization in L&B's algebraic network. We conclude that, while the network was indeed successful at later learning on the basis of extremely small datasets, these successes cannot be described as instances of systematic generalization, so that F&P's challenge is in fact *not* addressed by L&B's MLC protocol, in its current shape.

One counter to our conclusion, consonant with the methodology of the article concerning behavioral data from humans, is to propose that neural networks only need to be as systematic as humans are *in performance*. There are all manner of reasons why a particular human in a particular circumstance might behave differently on the basis of the two stimuli "Ann introduced Bill to Claire" and "Claire introduced Dan to Ed." Having had an unpleasant experience with someone named Bill might do the trick. Cognitive science recognizes that a faculty may be systematic, while particular circumstances of use of a faculty might not show it.

Now, we strongly suspect that the failures of systematicity we presented here would not be observed in humans. Authors and readers might not share our expectation, but we submit that our findings pose an important challenge to the conclusions of the target article nonetheless. Because the puzzle the article ostensibly attempts to solve concerns *competence*, one of two positions would have to be argued for. The first option is to accept the classical points about systematicity, despite performance complications, and to somehow

investigate the *competence* of MLC networks, to show that *it* is systematic in the required sense [6]. At points, the authors imply that the algebraic model is a model of competence, while the models supplemented with heuristics and human performance data would be models of performance. Yet, our discoveries here concern precisely the algebraic model, and show that on it its face it is not systematic. The second option is to reject the classical arguments: perhaps classical systematicity is actually *not* a property of human cognition, or competence is too idealized a notion to be of real interest, or any number of other responses that argue F&P's challenge away. The points made by the authors about systematicity in the classical sense would be moot, and the interest of the work would shift from a strong claim about the human-like generalization powers of MLC networks to the far more modest point about successes producing human-like performance partly on the basis of human performance data, which we haven't the space to comment on properly in this brief reply.

However, neither of these options is pursued in the article, which instead positions itself clearly and boldly as a solution to the systematicity challenge. Consequently, our conclusions stand, calling into question the success of the authors' MLC technique, irrespective of behavioral experimentation investigating precisely how humans fare with the failures of systematicity we document here.

Author contribution statements

Both authors were involved in the study's conception and design. Programming and running experiments was handled by MG. Both authors contributed to the writing and editing of the article.

Code availability

All relevant code for these experiments, allowing for full reproduction of our results, can be found at `https://github.com/MichaelGoodale/mlc-reply`.

Data availability

All relevant data is included in the repository of our code or can be re-created by running our code.

Funding acknowledgment

The work reported here was funded in part by Agence Nationale de la Recherche grant ANR-19-P3IA-0001 (PRAIRIE 3IA Institute, PI: Mascarenhas) and by a grant from Ecole Doctorale Frontières de l'Innovation en Recherche et Education—Programme Bettencourt (PhD funding for Goodale).

References

[1] Brenden M. Lake and Marco Baroni. "Human-like systematic generalization through a meta-learning neural network." In: *Nature* 623.7985 (2023), pp. 115–121. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06668-3.

[2] Jerry Fodor and Zenon W. Pylyshyn. "Connectionism and cognitive architecture: A critical analysis." In: *Cognition* 28.1–2 (1988), pp. 3–71. DOI: 10.1016/0010-0277(88)90031-5.

[3] Steven Phillips and William H. Wilson. "Categorial compositionality: A category theory explanation for the systematicity of human cognition." In: *PLoS Computational Biology* 6.7 (2010). DOI: 10.1371/journal.pcbi.1000858.

[4] David Chalmers. "Why Fodor and Pylyshyn were wrong: The simplest refutation." In: *Proceedings of the 12th Annual Meeting of the Cognitive Science Society*. Ed. by Massimo Piattelli-Palmarini, Beth Adelson, Stephen M. Kosslyn, Steven Pinker, and Ken Wexler. 1990, pp. 340–347. URL: https://cognitivesciencesociety.org/wp-content/uploads/2019/01/cogsci_12.pdf.

[5] Nur Lan, Michal Geyer, Emmanuel Chemla, and Roni Katzir. "Minimum description length recurrent neural networks." In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 785–799. DOI: 10.1162/tacl_a_00489.

[6] Chaz Firestone. "Performance vs. competence in human-machine comparisons." In: *Proceedings of the National Academy of Sciences* 117.43 (2020), pp. 26562–26571. DOI: 10.1073/pnas.1905334117.