Squib

Dimensions of Explanatory Value in NLP Models

Kees van Deemter
Utrecht University
Department of Information and
Computing Sciences
c.j.vandeemter@uu.nl

Performance on a dataset is often regarded as the key criterion for assessing NLP models. I argue for a broader perspective, which emphasizes scientific explanation. I draw on a long tradition in the philosophy of science, and on the Bayesian approach to assessing scientific theories, to argue for a plurality of criteria for assessing NLP models. To illustrate these ideas, I compare some recent models of language production with each other. I conclude by asking what it would mean for institutional policies if the NLP community took these ideas onboard.

1. Introduction

Much recent work in Natural Language Processing (NLP) has focused on the performance of NLP models, as measured by various intrinsic or extrinsic evaluations, neglecting some vitally important other dimensions of model quality. I argue that this practice risks building models that are *ad hoc*, that are implausible and unwieldy, that are not connected to existing insights, and that may not generalize to other domains, genres, and applications. I argue that an antidote to these risks can be found in a widely accepted view of *scientific explanation*, and can be harnessed by a Bayesian perspective on assessing scientific hypotheses and theories.

I will try to be generic, covering NLP theories and models of every kind, using the word "model" throughout; following the terminology of Sun (2008), this will include both **process models** (which aim to mimic mental processes) and the more ubiquitous **product models** (which focus on the relation between inputs and outputs, agnostic about how these outputs are produced). I start by focusing on models that are constructed with the aim of expressing, testing, and ultimately enhancing humankind's grasp of human language and language use (i.e., "NLP-as-Science"). In Section 4, I will argue that my argument can be extended to cover many practically oriented types of NLP, too.

One might think that a lot of work in NLP already focuses on explanation, because it aims for explainability (e.g., Lyu, Apidianaki, and Callison-Burch 2023; see also Ghassemi, Oakden-Rayner, and Beam 2021). This argument, however, confuses (1) explaining a natural phenomenon (e.g., an aspect of language use) and (2) explaining a

Submission received: 12 September 2022; revised version received: 6 March 2023; accepted for publication: 9 April 2023.

https://doi.org/10.1162/coli_a_00480

model (i.e., a piece of software). The difference is starkest when the model doesn't match the phenomena well. Suppose a model classifies student essays into good (pass) and bad (fail). Suppose the model has terrible performance but admirable explainability, for instance via computer-generated "rationales" that highlight essay fragments that were particularly important for each classification decision (as in Lei, Barzilay, and Jaakola 2016, for example). These rationales can be useful for a stakeholder who wonders whether to trust its decisions, or a developer wanting to improve it. Yet they cannot tell us what makes an essay good or bad, because (we assumed) the model does not know the difference.

2. Dimensions of Explanatory Value

Explanation lies at the heart of the scientific enterprise (Overton 2013; Woodward and Ross 2021; Hepburn and Andersen 2021). But what is scientific explanation? Most theorists believe that scientific explanation should involve a reduction of something *not yet known* (e.g., previously unobserved facts) to something *known*, such as an existing law or model. A famous version is the Deductive-Nomological Theory (DNT). The DNT (also known as the Deductive-Nomological Model), which originated in the 1940s and was further elaborated in Hempel and Oppenheim (1965) and Hempel (1965), asserted that an explanation of data D (e.g., two billiard balls colliding) should take the form of a model M (e.g., Newton's laws of motion) that meets two requirements: first, D should follow logically from M, and second, M should be true.

These ideas may be similar to everyday notions of explanation, yet it is worth noting that Hempel's perspective on explanation is more scientific than psychological: It does not require that an explanation should give humans the feeling that they understand the data; it does not even require that M should be intelligible to us at all.

Later offshoots of DNT, also discussed in Hempel (1965), have replaced the notions of truth and logical consequence (i.e., the "follow logically from" relation above) by their probabilistic counterparts. In particular, Hempel's Inductive-Statistical Theory (IST) requires that, firstly, D should be *probable* given M (compared to other models) and, secondly, M itself should be probable. This idea is most at home in a Bayesian conception of probability,[1] which allows probabilities to be based on any kind of information, instead of only frequencies. In what follows, I will flesh out these ideas a bit more, focusing on NLP models. I will argue that IST gives rise to three dimensions of model quality, which we call performance, indirect support, and parsimony.

1. Performance. The IST requirement that data D should be probable given a model M is often measured by the performance of the model M on D, by means of metrics such as BLEU or Moverscore (Sai, Mohankumar, and Khapra 2022; Celikyilmaz, Clark, and Gao 2020), which allow us to compare a model's predictions with a gold standard. From our present perspective, evaluations based on human judgments (van der Lee et al. 2019) and task performance are likewise varieties of performance. If model A has better performance than model B then, *other things being equal*, A has greater explanatory value (in the sense of Hempel's IST) than B, because its predictions are more reliable, at least on data drawn from the same distribution as the test data.

