ARTICLE REVIEW: Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*.

For decades now, a number of scientists across different fields have researched the question: how does communication between human beings work? Indeed, the fact that humans are able to communicate at all is in itself an astounding phenomenon: how do we do it, despite the ambiguity present in all languages, the limited memory capacity of our brains and, just to cite another, our ever-incomplete knowledge, of the interlocutor's intentions and of the state of the world? (8)
In the realm of cognitive modelling, the answer to this question divides the field between proponents of the *symbolic* approach and those who adhere to the *connectionist* approach. While symbolists agree that there exists some shared concept of *rationality* or *optimality* in humans (3), connectionists believe these structures simply arise from statistical regularities in our shared experiences of the world.

In their review, Goodman and Frank summarized the 2016 state-of-the-art in structured probabilistic Bayesian models for language interpretation: the Rational Speech Act framework (RSA). This class of symbolic computational models relies on Gricean *semantics*, i.e. the study of meaning in language developed by Grice. However, it abandons Grice's belief in maxims as the guiding principle of conversation; instead, it approaches the structure of human language by treating humans as agents who try to maximize the probability of realization of their communication goals (4).

The RSA framework views communication as a recursive game between speaker and listener, as will be discussed in Part I of this review. While Goodman and Frank present two models of language interpretation (simple RSA and uncertain RSA) both focusing mainly on the listener side, later developments in the field have seen the emergence of speaker-based models (e.g., 12).

Part II of this article review will focus on competing approaches to language understanding, summarizing the latest development in Natural Language Processing (NLP) following advances in attention-based network architectures (15), and focusing particularly on the approach proposed by McClelland et al. (9).

Later developments of RSA, including some to which Goodman and Frank themselves contributed, novel implementations and criticisms will be discussed in Part III of the review, after which only one conclusion will be drawn: as it often happens, the best attempt at explaining human cognition could turn out to be not one approach or the other, but a mixture of the two.

I: The Rational Speech Act framework and uncertainty

Grice concluded that, during a conversation, speakers and listeners follow a series of maxims - such as "*Be relevant!*", the Maxim of Relation - which, together, constitute a kind of "cooperative principle" of communication (5). The RSA framework belongs to the domain of probabilistic pragmatics; although it "*follows Grice in assigning an important role to goal-oriented, optimal behaviour*" (3), probabilistic pragmatics does not itself rely on maxims to explain communication processes. It is rather more interested in how goals and beliefs influence our inferences about meaning, shifting the focus of its symbolic approach from rules to probability.

In structured probabilistic Bayesian models, we often assume knowledge of which variables represent each hypothesis (here, each meaning) and that it is possible to represent this knowledge explicitly in a graph structure, such as the one in Figure 1 below. The core advantage of these models is that we can recursively update our beliefs, represented as probability distributions, as we gain more information. RSA models are inherently recursive (the speaker reasons about a listener, who reason about the speaker reasoning about the listener and so on…), hence we will need the formal tools of mathematics and, later, of computer science in order to fully represent them.

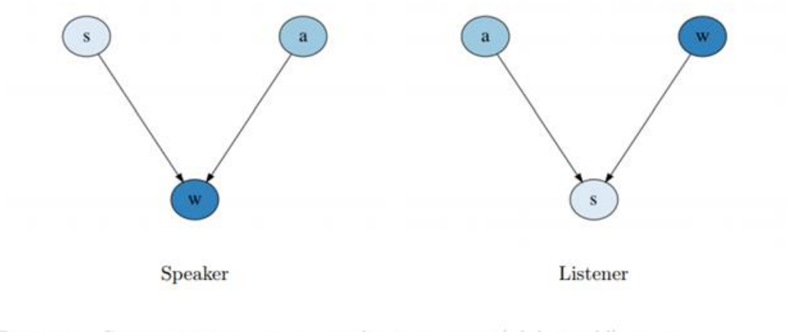


Figure 1: Simple RSA model (17)

