Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-01643-9 OPEN

Semantic noise in the Winograd Schema Challenge of pronoun disambiguation

S. de Jager [1]✉

The Winograd Schema Challenge (WSC) of pronoun disambiguation is a Natural Language Processing (NLP) task designed to test to what extent the reading comprehension capabilities of language models (LMs) can be compared to those of human subjects. It is generally assumed across the NLP literature that human subjects are capable of resolving this task because of their acquired commonsense knowledge, thus setting a commonsense benchmark for LMs, one which has even been proposed as an alternative to the Turing test. In the context of complex natural language communications, Shannon and Weaver observed that the act of semantic interpretation is subject to *semantic noise* (Shannon and Weaver, 1964 (1949)). Semantic noise is a constraint that ensues from terms exhibiting variable interpretations across contexts, presenting a challenge to the resolution of tasks such as the WSC. However, the main argument of this paper is that rather than seeing semantic noise as a challenge to otherwise unambiguous communication, it can also be understood as a functional quality of natural language, given that it results in the conceptual negotiation of terms. Failing to theoretically attend to this linguistic matter of fact leads to unintended problems in instances where NLP applications are offered as unbiased or objectively applicable solutions. To address this, this article offers a renewed and original analysis of a series of Winograd Schemas, in order to demonstrate how they are not as straightforwardly solvable by human subjects as is commonly claimed across the NLP literature. The methodology employed is that of historical contextualisation in information theory, and qualitative cultural analysis drawing on examples from a wide variety of recent NLP literature.

[1] Erasmus School of Philosophy, Rotterdam, Netherlands. ✉email: dejager@esphil.eur.nl

Introduction

Natural Language Processing (NLP) is a particular subfield of artificial intelligence (AI) with the aim of developing language models (LMs) that are able to process linguistic inputs and produce human-like outputs. Its applications are pervasive: from chatbots and sentiment analysis to explainable AI and generative tools such as text-to-image or video. While NLP can be further subdivided into various domains, we will focus on the specific NLP task of pronoun ambiguity resolution in Winograd Schemas. Winograd Schemas (WSs) originated in Terry Winograd's (1972) dissertation *Understanding Natural Language*, where he presented the problem of anaphora: indefinite instances of natural language—in English: demonstrative pronouns in particular—where contextual reasoning is required in order to assess their specific meaning.[1] A common example of a WS would be: "the nurse helped the patient even though she was upset", where "she" is anaphoric. While Winograd recognised the semantic plurality emerging from indefinite cases, he also believed that a human subject "filters out all but the most reasonable interpretations" when presented with ambiguity (ibid., p. 31). More recently, Luciano Floridi and Massimo Chiriatti have criticised OpenAI's GPT-3[2] for its mathematical, semantic and ethical shortcomings, albeit against the tacit background of an elusive "common sense" supposedly possessed by humans. In the context of a failed semantic test, they argue that "Confused people who misuse GPT-3 to understand or interpret the meaning and context of a text would be better off relying on their common sense." (2020, p. 689). We will refer to this belief—which is found across the NLP literature, spanning both proponents and detractors, early and current—as *commonsense bias*, so termed because it is assumed to apply invariably across human subjects. This bias, upholding a general image of human agents to a largely undefined standard, seems to pervade over the field of NLP. As we will observe in our specific case, the main body of NLP literature on WSs proposes that these anaphoric cases are unambiguous to human readers, but (possibly) ambiguous to language models. This article aims to modestly elucidate how this ambiguity is just as present to human readers, and how failing to attend to this admits a variety of serious issues through the back door. The analysis of the WSC proposed here will focus on the concept of *semantic noise*: the contextually-dependant semantic variability of terms, originally presented by Warren Weaver in *The Mathematical Theory of Communication*, (TMTC, 1964 (1949)) as noise at the level of intent, interpretation and its resulting behaviour. We will observe how and why this dynamic effect is not just a challenge to the statistical distributions of "meanings" in language corpora—where meanings should in principle simply derived from local "commonsense" knowledge—but a fundamental feature of linguistic reasoning. Statistical distributions in large language models are unavoidably biased in many respects, but most of all: their functional conceptualisation in the examples we will see is biased *against* the dynamic semantic negotiations that characterise the functions of natural language as a social phenomenon.

The WSC has received considerable attention as a commonsense benchmark (Elazar et al., 2021; Levesque, 2012; Morgenstern, 2016; Sharma, 2019; Speer et al., 2017; Rahman and Ng, 2012; Wolff, 2018; Brown et al., 2020; Kocijan et al., 2022). At its inception, WS disambiguation has even been proposed by renowned AI researcher Hector Levesque as an alternative to the Turing test, because it requires the "commonsense knowledge" of a human interpreter to link which predicate coincides with which subject best (Levesque, 2011), and should be "designed so that the correct answer is obvious to the human reader, but cannot easily be found using selectional restrictions or statistical techniques over text corpora" (Levesque et al., 2012). Despite recent successes revealing high degrees of human-level interpretative capabilities (Kocijan et al., 2022), it has also been observed that disambiguation success does not reveal actual commonsense reasoning but rather language model alignment with the specific commonsense expectations of its designers (Elazar et al., 2021; Kocijan et al., 2022). What has not yet been specifically addressed and thoroughly analysed, however, is the fact pronoun disambiguation in many WS cases actually promotes highly problematic social, cultural and political assumptions about the capabilities of human subjects. It is the intention of this article to highlight some of these issues and their consequences.

In the recent, highly influential OpenAI paper "Language Models are Few-Shot Learners" (Brown et al., 2020), the authors present what we have termed *commonsense bias* quite clearly:

> "...humans do not require large supervised datasets to learn most language tasks—a brief directive in natural language [...] is often sufficient to enable a human to perform a new task to at least a reasonable degree of competence [...] To be broadly useful, we would someday like our NLP systems to have this same fluidity and generality."

But where do these assumptions rest? The different concepts employed here (*useful*, *reasonable*, *competence*, *fluidity*, *generality*) are each deserving of in-depth analyses which exceed the scope of this article, but it should suffice to say that not only do they imply vague and highly contested notions, but also a certain level of circular reasoning in how these terms relate to and define each other. On the less optimistic side of the spectrum, the "Stochastic parrots" criticism presented by Bender et al. suggests that "an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot." (2021, p. 617). While partly in agreement with this statement we also observe the following: the "reference to meaning" is semantically noisy and thus not entirely straightforward for human speakers either. As observed by Karen Spärck Jones in 2004 already, in the context of information retrieval and automated text summarisation: a human language model (LM) user cannot even be guaranteed to be able to express their needs and/or use adequate expressions even if they have a goal in mind (2004, p. 8). This last point becomes an increasingly acute symptom to attend to in the rapidly expanding landscape of generative AI technologies, where short queries and prompts guide their development and determine their complex products.