However, performance is best seen as a broader concept, which includes a range of ways to test a model; I will use the term **direct support** to refer to these collectively. One type of direct support is replication. Replications can vary considerably. For example,

[1] As is common in the Bayesian literature, I will use "probability" and "plausibility" interchangeably.

they may or may not be conducted in a different environment, by other researchers, and with a different data sample. Replications are important because they can reveal problems with the original experiment (Belz 2022).

Often, our ultimate aim with a model is not to shed light on just one set of data, but on a much wider class of phenomena (see also Section 4). It is therefore often necessary to test a model on data drawn from different distributions. Suppose, for example, a caption generation model in the tradition of Hodosh, Young, and Hockenmaier (2013) and Agrawal et al. (2019) is initially tested on a set of outdoor photos, showing good performance. To explore the reach of the model, one could test the model on a corpus of indoor photos, with captions in a different language, and where the captions serve a different practical purpose than in the original experiment; in the terminology of Hupkes et al. (2022), these experiments would count as cross-domain, cross-lingual, and cross-task *generalizations*. To complicate the picture, generalizations can change the model. For example, when a neural "foundation model" is fine-tuned to perform a new task, this can be considered a generalization of the model as well, and this perspective has given rise to new metrics in transfer learning and domain adaptation, which aim to assess how easy it is to transfer knowledge learned from one task to another (e.g., Tsuong et al. 2020; Tan, Li, and Huang 2021).

2. Indirect Support / Theory Support. These ideas are not always adhered to, but they are widely accepted in principle. This is not true for the second IST requirement, that the model should itself be plausible. The fact that human problem solvers often forget to consider the *a priori* probability of an idea when new evidence is considered is a well-attested finding (e.g., Bar-Hillel 1980). Bayesians have argued that when a model is assessed, the same should apply: One should consider the probability of the model not just in light of the data, but in light of everything else we know or believe (Jaynes 2003; Clayton 2022). The fallacy of "base-rate neglect"—which occurs when we fail to take the *a priori* probability of a model into account—occurs for example when Null Hypothesis Significance Testing is performed in such a way that $P(\text{Data}|\text{Hypothesis})$ is computed without also taking into account the (im)plausibility of the Hypothesis, leading to incorrect estimates of $P(\text{Hypothesis}|\text{Data})$.[2] I will call the extent to which a model is supported by our existing knowledge **indirect support**, or **theory support** (because existing theories are often involved). Many scientists accept that indirect support is important, although what role it should play can be a matter of fierce debate.[3]

Indirect support can come from many sources. For example, if a model M based on BERT (Devlin, Chang, and Toutanova 2019) is good at predicting what type of expression human speakers use, and if linguistic theory suggests that the wider linguistic context affects this choice, then the fact that BERT was constructed to handle context dependencies is indirect support for M as a model of the phenomenon in question.

Because language use is a human activity, indirect support in NLP can sometimes involve considerations of *cognitive* plausibility. Suppose, for example, two Natural Language Generation (NLG) models have similar performance, but one of them is more in line with existing insights into the human language production architecture (e.g., as per Vigliocco and Hartsuiker 2002); then that model's alignment with psychological theory lends indirect support to it if the aim of the work is to shed light on human language processing; if the aim is otherwise, the alignment is irrelevant. In the case

[2] To illustrate this point, Clayton discusses cases in which unthinking researchers reported confirmation of some highly implausible hypotheses regarding paranormal behavior (Clayton 2022, chapter 6).

[3] For instance, action-at-a-distance (where objects influence each other without "touching" each other) was viewed with suspicion for some time, but has become an accepted part of physics (see, e.g., French 2005).

study of section 3.1 we shall encounter a situation where the support for a bunch of NLP models hinges on changing insights into the limits of human rationality.

3. Parsimony. An idea that comes up frequently when scientific theories are compared is parsimony: Other things being equal, a simpler model is a better model. Parsimony is closely related to Occam's Razor and to the idea that models should be as *elegant* as possible (see, e.g., Greene [2000] for a defense).

