Contrafactives and Learnability*

David Strohmaier[1] and Simon Wimmer[2]

[1] Department of Computer Science and Technology
ALTA Institute
University of Cambridge
ds858@cam.ac.uk
[2] Institut für Philosophie und Politikwissenschaft
Technische Universität Dortmund
simon.wimmer@tu-dortmund.de

Abstract

Richard Holton has drawn attention to a new semantic universal, according to which (almost) no natural language has contrafactive attitude verbs. This semantic universal is part of an asymmetry between factive and contrafactive attitude verbs. Whilst factives are abundant, contrafactives are scarce. We propose that this asymmetry is partly due to a difference in learnability. The meaning of contrafactives is significantly harder to learn than that of factives. We tested our hypothesis by conducting a computational experiment using an artificial neural network. The results of this experiment support our hypothesis.

1 Introduction

Holton (2017) has drawn attention to a novel semantic universal, according to which (almost) no natural language has contrafactive attitude verbs. Contrafactives are the mirror image of factive attitude verbs, such as *know*, *remember*, *see*, and *regret*. Although both factives and contrafactives entail a belief, contrafactives differ from factives in presupposing the falsity, as opposed to truth, of their declarative complements. To illustrate, suppose we extend our language with the contrafactive *contra*. Now, whilst both *Dan knows that Maggie is dancing* and *Dan contras that Maggie is dancing* entail that Dan believes that Maggie is dancing, the former presupposes that it is raining, the latter that it is not. This difference in presuppositions between the factive and contrafactive surfaces in the following diagnostics.

1. # Umut knows that it's raining, but it isn't.
2. # Umut contras that it's raining, and it actually is.
3. (a) Does Eylem know that we've won?
   (b) Eylem doesn't know that we've won.
   → We've won.
4. (a) Does Eylem contra that we've won?
   (b) Eylem doesn't contra that we've won.
   → We haven't won.

*This paper reports on research supported by Cambridge University Press and Assessment, University of Cambridge. We thank the NVIDIA Corporation for the donation of the Titan X Pascal GPU used in this research. Simon Wimmer's work on this paper was supported by a postdoc stipend of the Fritz Thyssen Foundation. We thank audiences in Bochum, Dortmund, Essen, Tokyo, and Utrecht for discussion of related material. David Strohmaier designed and ran the computational experiment, Simon Wimmer did conceptual work to connect philosophical and linguistic discussions to the experiment.

The diagnostic in 1 and 2 shows that the inference to the truth/falsity of the verb's declarative complement cannot be cancelled; the diagnostic in 3 and 4 suggests that the inference projects through entailment-cancelling environments, such as question and negation.

An important further feature of a contrafactive, according to Holton, is that it is an "atomic propositional attitude verb" (2017, p.248).[1] Thus, the inference to the falsity of its declarative complement is not the result of a compositional method: that would make the target expression "molecular". This means that Anvari, Maldonado, and Soria Ruiz's (2019) Spanish *creerse* 'wrongly believe' (as well as Shatz et al.'s (2003) Puerto Rican Spanish *creerse*) does not count as a contrafactive. Although *creerse* performs much as *contra* on the above diagnostics (with the exception of what Anvari, Maldonado, and Soria Ruiz call "polarity reversal under negation"), *creerse* is built by adjoining the reflexive pronoun *se* to the non-factive verb *creer* 'believe'.[2] And, since adjoining the reflexive pronoun has similar effects in the case of *pensarse* and French *s'imagine*, Anvari, Maldonado, and Soria Ruiz (2019, p.72) suggest that the inference to the falsity of *creerse*'s declarative complement results, at least partially, from composing the meanings of *se* and *creer*. Hence, *creerse* is not atomic, and so not a contrafactive.