In order to consider these issues we will first analyse the problem of the pronoun disambiguation task in WSs. We will then explore the concept of *semantic noise* in order to challenge what we have identified as *commonsense bias*. The phenomenon of semantic noise will be used to interrogate assumptions admitted in the supposed resolution of WSs. As we will observe, we find a tendency towards the elimination of semantic noise in NLP, and it can be conjectured that this is a result ensuing from the "need" for cheap, technical and applied solutions (Floridi and Chiriatti, 2020). The intent in the present article is not to take away from the admirable feats of the field, but to interrogate some of its inherent biases. The main questions considered by this article are thus: What are the consequences of considering human subjects as capable of resolving a task they are not actually fully capable of resolving? And can the concept of semantic noise help identify these problems differently if we consider it a fundamental property of natural language communications and not an impediment? As stated earlier, in order to answer these questions, the methodology employed will be that of historical contextualisation in information theory, and qualitative cultural analysis which draws on examples from a wide variety of recent NLP literature.

This article intends to pick up where Bender et al. left off with their concluding remark, calling "on the [NLP] field to recognise that applications that aim to believably mimic humans bring risk of extreme harms." (Bender et al., 2021, p. 619). These harms may be as banal as the inability to determine "who was upset" in a specific utterance, to vastly reaching adverse social effects across a wide range of increasingly pervasive NLP applications.

What humans (still) can't do: Winograd Schemas

A prominent instance of ambiguity in NLP—i.e. a situation presenting the possible challenge of semantic noise—is that of WSs: sentences (in English) in which demonstrative pronouns like "it" or "they" can be related to multiple predicates. In what follows, we will analyse a series of WSs in order to question what is generally assumed about the "typical" human interpreters who are capable of resolving these challenges putatively effortlessly. According to the developers of the 2020 GPT language model, WSs are "a classical task in NLP that involves determining which word a pronoun refers to, when the pronoun is grammatically ambiguous but semantically unambiguous to a human" (Brown et al., 2020). We will challenge this claim by demonstrating that these anaphoric cases are ambiguous and semantically noisy for human interpreters as well.

Linguistic ambiguity, denotative plurality, contextual fuzziness or lack of specificity—all instances of Weaver's *semantic noise*, as we will see in the following section—are certainly a pressing issue arising from any attempt to design an inferential model the purpose of which is providing "objective" answers to (reversible) questions regarding translation, interpretation, etc. However, one could actually make the claim that when *human* agents deal with semantic noise (due to the effects of cognitive biases, sociocultural partiality, etc.) they are presented with the same (ethical, belief-updating) challenges. Let us begin our exploration of why this is so with the original, and most famous of all Winograd Schemas:

> "The city councilmen refused the demonstrators a permit because they feared violence."
>
> **or:**
>
> "The city councilmen refused the demonstrators a permit because they advocated violence."

This foundational WS is presented in Winograd's dissertation (1972, p. 33). The idea should be that in order to make an instantaneous assessment of who "they" are, a human reader reasons by elimination that "fear" is most likely experienced by the city councilmen, whereas the advocacy of violence most likely relates to the demonstrators. However, something that has yet to be remarked in the NLP literature, is the fact that depending on one's political views and, perhaps, on how many protests one has witnessed, it might very well be that one is partial to arguing for the *exact opposite*.[3] This alternative meaning is relatively more "far-fetched" but certainly not an impossible interpretation given the appropriate context. For example, the first sentence could be in the context in which demonstrators fear violent (police) retaliation, and the city councilmen rightly refuse them a permit in order to avoid said violence. Or, it could be in the context in which the demonstrators fear violence and the city councilmen *expect* them not to, which is why their permit is refused. The second sentence can be interpreted as having the meaning that the city councilmen themselves advocate violence by limiting citizens' freedoms (such as the right to protest), and therefore refuse the demonstrators a permit. As we will see, there have been many proposals for disambiguation towards *one* particular instance of meaning in this and other WSs, a quest that is still ongoing, despite claims from influential AI scholars such as Kocijan et al. who propose that the WSC has been conceptually defeated as a commonsense benchmark (2022, pp. 31–33).

In the present article, we are in full agreement with Kocijan et al. that the WSC cannot be considered a yardstick for commonsense reasoning. However, we would like to point out that while Kocijan et al. suggest that (commonsense) knowledge about "stereo-typical attitudes toward (non-state-sanctioned) violence no doubt also plays a role in the disambiguation." (ibid., p. 2), in the example of the above-mentioned WS, they also quote Levesque's (2011) proposed criteria for WSs, and devote special focus to the criterion that "Both sentences must seem natural and must be easily understood by a human listener or reader; ideally, so much so that, coming across the sentence in some context, the reader would not even notice the potential ambiguity." (ibid., p. 3). The authors fail to provide a reason for why the "natural interpretation" is, in fact, supposed to be natural. While they rightly acknowledge that the "commonsense reasoning problem remains" given that the WSC is not an adequate test of "common" sense (ibid., pp. 31–33), as well as point to the work of linguists who study the complexity of pronoun disambiguation and its sometimes impossible resolution (ibid., p. 25), they also conclude their paper with the proposal that AI language systems seem more prone to spatiotemporal commonsense reasoning errors than errors of a higher degree of abstraction (ibid., p. 33). However, as we will see in the following sections, it is not that "AI tends to stumble over basic concrete realities much more than over abstractions" (ibid.), it is that human readers interpret abstractions generatively, as words denoting concepts with potentially vast semantic possibility spaces. This inclines human readers of language system results to read these with a certain degree of charity which, we will argue, ensues from the human capacity to handle semantic noise by making meaning-estimating inferences.

Without contextual information, which is much too often hastily disregarded as (semantic) noise, the disambiguation of the WS above can in fact be said to be unsolvable. As Altaf Rahman and Vincent Ng suggest in their 2012 paper: "when given these sentences, the best that the existing [NLP] resolvers can do to resolve the pronouns is guessing" (p. 2). As it will become clear with the next examples, human interpreters resolve ambiguous schemas in a similar way: by contextualising and estimating, which cannot be said to be anything different than an inference to the best explanation, i.e. an inference based on their best and *preferred* possible guesses. Ng and Rahman address different types of schemas in their paper, and the ones that are easily resolved involve object-language situations such as "Lions eat zebras because they are predators" or "The knife sliced through the flesh because it was sharp" (ibid.). Kocijan et al. also mention that of all the WSs proposed, "The trophy doesn't fit in the brown suitcase because it's too [small/large]," has become the most prominent, standard example (2022, p. 4).[4] However, whenever abstract concepts are involved ("popular," "beautiful," "angry," etc.), the resolution of the schemas becomes increasingly difficult. The argument for the appreciation rather than the dismissal of semantic noise here is that these unstable, context-dependent concepts are constituted as they are applied. As we will see, this occurs just as much in scientific literature as it does in conversation, and can thus be said to exhibit a *dialogical*, and thus fundamentally socially-distributed character.