Let us look at the equations in Figure 2: in this simple model, only two variables are considered: the state of the world (w) and the utterance made by the speaker (u). Their relation is defined by P_S (the speaker's probability of choosing a certain word to convey some meaning) and P_L (the listener's probability of choosing a certain meaning upon hearing some word), both dependent on a hypothetical literal listener probability P_{LIT}.

$$\boxed{P_L(w|u) \propto P_S(u|w)P(w).}$$

$$\boxed{P_S(u|w) \propto \exp(\alpha U(u;w)).}$$

$$\boxed{U(u;w) = \log P_{\text{Lit}}(w|u).}$$

$$\boxed{P_{\text{Lit}}(w|u) \propto \delta_{[\![u]\!](w)}P(w).}$$

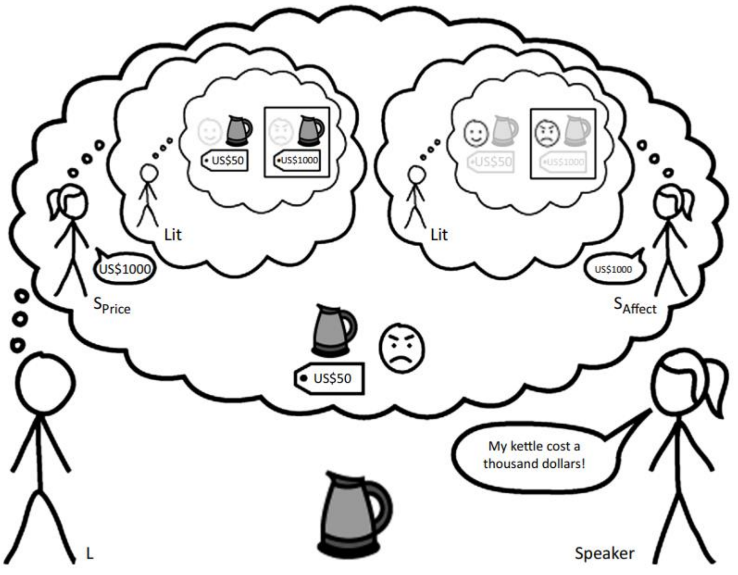
Figure 2: Simple RSA Bayesian inference model (4)

The cognitive implausibility of the underlying assumption of this simple model, i.e. the fact that a speaker always behaves as a rational agent, is extensively discussed in Goodman and Frank's review (4). It is also addressed through a small adjustment to the first equation of Figure 2, shown in Figure 3: by positing a joint inference, i.e. by making the listener infer the kind of speaker she is faced with, given the state of the world and the utterance heard, they develop what they call uncertain RSA (uRSA).

$$\boxed{P_L(w,s|u) \propto P_S(u|w,s)P(s)P(w)}$$

Figure 3: uncertain RSA model equation (4)

This framework ameliorates RSA in that it explains more linguistic phenomena, such as nonliteral and figurative language (metaphors, hyperboles, sarcasm…), as shown in Figure 4. By introducing what is usually referred to as the "question under discussion" (i.e. by iteratively tuning the model to the changing context), many other phenomena have been efficiently modelled by uRSA, such as the vagueness of adjectives (e.g. "Bob is tall" and "The Big Ben is tall" require different thresholds for the qualifier "tall") or the evolution of vocabulary patterns of existing languages over time, often called "evolutionary approach" (12). Moreover, the speaker profile that emerges from this new framework is more realistic as a cognitive agent, in that she is not restricted to one specific context and goal (4).



Trends in Cognitive Sciences

Figure 2. Uncertain Rational Speech Act-Style Reasoning Applied to Hyperbole. Listener L reasons jointly about the price of the item and the speaker's affect. In doing so, he considers two speakers, one who is primarily interested in conveying her affective response to the kettle, and one who is primarily interested in conveying the actual price. (The full model also considers speakers, not pictured, who wish to convey approximate price and combinations of these goals.) Each of these speakers is modeled as reasoning about a literal listener who interprets the utterance literally (indicated by the box selecting the 'US$1000' state), but focus on different aspects of the situation (price on the left and affect on the right).