Holton (2017, pp.245-9, 262–4) considered several apparent counterexamples to his universal found in non-Indo-European languages. However, none is a genuine contrafactive. For instance, the Mandarin verb *yǐwéi*, glossed by Lee, Olson, and Torrance (1999) and Cheung, Chen, and Yeung (2009) as 'believe wrongly', has been found to carry a post-supposition that the reported belief must not be added to the common ground, rather than a presupposition that its declarative complement is false (Glass, 2022): the inference *yǐwéi* triggers can be cancelled and does not project through negation. Likewise, the inferences triggered by Shatz et al.'s (2003) Turkish belief verbs *san* and *zannet* can be cancelled, too.[3]

That no natural language has contrafactives raises the question: why do natural languages universally have factives like *know* (Goddard, 2010; Haspelmath and Tadmor, 2009), but universally lack contrafactives? Importantly, the issue here would remain, even if some counterexamples to Holton's universal were eventually found. For even if there were some contrafactives, an asymmetry between factives and contrafactives would persist: factives would be abundant, contrafactives scarce.

Our aim here is to uncover one reason for the asymmetry between factives and contrafactives. Drawing on recent discussions of other semantic universals, like the veridical uniformity universal for responsive verbs (Steinert-Threlkeld, 2019), the conservativity, monotonicity, and quantity universals for determiners (Steinert-Threlkeld and Szymanik, 2019), and the convexity universal for color terms (Steinert-Threlkeld and Szymanik, 2020), we explore the hypothesis that the asymmetry between factives and contrafactives arises partly because the meaning of a contrafactive is harder to learn than that of a factive. Our hypothesis is inspired by the intuitive idea that languages have words for meanings that are easier to acquire and use compositional methods to express meanings that are harder to learn (Steinert-Threlkeld and Szymanik, 2019, p.4). We tested our hypothesis by conducting a computational experiment using an artificial neural network. As we explain below, the results of the experiment support our hypothesis.

[1] Holton adopts two further necessary conditions an expression must satisfy in order to count as a contrafactive. In parallel with *know*, he would regard *contra* as a mental state verb and as responsive (embedding both declarative and interrogative complements). For present purposes, however, we set these conditions aside. For one, we take the question of why no natural language has a verb with the features noted in the text to be of independent interest. For another, we expect the work we present here to also go some way toward addressing why no natural language has a verb that satisfies all of Holton's conditions.

[2] By contrast with factives and contrafactives, a non-factive, like *believe* or *think*, triggers neither an uncancellable inference to the truth/falsity of its declarative complement nor an inference to truth/falsity that projects through entailment-cancelling environments.

[3] We are grateful to Dilara Malkoc for discussion of the Turkish data.

2 Hypothesis

Before we turn to our computational experiment, we want to provide an intuitive motivation for our claim that factives are easier to learn than contrafactives. This motivation is inspired by Phillips and Norby's (2021) work on differences between factive and non-factive mental state attribution (see also Nagel, 2017; Phillips et al., 2020).

Suppose a speaker utters the factive attitude ascription *Dan knows that Maggie is dancing.* Since factives presuppose the truth of their declarative complements, this utterance commits the speaker to it being true that Maggie is dancing. Further, since factives entail a belief, the ascription entails that its subject, Dan, is also committed to it being true that Maggie is dancing. For this reason, the factive attitude ascription represents the ways the speaker and the subject of the ascription take the world to be as converging. Put in terms of Phillips and Norby's (2021) map analogy, the speaker's map of the world and the map they attribute share certain parts; our speaker simply copy-pastes these parts from their own map onto the other.

The ascription of non-factive attitudes is significantly less constrained. To see this, suppose a speaker utters the non-factive attitude ascription *Dan believes that Maggie is dancing.* Since non-factives do not presuppose the truth of their declarative complements, this utterance does not commit our speaker to it being true that Maggie is dancing. However, the ascription does entail that its subject, Dan, is committed to it being true that Maggie is dancing. For this reason, the non-factive attitude ascription by itself leaves open whether the ways the speaker and the subject of the ascription take the world to be converge or diverge. Thus, the speaker cannot simply copy-paste parts of their own map of the world onto the map they attribute. In this sense, their own take on what the world is like does not constrain, and so simplify, their ascription of a non-factive attitude to another person.