Taking the idea of language as context-dependent, and primarily as changing over time, Carlson et al. (2010) have suggested an interesting knowledge architecture for the *Never-Ending Language Learning* (NELL) project,[5] which is defined by the fact that it incessantly updates itself by searching the web for semantic change (new categories, associations, terms, etc.). While the *never-ending* part seems like the right approach, NELL still had the drawback that its focus remained much too grounded on

object-language descriptions, and relied on web pages as its only source, which significantly influenced the type of grammar, symbolism, slang, etc. analysed. In a paper addressing the strengths and weaknesses of the NELL approach, Mitchell et al. (2018) tackle the concept of *never-ending learning* and suggest that in order to arrive at a true understanding of machine or human learning we should have a working structure for learning across "many different types of knowledge or functions; from years of diverse, mostly self-supervised experience; in a staged curricular fashion, where previously learned knowledge enables learning further types of knowledge; where self-reflection and the ability to formulate new representations and new learning tasks enable the learner to avoid stagnation and performance plateaus" (Mitchell et al., 2018). Because of these hard-to-achieve features, they observe that *plasticity* is also an issue for NELL, given that some of its semantic structures are "set in stone" and cannot change (ibid.). This is in part due to the fact that it cannot self-reflect and "reason" about whether it lacks the correct knowledge or not, which is tied up with the issue that NELL lacks, for example, an understanding of space and time, but also with the authors' claim that an autonomous learning agent has no capacity to distinguish whether it has sufficient or correct information about something, the only thing it can do is detect whether what it "knows" is internally consistent (ibid.).

What these observations might be missing is that these limitations are perfectly applicable to the case of human learners as well, with the key that difference humans exist in a dialogical network with already-given, specific types of semantic orientations. What NELL lacks, thus, is a multi-agential perspective constrained by biases *and* capable of observing them *as* biases. Without any context with regard to why things should be ranked as less or more relevant in semantic terms, NELL lacks a determining vantage point which would make its assessments gravitate around the different cores of its belief-system. This would unavoidably frame it within specific epistemological limits, just as is the case for human beings, but it would render its assessments *specific* towards a particular future-oriented goal. Any quest for an "objective" (commonsense) knowledge base in NLP should be considered as elusive as the fantasy of the complete elimination of noise or bias in natural language communication. In our argument: these are not glitches but *features* of the system.

Other proposed solutions to the programmability of interpretation in WSs exist, most of which involve adding the "correct" categorical associations to concepts. In "Interpreting Winograd Schemas Via the SP Theory of Intelligence and Its Realisation in the SP Computer Model" (2018) Wolff suggests labelling terms with associations such as "peace-loving" to "city councilmen", in order to supplement the language system with the "commonsense" knowledge necessary to arrive at unambiguous interpretations. This proposal is a very clear example of the much criticised unattentiveness to many forms of *biases* (Bender et al., 2021), as it contrabands a specific sociopolitical opinion posing as a neutral position. Bias, while ultimately unavoidable, is usually discernible in examples in which attributes such as "neutrality" or "common" sense are employed in favour of a certain normalcy. Another such effort is ConceptNet, a knowledge/semantic graph that connects "words and phrases with labelled, weighted edges (assertions)" (Speer et al., 2017). ConceptNet was originally part of MIT's *Open Mind Common Sense* project, initiated in 1999 towards the construction of a commonsense knowledge base. Yet another example of an attempt to resolve the WSC by means of semantic graphs is presented in a 2019 paper by Arpit Sharma, which observes the following schema as unambiguous to human readers:

"The man couldn't lift his son because he was so heavy."
or:
"The man couldn't lift his son because he was so weak."

Again, without context, we cannot truly say that we are unambiguously able to distinguish between possible meanings. The context could be that the man himself was too heavy and could thus not lift his son, or the reverse (this is the implied "commonsense" meaning attributed to the sentence). Similarly, in the second sentence, albeit requiring some far-fetching, it is not far from possible to say that the man could not lift his fragile son, because his son was, indeed, too weak. What we would like to draw attention to by presenting these instances is that the semantic noise inherent in the inferences that human agents make about other agents can perhaps be said to be more interesting in the speculative quest for a *common* concept of "reason", than the desire to automate "objective", functional commonsense reasoning. Current NLP proposals for the automation of the latter sense all imply a limited variant of these inferences (e.g. demonstrators are inherently violent), without even addressing them *as* inferences.[6] In a similar vein to the "Stochastic parrots" argument (Bender et al., 2021), where the authors observe that linguistic coherence is not all that matters in NLP results, Elazar et al. (2021) suggest that the majority of disambiguation attempts aimed at discerning the possibility of *the learning of common sense* should in fact disentangle the concept of "actual" commonsense reasoning from the *learned* common sense presented in supposed WS resolutions. The common sense purportedly presented is not only a probabilistic assessment based on a limited corpus, but it is especially concerning if phrased as "commonsense reasoning" when—as in the case of GTP-3, for example—the training data includes WS challenge materials: meaning that coherence is to be expected, if the corpus already contains the variety of common sense its makers expect to find.

Below we can observe a small, non-exhaustive sample of WSs that are not as straightforwardly unambiguous as they are proposed in accounts of the WSC. All these schemas are taken from Ernest Davis' collection (2011, last update: 5/4/2018). Except for the last one, the schemas where explicit mention of the resolution difficulty is already addressed by the author in the collection itself, have been excluded.

> "Jim [yelled at/comforted] Kevin because he was so upset. Who was upset?"

In this case, one could make the claim that the gamut and complexity of human emotions is considered under a rather reductive light if "yelling" is only to be expected when someone is upset. In our alternative reading, Kevin could be the one who is upset, which might inspire Jim to yell. Additionally, Jim could also resort to comforting Kevin because there was nothing else he could do due to his being upset in light of a contextually inaccessible state of affairs.

> "I was trying to balance the bottle upside down on the table, but I couldn't do it because it was so [top-heavy/uneven]. What was [top-heavy/uneven]?"

The alternative interpretative situation would necessitate that the speaker is formulating this somewhat clumsily, but, by all means: the bottle could, in this case, be uneven, and the table top-heavy. This interpretation would result in an awkward but not impossible to imagine physical situation, where, whoever is doing the balancing, is perhaps trying to shuffle the top-heavy table in order to balance the uneven bottle.

> "Susan knows all about Ann's personal problems because she is [nosy/indiscreet]. Who is [nosy/indiscreet]?"

Let us consider the following scenario: Susan could be indiscreet, as someone who imprudently *interferes* in other people's lives. The meaning of indiscreet is approximately that

of a quality possessed by someone lacking in good judgement and/or manners, not simply as someone who *relays* information carelessly. At the same time, Ann could be nosy, as in: someone with an intrusive, meddlesome personality, who incautiously discloses her problems.

> "Fred covered his eyes with his hands, because the wind was blowing sand around. He [opened/lowered] them when the wind stopped. What did Fred [open/lower]?"