Figure 4: Example of how hyperbole can be explained by uRSA models (4)

II: An alternative approach to language interpretation: the connectionist perspective

As our technical possibilities for developing neural networks grew, so did the appeal of the so-called *connectionist* approach, according to which human brains are nothing but a large population of massively interconnected neurons responding to the stimuli of the environment they live in (10), from which they learn higher-order structures. Language understanding, production and comprehension were thought to be too complex to be modelled by neural networks up until the late 2000s, but this quickly changed with the advent of GPUs, deep networks and big data, generating a "frenzy" over neural networks across research fields, which led to incredible advances in their architectural design.

It is not of much relevance to this review to dwell on the role played by theoretical and implementational advances (the back-propagation algorithm, PDP, Elman's networks…) that made today's neural networks so powerful. Instead, let us fast-forward to a very recent, major, breakthrough: the development of the so-called Transformer model, an attention-based architecture that revolutionized the field of NLP (i.e. the branch of AI that is concerned with giving machines the ability to understand text and spoken words in much the same way human beings can) (7).

Previous to the development of the first Transformer model, shown in Figure 5, most NLP models were trained to execute specific tasks through supervised learning. This made them unable to generalize the learned knowledge to other domains in the same way humans do. The Transformer architecture, relying on process called *self-attention* – "*an attention* *mechanism relating different positions in a single sequence in order to compute a representation of sequence*" (15) -, rather than on recurrence or convolution, is the direct ancestor of the Generative Pre-trained Transformer models (GPTs) developed by OpenAI: GPT-1 (13), GPT-2 (14) and GPT-3 (1), itself one of the most powerful NLP models ever developed with 175 billion parameters, compared to "only" 1,5 billion parameters in GPT-2.

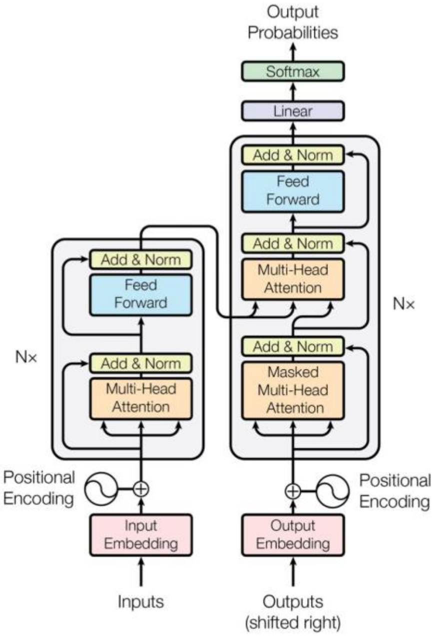


Figure 5: The Transformer - model architecture (15)

The above statistics make it hard to substantiate the claim that GPT-3 is a "few-shot learner" (i.e. only needs few training samples to learn to predict the correct output), as does the fact that its performances have been achieved after training on more than 500 billion Gb of text samples (1Gb of text being roughly equivalent to 60,000 pages). Such a model is therefore unlikely from a cognitive science perspective, while it remains true that it can deliver human-like performances and has generated highly significant computational advances in attention-based architectures.

A more interesting example with respect to cognitive plausibility is McClelland et al.'s proposition of a potential model that "*places language in an integrated understanding system*", i.e. that exploits indefinite length (and types) of context in order to produce and understand language. This model is still part of the connectionist approach to language understanding, in that it is based on "*continuous, multivalued arrays of connection weights*" between pattern vectors that map patterns (encoding language symbols) to other patterns (9), rather than on syntactic and semantic rules. Yet, the authors attempt to devise a means of solving the common shortcomings of neural networks in language tasks (e.g. inability to generalize) which would not, in principle, require training on such huge datasets. By means of a mix between Query-Based Attention methods and Long-Short-Term Memory cells, they propose to use integrated context, i.e. information about the entire "situation" (including visual stimuli, long-term memory, etc.) in which the language exchange is taking place as a way to overcome these problems (9). This seems to build a bridge between the symbolic and the connectionist approach: while it may be true that high-order cognitive abilities require an ability to reason by means of structures, the only way to determine whether these structures are acquired through education or innate is - funnily enough - to find out how whether, and how, they can be learned (10).