We can extend the point from non-factives to contrafactives. Suppose a speaker utters *Dan contras that Maggie is dancing.* Since contrafactives presuppose the falsity of their declarative complements, this utterance commits our speaker to it being false that Maggie is dancing. Yet the ascription entails that its subject, Dan, is committed to it being true that Maggie is dancing. Given this, the contrafactive attitude ascription represents the ways the speaker and the subject of the ascription take the world to be as diverging. And so, the speaker's map of the world and the map they attribute have incompatible parts. Consequently, they cannot copy-paste parts of their own map of the world onto the map they attribute. As for non-factives, the speaker's own take on what the world is like does not constrain, and so simplify, their contrafactive attitude ascriptions in the way in which it constrains their factive attitude ascriptions.

Importantly, whether a speaker's own take on what the world is like constrains, and thus simplifies, their ascription of factive and contrafactive attitudes is arguably (at least partly) due to the meaning of factives and contrafactives. This, however, leads us to expect the meaning of a contrafactive to be harder to acquire than that of a factive.

Building on Phillips and Norby (2021), we also expect the meaning of a contrafactive to be slightly easier to learn than that of a non-factive. For there is a sense in which a speaker's take on what is the case does constrain, even if only slightly, their contrafactive, but not their non-factive, attitude ascriptions. If, say, our speaker takes Maggie to be dancing, they cannot consistently claim that Dan contras that Maggie is dancing, but can consistently ascribe a belief to that effect. So, the speaker's take on what is the case rules out a contrafactive attitude ascription for them, but leaves open the corresponding non-factive attitude ascription. In this sense, information about the way the world is only contributes noise to the ascription of non-factive attitudes; noise one must learn to ignore. As we note below, our experimental results also support the claim that contrafactives are slightly easier to learn than non-factives.

The result that the meaning of a non-factive is harder to acquire than that of a contrafactive can seem puzzling, given that non-factives like *think* are universal across natural languages (Goddard, 2010). However, the added difficulty of acquiring the meaning of a non-factive just described does not entail that it is less common. Other factors that our current experiment does not model can make a contrafactive *overall* harder to learn than a non-factive. For instance, on the pragmatic syntactic bootstrapping model of how infants acquire attitude verb meanings (Hacquard and Lidz, 2022), the meaning of non-factive *think* is partly inferred from the parallel between the use of *I think P* as an indirect assertion and the primary use of *P* as an assertion, and the meaning of factive *know* is partly inferred from the parallel between the use of *Do you know Q?* as an indirect question and the primary use of *Q?* as a question. Yet unlike in the case of factives and non-factives, no such parallels would hold for a contrafactive. One cannot use *Dan contras that Maggie is dancing*, say, as an indirect assertion that Maggie is dancing. Thus, use of a contrafactive attitude ascription would not match the primary use of its complement. And so, we expect pragmatic syntactic factors to make it harder to acquire contrafactives than non-factives and factives. We leave it to future work to explore how this pragmatic syntactic difference between factive, non-factive, and contrafactive attitude ascriptions and the differences suggested by our computational experiment combine to explain the difference in frequency of factive, non-factive, and contrafactive attitude verbs in natural languages.

3 Experiment and results

To test our expectation that the meaning of a contrafactive is harder to acquire than that of a factive, we conducted a computational experiment using an artificial neural network, specifically a Transformer encoder.[4] This network was trained to predict the truth value of factive, non-factive, and contrafactive attitude ascriptions, given a representation of a small world and a representation of the small world as the attitude holder takes it to be (which may or may not be accurate).[5] The network's predictions were expressed in a probability within [0,1] that the target ascriptions are true. The artificial language in which these ascriptions were formulated and which the neural network learned can be interpreted as a fragment that describes propositions about the relative locations of two objects to each other plus the attitude taken towards these propositions. The small world can be conceived of as a 3-by-3 grid containing 3 objects. All objects differ in shape and they sometimes differ in colour. A typical statement in the artificial language can be glossed as *contra red triangle above blue square*, so long as we bear in mind that the network lacks any real world knowledge about triangles, squares, etc.

To encode this artificial language as well as the mind and world representations, we used a Transformer encoder from the pyTorch library. Transformers, based upon the so-called attention mechanism that allows contextualised processing of word information, are the foundation of current state-of-the-art results in natural language processing (Vaswani et al., 2017; Devlin et al., 2019; Rogers, Kovaleva, and Rumshisky, 2020). Our Transformer encoder used position embeddings and sequence embeddings to encode word order and distinguish the three types of input (attitude ascription, world representation, mind representation).