This particular WS reveals the oft-mentioned poetic effect which LMs can apparently reproduce but often fail to interpret: Fred could be lowering his gaze and/or opening his hands, there is simply no way to know.

> "The user changed his password from "GrWQWu8JyC" to "willow-towered Canopy Huntertropic wrestles" as it was easy to [remember/forget]. What was easy to [remember/forget]?"

This example is particularly interesting in terms of commonsense reasoning, as memory and information-retrieval means something rather different for a computing system than it does for a human agent. If the capacities of human and artificial systems are to somehow correlate, this schema becomes especially contrived if it is supposed to be comparably interpretable by a shared commonsense logic based on a knowledge framework which has a representation of the concept of *memory*.

> "The police arrested all of the gang members. They were trying to [run/stop] the drug trade in the neighborhood. Who was trying to [run/stop] the drug trade? Answers: The gang/the police."

Davis adds the comment: "Hopefully the reader is not too cynical" (ibid.). Thinking back of our first example about city councilmen and demonstrators, and considering the (socio-political) problems enabled by a lack of theoretical, linguistic criticality, we would suggest that hopefully the reader *is* rather cynical. As mentioned earlier, these examples are non-exhaustive, and if employing a variety of (sometimes far-fetched but still relatively fair) interpretative strategies, almost *all* WSs can be read differently than through the lens of what we have observed as *commonsense bias*. This does not mean "meaning" is completely up for grabs, on the contrary: the fact that it is multiply interpretable makes it negotiable, the dialogue and reasoning that ensues from it is precisely what sharpens discourse and elaborates conceptual prowess. On the other hand, as we observe in our examples, one particular type of meaning is grabbed up and instrumentalised in sometimes innocuous and sometimes dangerously biased directions if not given proper theoretical analysis. As we will see in the next section, it can be argued that the conceptual analysis that results from the semantic noise of many terms is precisely that which drives dialogical, investigative reasoning about them. This is argued for the exemplar case of the concept of "noise" by philosopher Cécile Malaspina in *An Epistemology of Noise* (2018),[7] a work to which the present article owes much of its inspiration. In a completely different context the same argument (i.e. indefiniteness as virtue rather than obstacle) is championed by physicist and philosopher Erik Curiel, in the case of variations in the notion of a "black hole" (2019). In many other cases, as we have seen, the drive to reduce semantic uncertainty installs objectivist projects which haphazardly reduce the complexity and thus possible functionality of the objects at hand.

Shannon and Weaver famously presented the idea that an increase in uncertainty can represent an increase in the "degrees of freedom" of a message, in terms of its information (1964, p. 16, p. 27). That is: the less that is known with exactitude, the more that is possible. While "meaning" is out of the question in this approach to information, we can certainly understand the case of semantic noise as meaning-generative in many ways, as it produces the linguistic conditions for possible social experience: when the word "demonstrator" is employed, it can be "noiselessly" taken for granted (as a naively sketched concept of a violent agitator, an interpretation assumed to be "commonsensically" shared) or it can be questioned, reconsidered, or even misunderstood: thus granting new perspectives on an unavoidably incomplete concept. We will now move on to the speculative, theoretical argumentation of why the notion of semantic noise can help us understand these NLP considerations better.

Semantic noise

A commonly presented yet hardly *conceptually* addressed concern in NLP is the one of *semantic noise* (Luo, 2022). It problematises the measurable specificity inherent in Shannon and Weaver's non-semantic, mathematical formulation of noise, as it implies the variable interpretations of semantic information, and their capacity to affect conduct (Shannon and Weaver, 1964 (1949), p. 5). Unsurprisingly, the term observes multiple definitions in contemporary NLP discussions, but was originally defined by Warren Weaver in TMTC as: "the perturbations or distortions of meaning which are not intended by the source but which inescapably affect the destination" (ibid., p. 26). Linguistic issues pertaining to ambiguity present natural language communications with said distortions, but this situation can hardly be said to be statistically resolvable. As we saw with the WSC, a common and long-standing challenge in NLP is the interpretation of ambiguous anaphora (i.e. a specific case dealing with semantic noise). The present paper asks to what extent seeking to resolve the "problem" of semantic noise is not in itself actually problematic, and revealing of an objectivist agenda. As we will observe, the interpretation of semantic noise as a problem to be overcome is pervasive in the technical literature, and owes much of its contemporary influence to developments in information theory (IT) and statistics. The promise that IT set the ground "for a real theory of meaning" (Weaver, 1964, p. 27) is presented in Shannon and Weaver's seminal account of information and its communication, but never delivered, semantics only appears in the form of questions or open-ended proposals. This article takes semantic noise to be a contingent contextualising element present wherever communicative certainty is sought after. Most importantly for our case study of WSs: this element is often unwillingly ignored as a result of what we have referred to as *commonsense bias*. It is crucial to consider that this meaning-avoidance or agnosticism in early IT is a root cause of the semantics problems that NLP, among other fields, is currently facing. These problems have often been framed as pertaining to bias, for example, by prominent AI scholars such as Gebru et al. (2021), and their criticisms have mostly emerged as calls to rethink large scale systems with claims to universality and neutral applicability. In our analysis of WSs, whenever they are claimed to be unambiguous to human readers, a variety of problems concerning semantic noise emerge: issues regarding spatiotemporal reasoning all the way up to political sentiments. If we observe semantic noise as an unavoidable condition *of* and *for* discourse, rather than seeing it as an obstacle, we can acknowledge that the possible disagreement between interpretative agents is what renders concepts interesting to scientific and speculative inquiry (Curiel, 2019; Malaspina, 2018). As we have argued, the WSC of pronoun disambiguation is conceptually and politically problematic because it assumes an ideal interpreter.

Weaver distinguishes different levels of noise-related problems in TMTC: at the technical and mathematically describable level of

interference, at the level of semantics and at the level of its ensuing behaviour. The two levels involving semantics and behaviour are intimately related, as we observed with our analyses of WSs: behaviour will vary tremendously depending on the semantic context of an agent. Imagine a situation in which the phrase "The city councilmen refused the demonstrators a permit because they advocated violence" is used as a prompt for writing a movie script with a text-generator. Different movies will ensue depending on the LM's interpretation of said sentence. Imagine a less innocuous situation in which legal documents including similar sentences are analysed by an LM, in the making of a case. The possible problems resulting from this can be, in some instances, quite fatal. Semantic noise and material/behavioural noise co-determine each other, they can hardly be understood as two—or three, including statistically measurable noise—hierarchically stacked levels. As Weaver points out: "Here again a general theory at all levels [the mathematical, the semantic, and the behavioural ones] will surely have to take into account not only the capacity of the channel but also (even the words are right!) the capacity of the audience." (Shannon and Weaver, 1964 (1949), p. 27). This begs the question of the capacity of the *sender*, as pointed out by Spärck Jones earlier: the fact that "the words are right" does not imply the original message is "correct," given the fact that a message assumes an interpreting receiver, whose capacity to discern the semantic noise-to-signal ratio is just as relevant as the capacity of the message sender to *accurately* formulate what they want to say. Of course, again, our analysis is limited to pronoun disambiguation in Winograd schemas, and there would be a lot more to say for other situations.