Part II: Later directions and criticisms of the RSA framework

Despite the rise in popularity of the connectionist approach, the RSA framework definitely did not exhaust its investigative power for language understanding in 2016. Goodman and Frank themselves have collaborated to many developments to further ameliorate it, confirming their intent to explore how RSA would model speakers that had different goals from pure information delivery. In one of their latest publications, they present a speaker-oriented RSA model that successfully explains the oddities of politeness in communication (16), depicted in Figure 6.

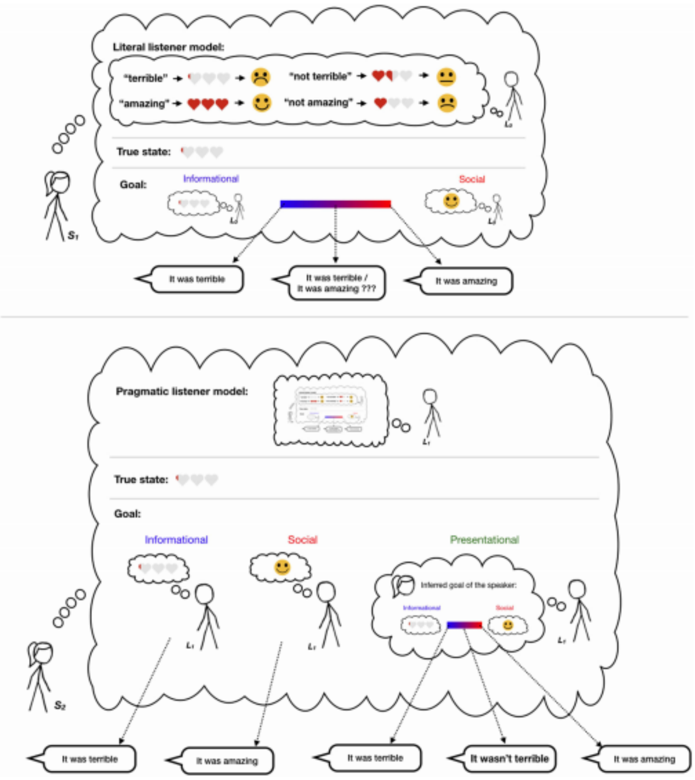


Figure 1. Diagram of the model, showing S_1 (a first-order polite speaker) and S_2 (a higher-order polite speaker capable of self-presentational goals). Top: First-order polite speaker (S_1) produces an utterance by thinking about: (1) the true state of the world (i.e., how good a given performance was), (2) the reasoning of a literal listener who updates his beliefs about the true state via the literal meanings of utterances (e.g., "not terrible" means in expectation 1.5 out of 3 hearts) and the effect associated with the state implied by the utterance, and (3) her goal of balancing informational and social utilities. Bottom: Second-order polite speaker (S_2) produces an utterance by thinking about (1) the true state, (2) the pragmatic listener L_1 who updates his beliefs about the true state and the first-order speaker S_1's goal (via reasoning about the S_1 model), and (3) her goal of balancing informational, prosocial, and self-presentational utilities. Different utterances shown correspond to different weightings of the utility components.

Figure 6: Pragmatic listener (informational, social or presentational) RSA model (16)

However, two main issues still remain. Even in the latest work cited (16), many heavy assumptions are imposed on the way that people are supposed to evaluate their utility in producing a certain utterance; in the previous example, how plausible can it be that everyone assigns a certain "politeness value" to a sentence, despite difference in cultures, contexts, etc.?