To complicate matters, parsimony can concern different aspects of a model; accordingly, it has been motivated in different ways. (See Fitzpatrick [2022] for a survey.) Accounts that focus on the logical strength of the postulates used by the model, for example, have been defended on the grounds that because models that are more parsimonious in this sense are logically weaker than less parsimonious ones, they have a greater probability of being true: For if $M_1 \models M_2$ (e.g., because M_1 adds a postulate to M_2), then $P(M_1) \leq P(M_2)$. While this is a sensible idea, NLP models tend not to be formulated as conjunctions of postulates. Rule-based models of syntax are a case in point, with some syntacticians signing up to a version of parsimony that is based, essentially, on measuring the size of a grammar (Akmajian and Heny 1975; Brody 1995).[4]

Parsimony of neural models could focus on various aspects of the model, including its overall architecture, number of hidden layers, number of parameters set and learned, the number of training epochs, and the amount of training data. Implicitly, the idea of parsimony is acknowledged in neural practices such as induction of "causal" models (Geiger et al. 2021), knowledge distillation (Sanh et al. 2019), and pruning (Tessier 2021), where the idea is to get rid of parameters or layers that do not add to a neural model's performance. Invoking parsimony as a factor in choosing between models is less common.[5] Discussions in other sciences suggest that doing so can be risky, particularly when a complex model has better performance than its simpler competitors. Hossenfelder, for instance, has argued that parsimony has played too much of a role in discussions of string theory (Hossenfelder 2018)—for, why should nature care about simplicity? Nonetheless, the idea that parsimony enhances the value of a model is widely accepted in physics too, because unless we insist on some form of parsimony, a model could be called highly explanatory even if it was nothing more than a huge collection of isolated facts; this would be counter-intuitive because a model is of little scientific interest unless it allows us to compress data in some way or other.

A **Bayesian perspective** can help to elucidate how these dimensions relate to each other. Let D be the data on which a model M is tested, and X "everything else we know" (a common Bayesian construct). As scientists—and as engineers too, as I will argue in Section 4—we want to know the plausibility of a model M (in theory: every possible model M) in light of both X and D; that is, we want to know $P(M|D, X)$. How do we get there? Starting with the two requirements inherent in IST, the **performance** of a probabilistic model M may be based on an assessment of $P(D|M,X)$, the probability the data D would have if M and X were true. As we have seen, performance may also involve **direct support** (i.e., replications and generalizations), which is essentially Bayesian update, where n subsequent datasets $D_1, .., D_n$ are brought to bear on M, yielding $P(M|D_1, .., D_n, X)$. **Indirect support** for M is $P(M|X)$, the plausibility of M in light of everything we know (before considering D_1, .., D_n). Now Bayes' Theorem tells us that

[4] Akmajian and Heny (1975) use this example: *X liked you* is assigned the underlying form *X did like you*, because this allows one to generate tag questions (*X liked you, didn't he?*), negated sentences (*X did not like you*), and emphatic sentences (*X did like you*) using one and the same mechanism, thereby minimizing the complexity of the grammar, as measured by the number of rules in it.

[5] But see Bender et al. (2021), who criticize large models for requiring considerable power consumption.

$P(M|D,X) = P(D|M,X) * P(M|X)/P(D|X)$. Since $P(D|X)$ is the same for every model, it follows that $P(M|D,X)$ (i.e., the value we're interested in) depends on performance (i.e., $P(D|M,X)$) and indirect support (i.e., $P(M|X)$) alone. It is $P(M|X)$ that is often overlooked in NLP. **Parsimony**, finally, can be seen as the probability $P(M)$ of a model before any data or other information about the world are considered. It is a component of $P(M|X)$ but it can be examined separately. Models can be compared in terms of Solomonoff's Prior (see Solomonoff 1964; Hutter, Legg, and Vitanyi 2007; and Li and Vitanyi 2008, Chapters 4 and 5), for example; the idea is that, once all models are encoded in the same way, the parsimony of a model is a function of the length of its encoding.[6] In other words, our three dimensions can be seen as a closely knit family with roots in Bayesian as well as classical philosophy of science.

3. Case Study: Two Types of Referring Expressions Generation

To illustrate both the usefulness and the pitfalls of assessing NLP models in terms of our three dimensions, I examine two types of referring expression generation (REG). I choose REG because referring is an essential part of human communication that has been studied from many different angles, using very different types of models. The performance of REG models has been tested extensively, and the outcomes of these tests will inform our discussion of the explanatory value of these models. We discuss the two types of REG one by one, then we reflect on some lessons learned (Section 3.3).

3.1 Generating Referring Expressions in a Visual Domain

REG has been studied intensively in NLG (Dale 1989; Dale and Reiter 1995; Krahmer and van Deemter 2012; Yu et al. 2016; Luo and Shakhnarovich 2017) and elsewhere (van Deemter 2016). A dominant research question in this area is, given a visual scene composed of objects, and without any linguistic context, what properties do human speakers use when they refer to one of the objects in the scene; for example, when they call an object "the ball that is red", they express the properties *ball* and *red*. Here we concentrate on models that emerged from controlled experiments involving artificial scenes whose objects have well-understood properties (shapes, colors, sizes, etc.) that can be manipulated precisely by the experimenter and presented to participants on a computer screen. Such experiments trade away some of the complexity of real-world scenes to allow a maximum of experimental control.

