

Generalised Winograd Schema and its Contextuality

Kin Ian Lo Mehrnoosh Sadrzadeh

University College London
London, UK

{kin.lo.20,m.sadrzadeh}@ucl.ac.uk

Shane Mansfield

Quandela
Paris, France

shane.mansfield@quandela.com

Ambiguities in natural language give rise to probability distributions over interpretations. The distributions are often over multiple ambiguous words at a time; a multiplicity which makes them a suitable topic for sheaf-theoretic models of quantum contextuality. Previous research showed that different quantitative measures of contextuality correlate well with Psycholinguistic research on lexical ambiguities. In this work, we focus on coreference ambiguities and investigate the Winograd Schema Challenge (WSC), a test proposed by Levesque in 2011 to evaluate the intelligence of machines. The WSC consists of a collection of multiple-choice questions that require disambiguating pronouns in sentences structured according to the Winograd schema, in a way that makes it difficult for machines to determine the correct referents but remains intuitive for human comprehension. In this study, we propose an approach that analogously models the Winograd schema as an experiment in quantum physics. However, we argue that the original Winograd Schema is inherently too simplistic to facilitate contextuality. We introduce a novel mechanism for generalising the schema, rendering it analogous to a Bell-CHSH measurement scenario. We report an instance of this generalised schema, complemented by the human judgements we gathered via a crowdsourcing platform. The resulting model violates the Bell-CHSH inequality by 0.192, thus exhibiting contextuality in a coreference resolution setting.

1 Introduction

The Winograd Schema Challenge (WSC) originated from the ideas of the American computer scientist Terry Winograd in the 1970s. Winograd was interested in situations where machine understanding could fall behind human understanding. He constructed hypothetical experiments where humans and machines would read a given description, and then answer some questions about it. The descriptions would provide humans with enough context and thus they could answer the questions correctly. However, machine understanding would fall short, as machines did not learn from the context in the same way as humans did. An example description is the sentence “The city councilmen refused the demonstrators a permit because they feared violence.”. The question following it is “Who feared violence?” and the correct answer is “The city councilmen”. If we change the word “feared” to “advocated”, the question will have the opposite answer, namely “the demonstrators”. Winograd’s examples were picked up by the Canadian AI scientist Hector Levesque in 2011. He created a suite of descriptions and questions, proposing them as a test of machine intelligence - an alternative to the Turing Test [26]. Later, the AI company Nuance put forwards a cash prize of USD 25,000 for any AI that could solve the challenge with an accuracy close to humans, 92-96%. No AI system managed to achieve the target, and as a result, the prize was withdrawn in 2018. It was not until the 2020s that large pre-trained language models, employing transformer architectures, eventually reached a performance level comparable to human accuracy [24]. Despite these advancements, the WSC continues to present significant challenges for AI systems lacking extensive data resources and computational power.

In previous work, we showed how natural language notions of context can be modelled by the mathematics of quantum contextuality [36, 37, 34]. In particular, we modelled anaphoric context in [28]. Inspired by the reliance of the WSC on anaphoric context, we decided to explore whether quantum contextuality could potentially provide a solution to the challenge.

Our initial examination found that the WSC in its original form lacked the complexity required to be of interest from a quantum contextuality standpoint. Upon modelling the WSC within the sheaf theoretic framework, it became evident that the scenario was too simplistic to exhibit contextuality, as the models derived from it were deterministic.

This motivated us to extend the schema and allow it to be non-deterministic such that it can, in principle, host contextuality. This was achieved by introducing additional linguistic context, namely, (1) two special words rather than one and (2) two ambiguous pronouns instead of one. Consequently, we obtained more observables and more measurement contexts, leading to a scenario that resembles the Bell-CHSH scenario.

The above outlines the first contribution of this paper. Our second contribution lies in the instantiation of our generalized Winograd Schema and the collection of human judgments via a crowdsourcing platform. This allowed us to calculate the violation of the Bell-CHSH inequality and thereby establish the contextuality of our model, which was constructed based on human judgments. We also modelled the data using the Contextuality-by-Default (CbD) framework of contextuality and calculated a corresponding CbD degree of contextuality. It was found that our probabilistic model exhibited contextuality in both the Bell-CHSH and CbD models.