This is one instance of a very common criticism moved to the symbolic approach, that of being biased by virtue of the a priori assumptions necessary to develop, for example, RSA models. A possible, provocative, response to those criticisms was formulated a decade ago: "*When a neural-level understanding of human knowledge and its origins is eventually achieved, we predict that it will build on a deep understanding of [how rich knowledge structures can be implemented in neural circuits] at the computational level – and that this understanding will be best framed using the concepts and principles of probabilistic inference*"(6). Outside of this provocation, the Bayesian perspective in cognitive science is also supported by extensive evidence that cannot be disregarded.

The advances in probabilistic programming and in competing approaches to language understanding and production described in Part II of this review have actually enriched the potentialities of the RSA framework. For example, Goodman collaborated in the development of a model that relaxes the assumption of Boolean values for the literal meaning of words, thus strengthening the inference potential of RSA models with a similar approach to that of NLP, which treats word and sentence meanings as vectors of real numbers, rather than variables assuming binary values (2).

Computationally speaking, the review correctly predicted that "*while […] it may emerge that the RSA approach is not able to capture some aspects of language understanding […] increasingly, methods in machine learning have been used to supplement RSA with powerful learning mechanisms [and] This cross-fertilization is among the most encouraging outcomes of work on RSA*" (4). In fact, the authors themselves collaborated to developing a pragmatic neural model for language understanding (11), thus fulfilling their own prophecies.

Word count: 2005

BIBLIOGRAPHY:

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. 2020 May 28.
2. Degen J, Hawkins RD, Graf C, Kreiss E, Goodman ND. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. Psychological review. 2020 Apr 2.
3. Franke M, Jäger G. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. Zeitschrift für sprachwissenschaft. 2016 Jun 1;35(1):3-44.
4. Goodman ND, Frank MC. Pragmatic language interpretation as probabilistic inference. Trends in cognitive sciences. 2016 Nov 1;20(11):818-29.
5. Grice, H.P. Logic and conversation. In *Speech acts*. Brill. 1975; 41-58.
6. Griffiths TL, Chater N, Kemp C, Perfors A, Tenenbaum JB. Probabilistic models of cognition: Exploring representations and inductive biases. Trends in cognitive sciences. 2010 Aug 1;14(8):357-64.
7. IBM Cloud Education, 2020. *Natural Language Processing*. Available at <https://www.ibm.com/cloud/learn/natural-language-processing>. Accessed 1st February 2021.
8. Levy, R. (2018). *Foundational Architecture of human language comprehension, production and acquisition.* Talk delivered at MIT's Center for Brain, Mind and Machines on August, 1st 2018 as part of the CBMM Summer Lecture Series. Available at https://cbmm.mit.edu/video/foundational-architecture-human-language-comprehension-production-and-acquisition-11301. Accessed 14 January 2021.
9. McClelland JL, Hill F, Rudolph M, Baldridge J, Schütze H. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. Proceedings of the National Academy of Sciences. 2020 Oct 20;117(42):25966-74.
10. McClelland, J., 2020. *Are people still smarter than machines?* Distinguished seminar delivered at University of Padua on Dec. 12th, 2020.

11. Monroe, W., Hawkins, R.X., Goodman, N.D. and Potts, C., 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, *5*, pp.325-338.
12. Qing C, Franke M. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In Semantics and linguistic theory 2014 Aug 5 (Vol. 24, pp. 23-41).
13. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018.
14. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019 Feb 24;1(8):9.
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. arXiv preprint arXiv:1706.03762. 2017 Jun 12.
16. Yoon EJ, Tessler MH, Goodman ND, Frank MC. Polite Speech Emerges From Competing Social Goals. Open Mind. 2020 Nov;4:71-87.
17. Gawron, J.M. Introduction to Rational Speech Act Theory. 2019; 1:11.