We compare two models. One is an application (which I will call RSA-REG) of Frank and Goodman's highly general Rational Speech Act (RSA) model (Frank and Goodman 2016, 2012).[7] RSA is a formalization of the Gricean idea that communication is always optimally rational; RSA-REG interprets this as meaning that a speaker model should emphasize efficiency: The probability that a property is chosen for inclusion in a referring expression is proportional to its discriminatory power (i.e., to the proportion of scene objects to which the property does *not* apply). As a result, RSA-REG tends to favor referring expressions that are efficient, and hence "rational".

The other model grew out of a research tradition associated with the notion of Bounded Rationality, which is skeptical about the idea that speakers routinely compute

[6] A related idea is Minimal Description Length (Solomonoff 1964; Gruenwald 2007; Voita and Titov 2020).

[7] The mechanisms of Degen et al. (2020) could lend RSA-REG better performance, but until a performance assessment of the resulting model is available, RSA-REG will serve our illustrative purposes.

discriminatory power for all the properties they consider for inclusion in their referring expressions. A well-known version of this experimentally well-supported idea (see, e.g., Belke and Meyer 2002) is the Incremental Algorithm of Dale and Reiter (1995), which assumes that some properties are intrinsically more "preferred" than others, and hence used more liberally. New findings (e.g., Koolen et al. 2011; Van Gompel et al. 2019) led to a new model in this tradition, called PRobabilistic Over-specification (PRO). PRO combines elements of the Incremental Algorithm with discriminatory power and a separate mechanism for over-specification. We compare PRO with RSA-REG.

Comparison 1: Performance. Van Gompel et al. (2019) reported an experiment in which the PRO model outperformed the other algorithms in terms of the human-likeness of their output. As for direct support, algorithms in the bounded rationality tradition have often been tested (Gatt and Belz 2010), but direct support for RSA-REG does not yet reach the level of the other models.

Comparison 2: Indirect Support. At first sight, there is much indirect support for RSA, given the intuitive plausibility of describing behavior as rational. On the other hand, behavioral economists have shown that human decision makers are affected by time and memory limitations that necessitate shortcuts (Elster 1983; Simon 1991; Gigerenzer and Selten 2002; Gershman, Horvitz, and Tenenbaum 2015) and many other deviations from rationality (Kahneman and Tversky 2013). REG evaluation experiments are broadly in line with these ideas (van Deemter 2016; Van Gompel et al. 2019); consequently, it may be argued that PRO matches theoretical results better than RSA-REG, and hence has better indirect support. This debate, however, is still ongoing.

Comparison 3: Parsimony. RSA-REG can be summarized in just two simple equations; by contrast, PRO's pseudo-code needs about a page. Since the two models are otherwise similar, it seems fair to say that RSA-REG is more parsimonious than PRO.

3.2 Generating Referring Expressions in Context

REG-in-Context is another well-studied area of NLG. It focuses on co-reference in discourse. It often starts from texts in which all referring expressions (REs) have been blanked out. The task is to predict, for each of these blanks, what RE should fill it. Other than the identity of the referent, the main information for the model to consider is the sentences around the RE, because this guides the choice between pronouns, proper names, and descriptions. The entities mentioned in the text play a role similar to the objects displayed on a computer screen in the previous section.

A long tradition of linguistic research has led to theories such as accessibility theory (Ariel 1990), the givenness hierarchy (Gundel, Hedberg, and Zacharski 1993), and Centering Theory (Brennan 1995). These theories emphasize the effect of the recency of the antecedent (e.g., in terms of the number of intervening words), its animacy (animate/non-animate), and the syntactic structure of the sentences (e.g., does the RE occur in the same syntactic position as the antecedent?) Computational accounts can be classified in terms of whether they use (1) handwritten rules, (2) hand-coded features and Machine Learning, or (3) an End2End neural architecture.

A wide range of models were recently compared in terms of their performance on this task (Same, Chen, and Van Deemter 2022). Models included (1) two rule-based ones, RREG-S (small) and RREG-L (large); (2) two models based on traditional Machine Learning (ML), called ML-S (small) and ML-L (large); and (3) three neural models, including two from Cunha et al. (2020) and one from Cao and Cheung (2019).

Comparison 1: Performance. Performance figures on these models have been reported, when feature-based models were tested in the GREC evaluation campaign (Belz et al.