The following quote in TMTC is worth analysing in full length, for the sake of historical context:

> "An engineering communication theory is just like a very proper and discreet girl accepting your telegram. She pays no attention to the meaning, whether it be sad, or joyous, or embarrassing. But she must be prepared to deal with all that come to her desk." (Shannon and Weaver, 1964 (1949), p. 27).

Again, to remark on the notion of meaning: a "very proper and discreet girl" will surely apply some degree of semantic interpretation if the telegram needs to be summarised, for example. The fact that most sentiment analysis tasks in NLP are aimed at things like translation and summarisation should make contemporary researchers particularly wary of this fact. The sidestepping of semantic noise, by the hand of IT and through major conceptual advances such as distributional semantics (Harris, 1970) or more recently *word2vec* (Mikolov et al., 2013), has resulted both in stochastic parrots, but also in the persistent yet vague idea of "meaning" as something objectively, unambiguously present somewhere. In order to model words "conveniently" one might opt for self-updating distributions of probabilities, but the semantic noise that most complex terms exhibit in dialogue remains an essential aspect of how they function, not an impediment to their "ultimate" meaning. In a 2020 paper by Xie et al., we are presented with the proposal of "objective" meanings quite clearly, as the authors aim "to develop intelligent communication systems by considering the semantic meaning *behind* digital bits to enhance the *accuracy and efficiency* of communications" (our emphasis). Contrary to this conception, we argue that—not always—but often enough it is the lack of accuracy in meaning (exemplified by how concepts, such as *noise*, are notoriously negotiated), which renders complex terms conceptually efficient, and our recommendation is that this phenomenon should play a more prominent role in discussions where the purportedly "unique" capacities of humans are compared to those of artificial systems.

In their 2022 paper "Semantic Communications: Overview, Open Issues, and Future Research Directions", Luo et al. recognise that not only is there insufficient theoretical research in the realm of communication systems and semantics, but also that the only way forward is to consider the implementation of technology and the theoretical analysis of semantics jointly (p. 216). However, at the same time, the authors also recommend that: "in order to interpret the meanings successfully at a semantic destination, we need to overcome not only physical channel noise, but also semantic noise in a semantic communication system" (p. 212). We may ask, however, how are we to "overcome" semantic noise, when it is such a fundamental phenomenon in the functions of language? Again: proposing its elimination is a claim to an unavoidably ideological objectivity. It is a problem which also impedes the unlocking of latent conceptual potentials in NLP technologies: if semantic noise represents an obstacle standing between the language model and its supposed correspondence to the natural language world, the solution should not be geared towards the reduction of noise but towards careful attention to the points where it generates relevant frictions. Conceptually conversely to this, most pursuits of language automation so far have often been considered as challenges pertaining to scope and data, rather than problems of conceptual underpinnings. This can be observed in the so-called "brute force" sizing up of language corpora in NLP: the data grow but the conceptual model stays the same.

Additionally, as we've just seen and will continue to argue: these assumptions about the possibility of semantic objectivity also present a specific ideological backdrop with regard to what counts and what does not count as *reasoning*. Kocijan et al. appeal to the psychology of Daniel Kahneman when they propose that: "in a sentence from a well-designed schema, human readers carry out the inference automatically [:] "System 1" (Kahneman, 2011)," and even though they mention that "this inference seems to require commonsense reasoning of some depth and complexity" (Kocijan et al., 2022, p. 4), they fail to address that however "automatic" it may seem, there's a high degree of complexity and depth in the *variability* of *possible* inferences. The reasoning problem remains: natural language is upheld as the reliable "outside" NLP aspires to, and human readers as the "bounded yet rational" (i.e. semantically noisy) users of language. However, neither one of these two assumptions can be said to rest on much else than unfounded sentiments about the capacities of human beings. Noise certainly *can* be characterised as an impediment to message relaying (Shannon and Weaver, 1964 (1949)), and semantic noise *can* present a further impediment, but if and only if we conceive of communication as requiring that utterances represent unambiguous, stable positions. The issue underlined here is not that the characterisation of noise as "unwanted impediment" is mistaken, rather that the promotion of an image of language[8] where most communication takes place unhindered by noise, and is otherwise semantically stable and generally coherent, is promoting an unquestioned appeal to a supposed normalcy of language. Semantic coherence in natural language is fundamentally noisy because it is dialogical: it functions *between* agents and is not *of* any one of them. The (social, political, etc.) relevance installed by any particular pattern (e.g.: "all demonstrators are violent"), is and should always be open to semantic change, a process which involves a great deal of semantic noise.

A recent influential attempt contemplating this supposed problem outside of NLP is Kahneman et al.'s approach to a "smooth" ethics unimpeded by the perils of noise. Their position is that: "Wherever you look at human judgements, you are likely to find noise. To improve the quality of our judgements, we need to overcome noise as well as bias" (Kahneman, 2021). This conceptualisation is subject to the same problems we've already

seen, and represents not only what we could call a "naive" view of (semantic) noise, but also an exemplar case of commonsense bias. In Kahneman's account of noise, crudely put: if we are able to somehow remove it, communication should improve because "Noise is the unwanted variability of judgements, and there is too much of it" (ibid.). Surely if our examples are credit scores and algorithmic bias, we can become inspired to remove certain frictions from our decision-making. But if speaking of human relations at large, which the authors do, what is hereby missed is that the very conditions of being a) situated (i.e. biased), and b) relatively uncertain about what is relevant (i.e. affected by semantic noise) are unavoidable parameters which often *positively constrain* dialogical agents, as they are forced to make inferences *and* communicate about them. Motivated by similarly "objective" ambitions: the notion of ambiguity continues to be simplified as a solvable challenge to computer-mediated communication, a challenge which human agents are supposedly capable of resolving almost instantaneously (Brown et al., 2020; Kocijan et al., 2022; Levesque, 2011, Levesque et al., 2012; Luo, 2022; Mitchell et al., 2018; Morgenstern et al., 2016; Speer et al., 2017; Wolff, 2018; Xie, 2020). As we observed, semantic noise is widely presented as a challenge, but the consequences of this interpretation are rarely analysed. Before moving on to the conclusion, in the section that follows we will present a few perspectives on why this may be so.