Generally speaking, the results of our experiment show the Transformer-encoder to perform better on factives than contrafactives. While the performance on non-factives was even worse, this is to be expected both from our intuitive motivation and the architecture of our network.

[4] The code of the model and further information are available on David Strohmaier's (2022) GitHub.

[5] Since our intuitive motivation did not touch on ascriptions in entailment-cancelling environments, we did not train the network to handle such ascriptions. Thus, our network does not model the presupposition projection of factive and contrafactive attitude ascriptions. We plan to fill this gap in a follow-up experiment.

Our network always processes both a mind and a world representation, although the latter only contributes noise in the case of non-factives. We suspect that, as humans learn non-factives, they develop a better input-gating mechanism than our model, which would increase their performance on non-factives. Below we focus on results that bear on the comparison between factives and contrafactives.

We evaluated 51 hyperparameter settings in an initial search.[6] Of these 18 performed below 60% accuracy and 27 exceeded 90% accuracy, suggesting that the network generally is able to learn our attitude verbs. In all except 4 of the 51 settings, the accuracy was higher for the factive than the contrafactive (but not significantly so, according to a Kruskal-Wallis test). We take this to suggest that the difference is due to the neural architecture rather than specific hyperparameter settings.

kind	accuracy	MAE
contrafactive	97.6%	0.0296
factive	97.8%	0.0255
non-factive	96.8%	0.0368

Table 1: Accuracy and MAE

The setting which performed best in the hyperparameter search, i.e. the one with the highest overall accuracy, was then applied to a hold-out test set. The results on this test set once again showed higher performance for the factives than the contrafactives. The difference in accuracy was small (0.2 percentage point, see table 1), because the model was trained on such a large sample of data (633981 examples) that it was able to successfully learn all attitude verbs.

Looking at accuracy, however, discards some information, since for an ascription that is true (not true), a prediction (not) above 0.5 is treated as accurate. By contrast, mean absolute error also considers how far the prediction strayed from the correct values of 0 (not true) and 1 (true). The differences for the mean absolute error are still small, but more striking (see table 1). A permutation significance test (resamples=9999) shows that the error is significantly larger for contrafactives than factives ($p < 0.01$), see figure 1. The training for the factive also proceeded faster than for the contrafactive, see figure 2, providing further support for a difference in how hard it is to learn the meaning of a contrafactive as opposed to factive. To give some numbers for intuition, after 100000 training examples the average loss for the factive is 0.39, while the average loss for the contrafactive is 0.54.

Post-experimental analysis suggests that many of the remaining errors by the network have the following source: The network struggles with evaluating a sentence like *contra red triangle above blue square* if the world or mind representation contains a red triangle next to a blue square, rather than one of the objects being missing altogether. The network was paying excessive attention to whether objects named in the artificial language sentences were present, ignoring whether the described relationship between the objects held. Put differently, the network struggles with reading the spatial relations from the linear enumeration of the 3-by-3 grid's 9 cells. This can be interpreted as a difficulty of dealing with word order, which is well-documented for Transformer models in the NLP literature (e.g. Pham et al., 2021).

To complicate the situation further, the target ascriptions differ in how their truth depends on the presence of objects. Notably, the factive attitude ascription can be true only if both objects mentioned (e.g. red triangle and blue square) are present in the grid, while the contrafactive ascription can be true regardless of whether both, only one, or neither of the two

[6] A list of available hyperparameters can be found in our online appendix on GitHub (Strohmaier, 2022).