2009), on a corpus of Wikipedia texts; neural models on Ferreira et al.'s (2018) version of the WebNLG corpus. As for direct support, Same, Chen, and Van Deemter (2022) tested each of these models on WSJ, the Wall Street Journal portion of the OntoNotes corpus (Gardent et al. 2017), arguing that the texts in WebNLG were too short to tell us much about referring expressions *in context*. In this generalization experiment, ML-L outperformed all other models, thereby diminishing the credentials of neural models of REG-in-Context while boosting those of feature-based models.

Comparison 2: Indirect Support. Indirect support varied widely across models, with larger models receiving the most support from the linguistics literature. RREG-L, for instance, rests on notions such as local focus (Brennan 1995) and syntactic parallelism (Henschel, Cheng, and Poesio 2000); ML-L makes use of the grammatical role of the RE.

Comparison 3: Parsimony. We have seen that parsimony can focus on different aspects of a model. The present set of models are alike in most respects but, as observed in Same, Chen, and Van Deemter (2022), the two rule-based models only have the current and previous sentence available to them as input; the two ML-based models look at the current and all previous sentences; the three neural models have the entire text available to them. The two "large" models, RREG-L and ML-L, contain more features than their smaller counterpart and are consequently less parsimonious.

3.3 Lessons from This Case Study

Our case study shows that the dimensions of model quality proposed in Section 2 can help one think clearly about NLP models, and to compare them fairly with each other. We saw how indirect support may or may not go hand in hand with superior performance. And although our first case study suggested a trade-off between parsimony and performance, in which an improvement in the latter was "bought" by sacrificing the former, the second study shows this is not always the case. It seems to me that attention to these dimensions is particularly beneficial in a case like that of indirect support, in Section 3.1, where it was debatable which model is most in line with current theories of rationality; such debates are important to have, and a focus on our three dimensions would stimulate such debates. Some challenges have come to the fore as well:

Performance. When judging direct support for a model, younger models (such as RSA-REG in Section 3.1) tend to be harder to judge, because they have been subjected to less scrutiny than older ones. Furthermore, when a model is subjected to new tests, for instance during generalization, it may be modified in the process. For example, although the PRO model (Section 3.1) incorporates many aspects of its predecessors, it is a new model nonetheless. So ultimately, perhaps the research community's focus of assessment should not be an individual model but the wider research programme of which it is a part, an idea that goes back to Lakatos and Musgrave (1970).

Indirect support for neural models can be debatable. These models are difficult to link with theoretical insights (Kambhampati 2021), at least unless they are combined with probing; (for probing in REG, see Chen, Same, and van Deemter 2021). On the other hand, neural models may be more cognitively plausible for being inspired by our knowledge of the brain; rather than either blithely rejecting or accepting this argument, this is an "indirect support" consideration whose validity deserves to be investigated rigorously, which is not often done yet (though see Ritter et al. 2017 and Momennejad 2022).

Parsimony. Comparisons of parsimony across different types of models can be problematic. For example, whereas traditional models tend to address one NLP task, neural “foundation” models such as BERT are adaptable to a wide variety of tasks, which makes a direct comparison with “single-task” models arguably unfair.

4. Conclusion

There is more to a model than performance; interpretability, novelty, and applicability, for example, are widely recognized as important considerations. I have argued that another set of dimensions is likewise important; they emerge naturally from Hempel’s theory of *scientific explanation*, and they are aligned with Bayesian thinking about assessing our beliefs and theories. I have argued that these ideas do not necessarily correspond to explanation in the psychological sense (e.g., as in Lombrozo 2006).

I have focused on models built for enhancing our grasp of language. It may be thought that my arguments are irrelevant to applied NLP models where, allegedly, performance is everything. However, the science and engineering aspects of NLP are thoroughly intertwined (as Ekbia [2008] argued about Artificial Intelligence in general), and some models that started out to solve a practical task were later studied as putative theories of human behavior (as I argued about REG models in van Deemter 2016).

To see how intertwined the two types of NLP are, let’s return to an example from Section 2. In order to couple a given collection of outdoor photos with useful text captions, a company decides to construct a model of human-produced image captions. Performance on the company’s photo collection is clearly paramount. Certain kinds of parsimony may be relevant (e.g., where they impact the time required for model training); replication of evaluation experiments will reduce the probability of error; but otherwise, direct support is irrelevant, and indirect support, too.

But if a company spends precious resources to construct a model, it may hope that the model keeps performing even as new photos are added to, or removed from, the collection. If the company is farsighted, it may therefore design its model not solely for the original dataset but, ideally, for all possible collections of the same kind (e.g., comprising both indoor and outdoor photos). Instead of focusing on one dataset, the company will thus target a far wider class of phenomena. If this happens, its work will start to resemble that of a scientist, in which case the argumentation that we put forward in favor of three dimensions of model quality applies to it.