Lack of theoretical reflection in language modelling. In an article titled "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences," Miller et al. survey the influence of philosophical, social and behavioural theoretical considerations in the realm of explainable AI—a field closely related to NLP—and the authors find little to no influence of the former on the latter (2017). They consider this unsustainable, given the fact that what software engineers end up designing are tools that—as has been often pointed out in fields such as (new) media or science and technology studies—work well in restricted domains, closed systems, narrow applications, etc. but are hardly applicable as *one-size-fits-all* solutions. This is one of the main points that Bender et al. also take up in their 2021 paper. As we have argued: this is not necessarily an avoidable problem, no model can capture things in their "totality" (as Weaver would have it), and even if this was possible such a model might not be very useful:[9] a model is often useful *because* it is a simplification, a representation: an abstraction. Modelling, generalising, is a fundamental strategy dialogical agents employ, as Bender et al. suggest: "human communication relies on the interpretation of implicit meaning conveyed between individuals [...] even when we don't know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words." (2021, p. 616). The real problem is idealising abstract aspects which lead to an over-reliance on the model's capacity to accommodate complex realities. As Malaspina phrases it: "nowhere in the empirical world is a closed system realised in absolute terms" (Malaspina, 2018, p. 48), we are always bound to some degree of excess. However, what is irrelevant to the model is certainly relevant to that which becomes modelled: restricting the semantic space where natural language users make choices is restricting thought, politics, material praxis, and beyond. Arguably, again, natural language users already exist within a highly restrictive sociocultural network: that of natural language itself. However, it is natural language's fundamental capacity for semantic change that, strangely, fails to become the focus point of systems attempting to somehow capture the functionality of its processes. In much of linguistics, semantics and philosophy we find analysis of this issue to be a common target, but, as Juan Luis Gastaldi notes: "epistemological and philosophical reflections are scarce, at best, in the literature of [NLP]" (Gastaldi, 2021).

If we start from the perspective that we need a solution to the "problem" of semantic noise (e.g. Brown et al., 2020; Xie, 2020; Kahneman et al., 2021; Kocijan et al., 2022; Luo, 2022), we are bound to remain stuck in an uncritical loop where language is understood as semantically equal for all. However, if we start from semantic noise as a *condition*, then language is rendered as a complex multi-agent landscape under permanent change. It can hardly be argued that the reality is the former and not the latter, why does this not prefigure the discussions on WS disambiguation? Elucidating a possible answer to this with the example of gender in language models can be helpful. Gender bias does not simply reside in language corpora, and the solution to this bias is not the diversification of the training data but a discussion of their curation, to begin with, which is an unavoidably semantically noisy affair. While this is brought up in the recent paper by Bender et al. (2021), the focus should not only be on the diversification and curation of language corpora, but in critical, careful attention to the sociopolitical backdrop that leads to this situation being an impasse in the first place: gender representation outside the realms of NLP. The construction of a model cannot begin until its elements have been thoroughly examined, otherwise we can hardly be tempted to call this "science". As Bender et al. conclude in their article, the labour of synthesising human behaviour requires that "downstream effects [are] understood [...] in order to block foreseeable harm to society and different social groups." They insist that research in this area is lacking, and recommend that what is much needed is "scholarship on the benefits, harms, and risks of mimicking humans." (2021, p. 619). This would require that issues pertaining to ambiguity or semantic change in NLP, such as pronoun disambiguation, become approached from said perspective, and not from the perspective which sees semantic noise as an obstacle to be overcome, as this will significantly entrench possibilities in the generative space that is the target of their model: natural language.

Another reason for the lack of theoretical attention is the market-driven solutionism of research areas such as NLP, as theoretical contemplation requires additional development time. Besides the obvious damaging influences of capital-driven enterprises on scientific research, a further problem can be observed in the evaluation of WSs with human subjects who become contracted for setting baselines for tasks such as the WSC: a disproportionately large amount of subjects are recruited on Amazon Mechanical Turk or similar platforms (Kocijan et al., 2022, p. 22). This presents the statistical interpretation of human disambiguation capacities with a limited cultural perspective both in terms of the logic of a specific background, but particularly in terms of it being based on the recruitment of people already biased towards rapid, unambiguous task-resolution, given they are naturally constrained by time, conditions, etc. in the context of performing tasks for platforms such as Amazon. This problem is comparable to the one often presented in the social sciences as the *WEIRD* sampling bias (where results reflect the Western, educated, industrialised, rich and democratic members of the sample analysed).

The importance of the effects of current developments in NLP is not to be underestimated. Given its wide range of applications, it is not an exaggeration to observe that we are witnessing the future (of language) to come. Natural language prompts will come to determine the creation of anything text-based (Floridi and Chiriatti, 2020), from movie scripts all the way to legal frameworks and scripts for virtual realities, a development which

ushers in a variety of ethical problems (of representation) in the process (Bansal et al., 2022). Cautionary observations about the potential harms resulting from the uncritical production of language models range from unintended "behavioural contamination" problems (i.e. as human agents interact with LMs, which on a surface level appear to communicate intentionally, the more humans learn to communicate in a way which works *for the system*, and vice versa: as we saw with the example of WSs being part of training text corpora) to the homogenisation of diversity in dialogical exchanges (i.e. the lack of minority representation, the loss of many dialects, etc.), or worse: to the instalment of novel (legal, infrastructural, etc.) realities which do not admit contestation. The main point we would like to stress is that current NLP interpretations of the function(s) of natural language cloud the *generative* and *social* aspects of human communication. However, the arguments presented here should be read neither as a plea for (1) the abandonment of current NLP enterprises, nor (2) the idolisation of human-language as superior or irreproducible. We simply present these arguments to underline the problems ensuing from a reductive image of language in the trajectory of NLP thus far, and to propose the possibility of conceptually improving it by paying closer attention to these issues.

This is what the argument from semantic noise entails: if primacy is given to stability over variability, i.e. if NLP wishes to resolve the "problem" of semantic ambiguity, then it is bound to an eternal game of linguistic catch-up with the dynamic linguistic landscape outside it, and its essential function will never exceed that of a dictionary, however responsive, dynamic and structurally complex. This is not to make a case for human language users as essentially different than artificial language users, quite the contrary: human language users should be considered just as limited and bound to error as the artificial systems they propose. What *is* different, however, is that semantic noise is not avoided by natural language speakers, it is often sought after and created. As mentioned earlier, in the context of the definition of a black hole, for example, philosopher and physicist Erik Curiel argues that it is an investigative virtue and not a problem for astronomy, physics and philosophy to have to admit variable definitions of the concept of a black hole, for the sake of furthering research (Curiel, 2019). Semantic noise is both structurally constraining and functionally exploitable: in the outlining of conceptual borders, as already mentioned, or for example in exploring emotional compatibility through the use of humour: in jokes, double entendres, puns, etc. But *also* in advancing a particular form of (sociopolitical) life by revealing ideological biases through collaborative discussion, as in the case of our leading example, where politically debatable terms such as "demonstrators" and "city councilmen" play a central role. As we have argued, attention to the semantic noise ensuing from Winograd schemas can reveal the processual, dialogical character of language functions.