2 Contextuality

The origins of contextuality research can be traced back to 1935, with the work of Einstein, Podolsky, and Rosen (EPR) [15]. In their work, they posited that the quantum mechanical description of physics was incomplete when two spatially separated parties were permitted to make measurements on an entangled system. A way of formalising such theories is in terms of hidden variables, which, if known, might fully determine the outcome that would result from any given measurement. Bell's theorem [7, 6] in the 1960s showed that no hidden-variable theory exists for quantum mechanics unless the measurement outcomes were allowed to be dependent on which other measurements are performed simultaneously. Around the same time, Kochen and Specker [23] independently demonstrated that there exists a set of measurements in a 3-dimensional Hilbert space such that a non-contextual hidden-variable theory cannot exist, regardless of the state of the system. These two results, collectively known as the Bell-Kochen-Specker theorem, showed that a hidden-variable theory for quantum mechanics must be contextual, providing some clarity to the debate on a more fundamental theory conforming to certain classical intuitions for quantum mechanics. The first attempt at experimentally verifying Bell's inequality was performed by Aspect et al. [5], with the most recent ones closing all known loopholes in the earlier experiments [18, 20, 32]. Thus it has been established that quantum physics is vastly different from classical physics – a description of quantum physics that agrees with our classical intuition must be contextual.

Other than the philosophical implications, contextuality has been shown to possess computational power through non-classical correlations. Anders and Browne first showed that certain measurements on GHZ states can be used to lift a linear classical computation into a universal classical computation [4]; Raussendorf later showed that the probability of success of such computation is bounded by the degree of contextuality [30], as measured by the contextual fraction [2, 1]. Subsequent work by Howard et al. revealed that contextuality is an essential ingredient for *magic state distillation*, a process that yields

specific quantum states known as *magic states* [21]. The current most promising fault-tolerant quantum computing scheme, the surface code [22], only permits fault-tolerant computation with a subset of quantum operations which can be efficiently simulated by classical computers. Via state injection, these magic states can be used with surface code to allow for fully fault-tolerant universal quantum computation. Thus, one might argue that contextuality carries an intrinsic computational power that is absent in non-contextual systems.

A variety of frameworks for modelling contextuality have been developed. These including the sheaf-theoretic framework [2, 3, 1], the Contextuality-by-Default (CbD) framework [11, 14, 12], the graph-theoretic framework [8], a framework based on simplicial sets [29]. Generally speaking, these frameworks enable the formalisation of the notion of measurement through the use of various mathematical structures. Bell's inequalities, or in general inequalities that witness contextuality, can be derived systematically within these frameworks. Although we will mainly use the terminology from the sheaf-theoretic framework to describe our examples, our results are framework-agnostic.

2.1 Sheaf Theoretic Framework

Here, we provide a concise overview of the sheaf-theoretic framework of contextuality proposed by Abramsky and Brandenburger [2]

A measurement scenario is defined as a triplet $\langle X, \mathcal{M}, O \rangle$, where X refers to a collection of observables, O is the possible outcomes, and \mathcal{M} denotes an abstract simplicial complex composed of subsets from X .

Every element in X is an observable of the system under consideration. Upon measurement, each observable yields one of the outcomes contained in O . The characterization of \mathcal{M} as an abstract simplicial complex implies a particular structural feature: if a subset C belongs to \mathcal{M} , then every subset nested within C must also be an element of \mathcal{M} .

A necessity of contextuality is that one cannot measure all the observables in X simultaneously, at least not without altering the state of the system. Thus, every framework for contextuality must provide a description of the compatibility between observables. Within the sheaf-theoretic framework, each simplex in the simplicial complex \mathcal{M} constitutes a subset of observables in X that are measurable simultaneously, i.e. they are mutually compatible. A *measurement context*, or simply *context*, is defined as a maximal simplex in \mathcal{M} , which is not a proper subset of any other simplex in \mathcal{M} .

For instance, the measurement scenario in the Bell-CHSH settings is specified by $X = \{a_1, a_2, b_1, b_2\}$; $\mathcal{M} = \{\{a_1, b_1\}, \{a_1, b_2\}, \{a_2, b_1\}, \{a_2, b_2\}\}$; $O = \{0, 1\}$. The simplicial complex \mathcal{M} can be geometrically realized as the boundary of a square, where each vertex corresponds to an observable and each edge represents a context (see Figure 1(a)). Two parties are involved in this scenario: Alice is allowed to measure either a_1 or a_2 , and Bob is allowed to measure either b_1 or b_2 . The measurements are dichotomic, i.e. the outcomes are either 0 or 1.