Researchers in NLP and other areas of Artificial Intelligence should learn to care about the dimensions discussed, without trying to reduce them to a monolith. We should learn to say, and elaborate on, things like, *“Model A has similar performance to B. Being newer than B, A has lower levels of direct support. However, A is more parsimonious and appears to have better indirect support than B.”* How the different dimensions should be weighed depends on the context in which the model is assessed. But when faced with a model, readers should at least be *cognizant* of its qualities in terms of the dimensions discussed, because this will help them decide how to use it, how much trust to have in it, both now and in future, and how to investigate the model further.

Different policy mechanisms may help our community to achieve this, similar to (or even as a component of) the *model cards* that are starting to be used to enhance the documentation of NLP models in terms of their intended use, evaluation details, and so on (Mitchell et al. 2019). Funding agencies should ask proposers to comment on the direct support, parsimony, and indirect support (*alias* theory support) of the models they propose to develop; this would be analogous to asking proposers to discuss the economic impact of their plans, as is often done. Analogous to the currently non-committal limitations paragraphs solicited by some conferences and journals such as *Computational Linguistics* should encourage authors to discuss the above quality dimensions for any models proposed in their articles (see the **Appendix**); which of these dimensions are most relevant in a given case will be for readers to decide.

5: Appendix: Reporting on the Explanatory Value of a Model

When discussing a model in light of the dimensions of explanatory value discussed in this article, the following guidelines may be helpful. The three dimensions are preceded by some preliminary information about the type and scope of the model, and the aims behind it. Where appropriate, the model should be compared with other models in the public domain.

Guidelines for reporting on the explanatory value of an NLP model

Preliminary information *about the model. Please specify what task(s) the model performs. Specify the* **type** *of model (neural, rule based, classic machine learning, etc.); if the model is neural, please specify the type of model, making it clear whether the model is pre-trained, finetuned, or otherwise. Please specify the* **scope** *of the model in terms of its intended domain(s) and text genre(s). Then specify the* **aims** *of the model, for example, whether it aims to make predictions, to elucidate the process of human language processing, or otherwise.*

Performance. *Please start by summarizing the most relevant evaluation results:*

1. *Please say briefly in what way(s) the model was evaluated, and summarize the results. If any* **replications** *were done, then discuss these as well, indicating the type of replication (see e.g., Belz 2022); please highlight any differences between the different evaluation experiments in terms of their experimental setup, and in terms of the results obtained.*
2. *If the model is a* **generalization**, *please specify what type of generalization (e.g., cross-domain, cross-lingual, cross-task), and what was the relationship between the original model and your generalization. If the model was modified, how was this modification performed?*

Indirect Support. *Should the model be regarded as inherently plausible (or inherently implausible) in light of common sense, theoretical insights, or a body of previously reported results? If so, please explain, and make sure to add references to the literature where these will help readers to assess your claims.*

Parsimony. *Please comment on how you would rate the parsimony of the model, addressing the simplicity or elegance of the model (or the lack thereof). For instance, for a rule-based model you could consider the complexity of the rule set; for classic machine learning, you could consider the amount of training data and the number of features; for a neural model, you could consider the amount of training data, the number of hidden layers, the number of parameters, the number of training epochs, and so on.*

Acknowledgments

I thank the reviewers and editor of *Computational Linguistics*, Guanyi Chen, Robert Dale, Mehdi Dastani, Yupei Du, Rembert Duine, Juup Essers, Albert Gatt, Roger van Gompel, Emiel Krahmer, Chenghua Lin, Pablo Mosteiro, Dong Nguyen, Paul Piwek, Ehud Reiter, Susanne

Riehemann, Fahime Same, Arno Siebes, and Anna Wegmann for discussions. I thank the audiences of the International Workshop on AI and Cognition (Örebro, Sweden 2020) and of the First Conference on Natural Language Generation and Intelligent Writing (Suzhou, China 2021).

References

Agrawal, Harsh, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957. https://doi.org/10.1109/ICCV.2019.00904

Akmajian, Adrian and Frank Heny. 1975. *Introduction to the Principles of Transformational Syntax*. MIT Press.

Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. Routledge.

Bar-Hillel, M. 1980. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Belke, Eva and Antje S. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, 14(2):237–266. https://doi.org/10.1080/09541440143000050

Belz, A. 2022. A metrological perspective on reproducibility in NLP*. *Computational Linguistics*, 48:1125–1135. https://doi.org/10.1162/coli_a_00448

Belz, Anja, Eric Kow, Jette Viethen, and Albert Gatt. 2009. Generating referring expressions in context: The GREC task evaluation challenges. In *Proceedings of ENLG 2009*, pages 294–327. https://doi.org/10.1007/978-3-642-15573-4_15

Bender, E., T. Gebru, A. McMillan-Major, and M. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT21*, pages 610–623. https://doi.org/10.1145/3442188.3445922

Brennan, Susan E. 1995. Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167. https://doi.org/10.1080/01690969508407091

Brody, Michael. 1995. *Lexico-logical Form*. MIT Press.