If the ultimate goal of synthesising natural language understanding is for it to possess the "fluidity and generality" of a human interpreter (Brown et al., 2020) then NLP needs to engage with semantic noise at a serious conceptual level, rather than dismissing its function and potential as incidental. In relation to this, we follow Gastaldi in proposing a shift in attention in NLP research, suggesting a move towards the exploration of new ways in which language-modelling can reveal something about the nature of dialogical activities in general:

> "[I]f we want to disclose the image of language animating the entire series of those [NLP] models, we need to consider their success as something more than a purely technical feat with respect to specific aspects of language, and redirect that question to the *nature of language itself*. In other terms, to the question "why can computers understand natural language?" we should direct our attention to natural language rather than to computers, and ask: *what must natural language be for the specific procedures of MMs and word embedding models to succeed in revealing some of its most essential aspects*?" (2021, emphasis in original).

While we agree that LMs reveal a great deal about the functions of natural language, our analysis of the WSC in NLP still begs some additional questions, the ones guiding our investigation: What are the consequences of considering human subjects as capable of resolving a task they are not actually fully capable of resolving? What linguistic functionality is lost when semantic noise is excluded from conceptual considerations in NLP? And can we arrive at a different interpretation of the semantic capacity of noisy linguistic phenomena if we recognise them as generative rather than unsolvably problematic? If we recognise them as generative (i.e. requiring cognitive effort to make an assertion *beyond* what they might *seem* to propose "commonsensically"), this acknowledgement can certainly shed light on the social production of knowledge. Among other things, it forces us to observe the—necessary but conceptually insufficient—critique of biases under a different light: biases are not problematic "glitches" to be avoided, incidental noise nuisances to be removed, but are actually fundamental to perception and (collective) meaning-making. It is the very negotiation (for lack of a better word) of these biases that drives natural language communication. This interpretation is opposed to the realist or "commonsense" interpretation of language as simply an ever-vaster repository of ever-more-accurate concepts, a view which, unfortunately, seems to be the dominating perspective in NLP today.

This is not to take away from the fact that research such as OpenAI's GPT is able to actually bring these questions to light, which is the reason for the generation of the current article. However, it seems like a missed opportunity to only reflect on the possible misuses of language-generation tools (such as plagiarism, redundancy, creativity-imitation, e.g. Floridi and Chiriatti, 2020, p. 681) when the fact is that not only are these problems already pervasively at play in the case of human beings, but also that aspiring to imitate "human-level" generativity, while at the same time acknowledging the bias inherent in it *and* pretending that it is possible to mitigate it, is a fundamental contradiction. Ignoring this not only impedes the possibility of said generativity, but also promotes a highly reductive image of language and all it can afford.

Conclusion

In this article, we have focused on a novel analysis of the WSC, and on the concept of semantic noise as presented in IT—as something often deemed as merely incidental—and proposed it as a contingent aspect present in all linguistic communication. We have made a case for promoting the acknowledgement of the dialogical engagement with this unstable conceptual variation of terms, which we repeatedly presented against the objectivist desire to procure stable definitions and disambiguated meanings in NLP research. Various cases of the object under analysis, WSs: anaphoric cases considered as challenges in NLP due to their interpretative ambiguity (e.g.: "The nurse helped the patient even though she was upset".), were reconsidered under the conceptual introduction of semantic noise. The standard assumption in NLP being that, in lacking the "commonsense" capacities that a human being possesses, a language model is not able to determine the "meaning" of many of these frequently occurring syntactic formulations (even in cases with enough contextualising evidence).

Ignoring how human agents can be said to struggle with the exact same disambiguation issues presents many problems. We have provided arguments for why ignoring these problems is a naive interpretation of (semantic) noise, one which proposes a specific "normalcy" of language as well as presents a specific ideological backdrop with regard to what (commonsense) reasoning is and how it functions through natural language users. The problem with the representation of knowledge in the WSC examples we have seen is the prominent reliance on the idea of meaning as *fixed*. Self-updating approaches (such as *NELL*) conceptually tackle this issue, but still fall short because of the presupposition of an objective semantics which the system could in principle learn from. The pervasiveness of semantic noise in natural language as a dialogical process seems to make its own case for the fact that it's relevant, as we saw in the examples presented, and in the proposals made by Malaspina (2018) and Curiel (2019). The argument we have been after, however, is in no way a glorification of semantic noise, but simply its acknowledgement.

Following the precedent set by a project such as *Datasheets for Datasets* (Gebru et al., 2021), where the authors suggest that AI training datasets can become supplemented with accompanying sheets which document their "motivation, composition, collection process, recommended uses, and so on" (ibid.), this paper suggests that NLP solutions (text or speech systems) could at the very least provide (extensive) disclaiming accounts, at the user interface level, of how they fail to capture certain fundamental aspects of natural language. While we already observe minute but important instances of this in initiatives such as ChatGPT (December 15 2022 version), an OpenAI project designed to engage in coherent human-like conversation,[10] it seems a serious cautionary effort is missing in the public presentation of these tools: most users engaging with these applications are not aware of their problematic limitations.[11] Besides a general lack of theoretical reflection, the market-driven solutionism that promotes their fast adoption is also what currently impedes NLP consideration of said matters and dangerously misleads the general public with regard to the "optimality" of LM performance (Bender et al., 2021; Floridi and Chiriatti, 2020). The material analysed, an exposition of current problems in the interpretation of WSs, elucidates how the "noisy" generative processes proposed are often overshadowed by an ideological technical impetus which simplifies language as an indexical, straightforwardly propositional "mirror of reality". Echoing Bender et al.: "we call on the [NLP] field to recognise that applications that aim to believably mimic humans bring risk of extreme harms" (Bender et al., 2021, p. 619). These harms may be as banal as the inability to determine "who was upset" in a specific utterance, to the needless perpetuation of rampant discrimination across a wide range of increasingly pervasive NLP applications, to the invasive technical implementation of an image of language which enforces its non-discursive and rather programmatic aspects, by dangerously idealising "commonsense" reasoning as it occurs in dialogical exchanges.

Data availability

The data analysed during the current study were derived from the following public domain resources: Davis, Ernest, CS NYU WS collection, (2011), available at https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection, accessed: August 12 (2022).