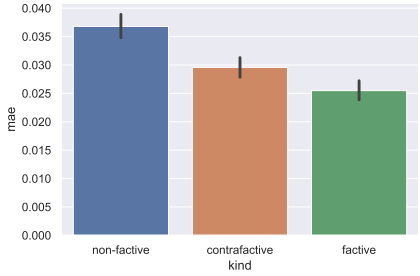

Figure 1: Mean absolute error on test set

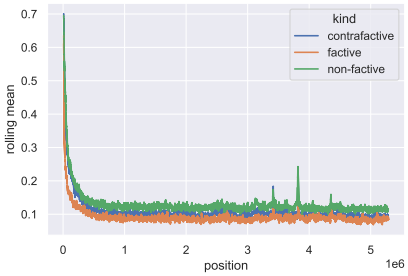

Figure 2: Rolling loss smoothed over 10000 instances during training

objects are present.[7] Given these findings, we are conducting a follow-up experiment in which the artificial language the neural network learns is simpler. The language we use here is interpretable as a fragment that merely describes a primitive, non-decomposable proposition and the attitude taken toward that proposition.

Our computational experiment improves on similar ones conducted by Steinert-Threlkeld (2019) and Steinert-Threlkeld and Szymanik (2019; 2020) in a number of ways. First, we report results based on a larger range of hyperparameters (e.g. training epochs, learning rate, etc.). For example, Steinert-Threlkeld and Szymanik (2019) only report the results in the case of two layers of LSTM cells and a hidden dimensionality of 12. We have explored 51 hyperparameter settings which can vary both in dimensionality and number of layers, among other hyperparameters, and can report that for all but four the accuracy for factives is higher than for contrafactives. This provides a better sense of the robustness of our experimental results. Second, while the cited research used feed-forward neural networks and LSTMs, we switched to the more advanced Transformer-architecture. Recent results suggest that, despite not being originally designed for cognitive plausibility, Transformer-based networks show greater convergence with human processing than other approaches (e.g. Caucheteux and King, 2022; Schrimpf et al., 2021). Given this, the results of our computational experiment likely reflect learnability for human language learners more closely than previous work.

4 Conclusion

Factives are abundant, contrafactives scarce in natural languages. We suggested that this asymmetry is partly due to a difference in learnability: the meaning of a contrafactive is significantly harder to learn than that of a factive. To support our suggestion, we reported the results of a computational experiment.

In closing, let us emphasize the scope of our discussion. We take our computational experiment to highlight one reason for the difference in frequency between factives and contrafactives in natural languages. But, most likely, this is not the only reason. We mentioned one other likely reason in section 2: the pragmatic syntactic difference between factive and contrafactive attitude ascriptions. We plan to explore this and other reasons in future work.

[7] This corresponds to the constraint on factive, but not contrafactive, attitude ascription noted earlier: the former is true only given a perfect match between world and mind, the latter can be true so long as there is no perfect match. So, there are many more ways for the latter to be true than for the former.

References

Anvari, Amir, Mora Maldonado, and Andrés Soria Ruiz (2019). "The puzzle of Reflexive Belief Construction in Spanish". In: *Proceedings of Sinn und Bedeutung* 23.1, pp. 57–74. DOI: 10.18148/sub/2019.v23i1.503. URL: https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/503.

Caucheteux, Charlotte and Jean-Rémi King (2022). "Brains and algorithms partially converge in natural language processing". In: *Communications Biology* 5.1, p. 134. DOI: 10.1038/s42003-022-03036-1. URL: https://www.nature.com/articles/s42003-022-03036-1 (visited on 08/15/2022).

Cheung, Him, Hsuan-Chih Chen, and William Yeung (2009). "Relations between mental verb and false belief understanding in Cantonese-speaking children". In: *Journal of Experimental Child Psychology* 104.2, pp. 141–155. DOI: 10.1016/j.jecp.2009.05.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022096509001155 (visited on 11/03/2021).

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. DOI: 10.48550/arXiv.1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 11/27/2022).

Glass, Lelia (2022). "The Negatively Biased Mandarin Belief Verb yǐwéi*". In: *Studia Linguistica* n/a.n/a. DOI: 10.1111/stul.12202. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/stul.12202 (visited on 08/15/2022).

Goddard, Cliff (2010). "Universals and Variation in the Lexicon of Mental State Concepts". In: *Words and the Mind: How words capture human experience*. Oxford: Oxford University Press. URL: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195311129.001.0001/acprof-9780195311129-chapter-5 (visited on 03/11/2019).