Cao, Meng and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3163–3172. https://doi.org/10.18653/v1/D19-1312

Celikyilmaz, Azli, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chen, Guanyi, Fahime Same, and Kees van Deemter. 2021. What can neural referential form selectors learn? In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 154–166.

Clayton, A. 2022. *Bernouilli's Fallacy*. Columbia University Press.

Cunha, Rossana, Thiago Ferreira, Adriana Pagano, and Fabio Alves. 2020. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In *Proceedings of the 28th International Conference on Computational Linguistics (ACL-2020)*, pages 2261–2272. https://doi.org/10.18653/v1/2020.coling-main.205

Dale, Robert. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89)*, pages 68–75. https://doi.org/10.3115/981623.981632

Dale, Robert and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263. https://doi.org/10.1207/s15516709cog1902_3

Degen, Judith, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2020. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, 127(4):591. https://doi.org/10.1037/rev0000186, PubMed: 32237876

Devlin, J., M. Chang, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ekbia, H. R. 2008. *Artificial Dreams: The Quest for Non-Biological Intelligence*. Cambridge

University Press. https://doi.org/10.1017/CBO9780511802126

Elster, Jon. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. MIT Press. https://doi.org/10.1017/CBO9781139171694

Ferreira, Thiago Castro, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176. https://doi.org/10.18653/v1/W18-6521

Fitzpatrick, Simon. 2022. Simplicity in the philosophy of science. In *Internet Encyclopaedia of Philosophy*, ISSN 2161-0002.

Frank, Michael C. and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998. https://doi.org/10.1126/science.1218633, PubMed: 22628647

Frank, Michael C. and Noah Goodman. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829. https://doi.org/10.1016/j.tics.2016.08.005, PubMed: 27692852

French, Steven. 2005. Action at a distance. In Edward N. Zalta, editor, *Routledge Encyclopedia of Philosophy*, Routledge.

Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 179–188. https://doi.org/10.18653/v1/P17-1017

Gatt, Albert and Anya Belz. 2010. Introducing Shared Tasks to NLG: The TUNA shared task evaluation challenges. In Emiel Krahmer and Mariet Theune, editors, *Empirical Methods in Natural Language Generation*. Springer. https://doi.org/10.1007/978-3-642-15573-4_14

Geiger, Atticus, Hanson Lu, Thomas F. Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*.

Gershman, S. J., E. J. Horvitz, and J. B. Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 49:273–278. https://doi.org/10.1126/science.aac6076, PubMed: 26185246

Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Health*, 3:745–750. https://doi.org/10.1016/S2589-7500(21)00208-9, PubMed: 34711379

Gigerenzer, Gerd and Reinhard Selten. 2002. *Bounded Rationality*. MIT Press. https://doi.org/10.7551/mitpress/1654.001.0001

Greene, B. 2000. The elegant universe: Superstrings, hidden dimensions, and the quest for the ultimate theory. *American Journal of Physics*, 68(2):199–200. https://doi.org/10.1119/1.19379

Gruenwald, Peter. 2007. *The Minimum Description Length Principle*. MIT Press. https://doi.org/10.7551/mitpress/4643.001.0001

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307. https://doi.org/10.2307/416535

Hempel, C. G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press.

Hempel, Carl G. and Paul Oppenheim. 1965. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175. https://doi.org/10.1086/286983

Henschel, Renate, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *Proceedings of the 18th Conference on Computational Linguistics-Volume 1*, pages 306–312. https://doi.org/10.3115/990820.990865

Hepburn, Brian and Hanne Andersen. 2021. Scientific method. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Hodosh, Micah, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899. https://doi.org/10.1613/jair.3994

Hossenfelder, Sabine. 2018. *Lost in Math; How Beauty Leads Physics Astray*. Basic Books.

Hupkes, D., M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, Ch. Christodoulopoulos, K. Lasri, N. Saphra, A. Sinclair, D. Ulmeru, F. Schottmann, K. Batsuren, K. Sun, K. Sinha, L. Khalatbari, M. Ryskina, R. Frieske, R. Cotterell, and Z. Jin. 2022. State-of-the-art generalisation research in NLP: A taxonomy and review. *arXiv preprint arXiv:2006.14799*.