Received: 20 June 2022; Accepted: 27 March 2023;
Published online: 11 April 2023

Notes

1 This has been a question regarding reference for modern logic and philosophy since Frege (See e.g.: Davidson, 1967).
2 Generative Pre-trained Transformer, version 3.
3 In the original context Winograd does remark, however, that: "Of course a semantic theory does not include a theory of political power groups, but it must explain the ways in which this kind of knowledge can interact with linguistic knowledge in interpreting a sentence." (ibid.). Even though he still promotes the disambiguating capacities of human agents with a greater degree of charity than the one we will expose, we find his reflection more charitable towards linguistic complexity than the general NLP proposals that followed from his observations, as we will see.
4 And, if only for the sake of insistence: even this schema can be considered ambiguous, e.g. in the context of a fictional story exploring mathematical and logical paradoxes, such as Lewis Carroll's much beloved *Alice in Wonderland* (1865). This is certainly not a needless mention, Kocijan et al. mention a wide variety of examples from fictional literature in their article (2022, pp. 14–15).
5 The *Never-Ending Language Learning* system is a semantic ML system developed by Carnegie Mellon University, with help from DARPA, Google, NSF, and CNPq and Yahoo! which began in 2010. At least up until 2018 it ran continuously for eight years Mitchell et al. (2018).
6 This lack of inferential reflexivity can be found in the responses presented by ChatGPT (December 15 2022 version). When asked about its interpretation of the Winograd Schema about city councilmen and demonstrators it incessantly returns variations of sentences such as: "While it is true that the meaning of a sentence can sometimes be influenced by broader contextual factors, in this case the meaning of the sentence is clear and specific based on the language of the sentence alone. The sentence states that the city councilmen refused to give the demonstrators a permit because they advocated violence, which directly implies that it was the demonstrators who were advocating violence, not the city councilmen. This interpretation is consistent with the language of the sentence and does not require any additional assumptions or inferences to be made. It is not necessary to consider broader contextual factors in order to understand the meaning of the sentence."
7 Other extensive treatments of the concept include Miguel Prado Casanova's *Noise and Morphogenesis* (2021), or Inigo Wilkins' *Irreversible Noise* (2023).
8 Gastaldi, 2021, echoing Gilles Deleuze's *image of thought*.
9 To briefly mention an influential account of this effect, we could think of the 1:1 scale geographical map presented in Jorge Luis Borges' "On Exactitude in Science", (1960 (1946)), the title of which is perhaps insufficiently translated as *exactitude*, given the Spanish original speaks of *rigor*, denoting a certain harshness in the discipline of modelling.
10 By and large, whenever asked anything speculative, politically complex or personal, ChatGPT responds with: "As a large language model trained by OpenAI, I am not capable of [x, y or z]. I am a machine learning model designed to generate human-like text based on the input I receive. My primary function is to provide information and answers to questions to the best of my ability based on the data I have been trained on". ChatGPT, ca. December 15 2022.
11 Which only exacerbates the problem as it is, among other things, the input these users contribute that continues to grow the datasets that train these tools.

References

Bacchus F, Halpern JY, Levesque HJ (1999) Reasoning about noisy sensors and effectors in the situation calculus. Artif Intell 111(1-2):171–208
Bender EM, Gebru T et al. (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency: 610-623
Bansal H et al. (2022) How well can text-to-image generative models understand ethical natural language interventions? Preprint at arXiv https://arxiv.org/abs/2210.15230
Borges JL (1960 (1946)) Del rigor en la ciencia, *El Hacedor*, Emecé Editores
Brown T et al. (2020) Language models are few-shot learners. Adv Neural Inform Process Syst 33:1877–1901
Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka Jr ER, Mitchell TM (2010a) Toward an architecture for never-ending language learning. AAAI 5:3
Crawford K (2016) Artificial intelligence's white guy problem. The New York Times, 25 Jun 2016
Curiel Erik (2019) The many definitions of a black hole. Nat Astron 3(1):27–34
Davidson D (1967) Truth and meaning. Synthese, Vol. 17, No. 3, Language in Use Including Wittgenstein's Comments on Frazer and a Symposium on Mood and Language-Games: 304-323
Davis E (2022) CS NYU WS collection (2011). https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection. Accessed 12 Aug 2022
Elazar Y et al. (2021) Back to square one: artifact detection, training and commonsense disentanglement in the Winograd Schema. Preprint at arXiv https://arxiv.org/abs/2104.08161
Fagin R, Moses Y, Halpern JY, Vardi MY (1995) Reasoning about knowledge. MIT Press

Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. Minds Machines 30(4):681–694
Gastaldi JL (2021) Why can computers understand natural language? Philos Technol 34(1):149–214
Gebru T et al. (2021) Datasheets for datasets. Commun ACM 64(12):86–92
Harris Z (1970) Distributional structure. Papers in structural and transformational linguistics Springer, Dordrecht: 775–794
Kahneman D, Sibony O, Sunstein CR (2021) Noise: a flaw in human judgment. Harper Collins
Kahneman, D. Thinking, Fast and Slow. Macmillan, 2011
Kocijan V, Davis E, Lukasiewicz T, Marcus G, Morgenstern L (2022) The defeat of the Winograd Schema Challenge. Preprint at arXiv https://arxiv.org/abs/2201.02387
Levesque H (2011) The Winograd Schema Challenge. American Association for Artificial Intelligence. www.aaai.org
Levesque H, Davis E, Morgenstern L (2012) The Winograd Schema Challenge. In: Thirteenth international conference on the principles of knowledge representation and reasoning
Luo X, Chen H-H, Guo Q (2022) Semantic communications: overview, open issues, and future research directions. IEEE Wireless Commun 29(1):210–219
Malaspina C (2018) An epistemology of noise. Bloomsbury Publishing
McCarthy J et al. (1978) On the model theory of knowledge. Department of Computer Science publication, Stanford University
Miller T, Howe P, Sonenberg L (2017) Explainable AI: beware of inmates running the asylum or: how I learnt to stop worrying and love the social and behavioural sciences. Preprint at arXiv https://arxiv.org/abs/1712.00547
Mikolov T et al. (2013) Efficient estimation of word representations in vector space. Preprint at arXiv https://arxiv.org/abs/1301.3781
Mitchell T et al. (2018) Never-ending learning. Commun ACM 61(5):103–115
Morgenstern L, Davis E, Ortiz CL (2016) Planning, executing, and evaluating the Winograd Schema Challenge. AI Magazine 37(1):50–54
Platanios EA, Blum A, Mitchell TM (2014) Estimating accuracy from unlabeled data. In: Proceedings of UAI
Shannon C, Weaver W (1964, (1949)) Mathematical theory of communication. University of Illinois Press, Urbana
Sharma A (2019) Using answer set programming for commonsense reasoning in the Winograd Schema Challenge. Theory Pract Logic Programming 19(5-6): 1021–1037
Spärck JK (2004) Language modelling's generative model: is it rational? In: Technical Report. Computer Laboratory, University of Cambridge
Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: an open multilingual graph of general knowledge. In: Thirty-first AAAI conference
Rahman A, Ng V (2012) Resolving complex cases of definite pronouns: the Winograd schema challenge. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning
Weaver W (1949) Translation. Memorandum, Rockefeller Foundation Archives
Wolff JG (2018) Interpreting Winograd Schemas via the SP theory of intelligence and its realisation in the SP computer model. Preprint at arXiv https://arxiv.org/abs/1810.04554
Winograd T (1972) Understanding natural language. Academic Press, Cambridge
Xie H et al. (2020) Deep learning based semantic communications: an initial investigation. In: GLOBECOM 2020, IEEE global communications conference

Acknowledgements

This article was partly made possible by the *Noise Research Union* (N.R.U.). The author would also like to acknowledge the dialogical, review-process exchanges that led to a legible version of the present article.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S. de Jager.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.