Hacquard, Valentine and Jeffrey Lidz (2022). "On the Acquisition of Attitude Verbs". In: *Annual Review of Linguistics* 8.1, pp. 193–212. DOI: 10.1146/annurev-linguistics-032521-053009. URL: https://www.annualreviews.org/doi/10.1146/annurev-linguistics-032521-053009 (visited on 05/11/2022).

World Loanword Database (WoLD) (2009). Tech. rep. München: Max Planck digital library. URL: http://wold.livingsources.org/ (visited on 03/28/2022).

Holton, Richard (2017). "I—Facts, Factives, and Contrafactives". In: *Aristotelian Society Supplementary Volume* 91.1, pp. 245–266. DOI: 10.1093/arisup/akx003. URL: https://0-academic-oup-com.pugwash.lib.warwick.ac.uk/aristoteliansupp/article/91/1/245/3897122 (visited on 11/06/2018).

Lee, Kang, David R. Olson, and Nancy Torrance (1999). "Chinese children's understanding of false beliefs: the role of language". In: *Journal of Child Language* 26.1, pp. 1–21. DOI: 10.1017/S0305000998003626. URL: https://www.cambridge.org/core/journals/journal-of-child-language/article/chinese-childrens-understanding-of-false-beliefs-the-role-of-language/B0A84B10F7113D94BB3330563EEB4248 (visited on 11/03/2021).

Nagel, Jennifer (2017). "Factive and nonfactive mental state attribution". In: *Mind & Language* 32.5, pp. 525–544. DOI: 10.1111/mila.12157. (Visited on 11/14/2017).

Pham, Thang et al. (2021). "Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks?" In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 1145–1160. DOI: 10.18653/v1/2021.findings-acl.98. URL: https://aclanthology.org/2021.findings-acl.98 (visited on 11/27/2022).

Phillips, Jonathan and Aaron Norby (2021). "Factive theory of mind". In: *Mind & Language* 36.1, pp. 3–26. DOI: https://doi.org/10.1111/mila.12267. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mila.12267 (visited on 02/19/2021).

Phillips, Jonathan et al. (2020). "Knowledge before Belief". In: *Behavioral and Brain Sciences*, pp. 1–37. DOI: 10.1017/S0140525X20000618. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/knowledge-before-belief/B434EF04A3EA77018384EABEB4973994 (visited on 10/07/2020).

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A Primer in BERTology: What We Know About How BERT Works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866. DOI: 10.1162/tacl_a_00349. URL: https://aclanthology.org/2020.tacl-1.54 (visited on 11/27/2022).

Schrimpf, Martin et al. (2021). "The neural architecture of language: Integrative modeling converges on predictive processing". In: *Proceedings of the National Academy of Sciences* 118.45, e2105646118. DOI: 10.1073/pnas.2105646118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2105646118 (visited on 11/27/2022).

Shatz, Marilyn et al. (2003). "The influence of language and socioeconomic status on children's understanding of false belief". In: *Developmental Psychology* 39.4, pp. 717–729. DOI: 10.1037/0012-1649.39.4.717.

Steinert-Threlkeld, Shane (2019). "An Explanation of the Veridical Uniformity Universal". In: *Journal of Semantics*. DOI: 10.1093/jos/ffz019. URL: https://0-academic-oup-com.pugwash.lib.warwick.ac.uk/jos/advance-article/doi/10.1093/jos/ffz019/5683663 (visited on 12/24/2019).

Steinert-Threlkeld, Shane and Jakub Szymanik (2019). "Learnability and semantic universals". In: *Semantics and Pragmatics* 12.0, p. 4. DOI: 10.3765/sp.12.4. URL: https://semprag.org/index.php/sp/article/view/sp.12.4 (visited on 11/18/2019).

— (2020). "Ease of learning explains semantic universals". In: *Cognition* 195, p. 104076. DOI: 10.1016/j.cognition.2019.104076. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027719302495 (visited on 03/18/2022).

Strohmaier, David (2022). *Contrafactives: Exploration of a Grid World*. URL: https://github.com/dstrohmaier/contrafactives_grid_world (visited on 11/30/2022).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *31st Conference on Neural Information Processing Systems*, pp. 1–11.