Hutter, M., S. Legg, and P. M. B. Vitanyi. 2007. Algorithmic probability. *Scholarpedia*, 2(8):2572. Revision #151509. https://doi.org/10.4249/scholarpedia.2572

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790423

Kahneman, D. and A. Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making*. World Scientific. https://doi.org/10.1142/9789814417358_0006

Kambhampati, Subbharao. 2021. Polanyi's revenge and AI's new romance with tacit knowledge. *Communications of the ACM*, 64(2):31–32. https://doi.org/10.1145/3446369

Koolen, Ruud, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250. https://doi.org/10.1016/j.pragma.2011.06.008

Krahmer, Emiel and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218. https://doi.org/10.1162/COLI_a_00088

Lakatos, I. and M. Musgrave. 1970. *Criticism and the Growth of Knowledge*, Cambridge University Press. https://doi.org/10.1017/CBO9781139171434

Lei, Tao, Regina Barzilay, and Tommi Jaakola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117. https://doi.org/10.18653/v1/D16-1011

Li, M. and P. M. B. Vitanyi. 2008. *An Introduction to Kolmogorov Complexity and its Applications. Third edition*. Springer. https://doi.org/10.1007/978-0-387-49820-1

Lombrozo, Tania. 2006. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470. https://doi.org/10.1016/j.tics.2006.08.004, PubMed: 16942895

Luo, Ruotian and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7102–7111. https://doi.org/10.1109/CVPR.2017.333

Lyu, Qing, Marianna Apidianaki, and Chris Callison-Burch. 2023. Towards faithful model explanation in NLP: A survey. *arXiv preprint arXiv:2209.11326*.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229. https://doi.org/10.1145/3287560.3287596

Momennejad, Ida. 2022. A rubric for human-like agents and neuroAI. *Philosophical Transactions B*, 378(20210446). https://doi.org/10.1098/rstb.2021.0446, PubMed: 36511409

Overton, James A. 2013. "Explain" in scientific discourse. *Synthese*, 8(190):1383–1405. https://doi.org/10.1007/s11229-012-0109-8

Ritter, Samuel, David G. T. Barrett, Adam Santoro, and Matt M. Botvinick. 2017. Cognitive psychology for deep neural networks: A shape bias case study. *arXiv preprint arXiv:1706.08606*.

Sai, Ananya B., Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys*, 55(2). https://doi.org/10.1145/3485766

Same, Fahime, Guanyi Chen, and Kees Van Deemter. 2022. Non-neural models matter: A re-evaluation of neural referring expression generation systems. In *Proceedings of ACL 2022*, pages 5554–5567. https://doi.org/10.18653/v1/2022.acl-long.380

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Simon, Herbert. 1991. Bounded rationality and organizational learning. *Organisational Science*, 2:125–134. https://doi.org/10.1287/orsc.2.1.125

Solomonoff, Ray J. 1964. A formal theory of inductive inference: Part I. *Information and Control*, 7(1):1–22. https://doi.org/10.1016/S0019-9958(64)90223-2

Sun, Ron. 2008. *The Cambridge Handbook of Computational Psychology*. Cambridge University Press.

Tan, Yang, Yang Li, and Shao-Lun Huang. 2021. OTCE: A transferability metric for cross-domain cross-task representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 15779–15788. https://doi.org/10.1109/CVPR46437.2021.01552

Tessier, Hugo. 2021. Neural network pruning 101. https://towardsdatascience.com/neural-network-pruning-101-af816aaea61

Tsuong, Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. 2020. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7294–7305.

van Deemter, Kees. 2016. *Computational Models of Referring: A study in Cognitive Science*. MIT Press. https://aura.abdn.ac.uk/handle/2164/18498

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368. https://doi.org/10.18653/v1/W19-8643

Van Gompel, Roger P. G., Kees van Deemter, Albert Gatt, Rick Snoeren, and Emiel Krahmer. 2019. Conceptualization in reference production: Probabilistic modeling and experimental testing. *Psychological Review*, 126(3):345. https://doi.org/10.1037/rev0000138, PubMed: 30907620

Vigliocco, G. and R. J. Hartsuiker. 2002. The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin*, 3(128):442–472. https://doi.org/10.1037/0033-2909.128.3.442, PubMed: 12002697

Voita, Elena and Ivan Titov. 2020. Information-theoretic probing with minimum description length. *arXiv preprint arXiv:2003.12298*. https://doi.org/10.18653/v1/2020.emnlp-main.14

Woodward, James and Lauren Ross. 2021. Scientific explanation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Springer. pages 264–293.

Yu, L., P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. 2016. Modeling context in referring expressions. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, pages 69–85. https://doi.org/10.1007/978-3-319-46475-6_5