Every subset of observables which is a context in \mathcal{M} can be measured jointly. Thus we can define a (local) joint probability distribution over the observables in the context. Such a joint probability distribution can either be estimated by performing the measurements in an experiment, or be calculated according to a theory of the system under consideration. A collection of all such joint probability distributions is called an *empirical model*, or simply *model*, of the system. For instance, using a set of appropriately chosen measurement bases, the Bell state $|\Psi\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$ produces the empirical model depicted in Figure 1(b). This state exhibits the highest violation of the Bell-CHSH inequality among all quantum states.

An empirical model is said to be *signalling* if the marginalised distribution of a set of observables

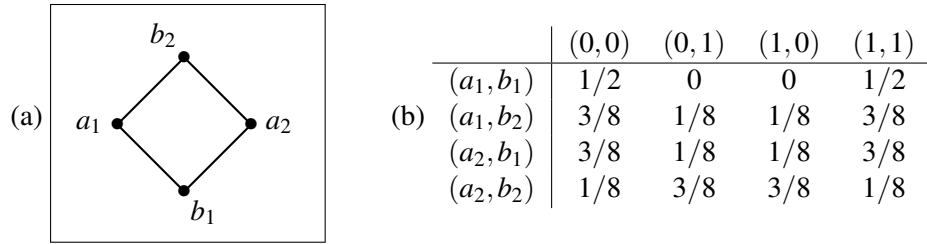


Figure 1: (a) The simplicial complex \mathcal{M} in the Bell-CHSH scenario. Every vertex represents an observable and every edge represents a context. Alice chooses between a_1 and a_2 ; Bob chooses between b_1 and b_2 . The absence of edges between a_1 and a_2 , and between b_1 and b_2 , indicates their incompatibility. (b) An empirical model of the Bell-CHSH scenario. Each row represents a joint probability distribution over the observables in the context. For example, the bottom-right entry $1/8$ is the probability of observing $a_2 = 1$ and $b_2 = 1$ when measuring the observables in the context (a_2, b_2) .

differs from one context to another. In contrast, non-signalling implies that the observed probabilities remain invariant under different contexts, thereby preventing the transmission of information through the choice of context.

A prevalent misconception is a belief that *signalling is contextuality*, often based on the incorrect reasoning that the *probabilities* in a signalling model are generally context-dependent, leading to the conclusion that the model is contextual. However, it is essential to recognize a fundamental distinction between the two concepts: signalling pertains to the observed probabilities, while contextuality relates to the underlying hidden-variable theories of the model.

The qualitative criterion for contextuality of a model in the sheaf-theoretic framework is based on Fine's theorem [17], which states that a model is contextual if and only if there exists a global probability distribution that is compatible with every local probability distribution in the model.

The quantitative degree of contextuality of a model is measured by the *contextual fraction* CF [1]. Given an empirical model e , the contextual fraction $\text{CF}(e)$ is defined as the minimum λ such that e admits a convex decomposition¹:

$$e = (1 - \lambda)e^{NC} + \lambda e^C, \quad (1)$$

where e^{NC} is a non-contextual (and non-signalling) empirical model and e^C is an empirical model that may be contextual.

Suppose a given model e is non-contextual, then λ can be set to zero by choosing $e^{NC} = e$. Otherwise, λ must be taken to be greater than zero to make the decomposition valid. Therefore, for non-signalling models, the sheaf-theoretic criterion of contextuality is

$$\text{CF}(e) > 0. \quad (2)$$

The calculation of CF can be reduced to solving for a linear program, for which numerical solvers are readily available. The CF of a model has a nice interpretation as the maximum amount of *normalised violation* of all possible general Bell's inequalities [1].

In the case of signalling models, the above decomposition cannot hold because e^{NC} and e^C are, by definition, non-signalling. We could consider allowing e^C to be signalling. However, this adjustment would lead to the misleading conclusion that all signalling models are contextual, assuming we maintain our interpretation of CF as a measure of contextuality for these models.

¹Here, we represent the empirical models as empirical tables. Addition and scalar multiplication are then interpreted as standard matrix operations, where the empirical tables are treated as matrices.

2.2 Contextuality by Default

In the setting of Contextuality-by-Default (CbD), there are two important notions: *contents*, denoted by q_i , which are measurements, or more generally, questions about the system; and *contexts*, denoted by c^j , which represent the conditions under which the questions are asked, e.g. their ordering. Every q_i in a c^j gives rise to a random variable R_i^j taking values in $\{\pm 1\}$, and representing possible answers and their probabilities. All random variables in a given context are jointly distributed.

A well-studied class of CbD systems are the cyclic systems [11, 12, 13], where each context has exactly 2 contents and every content is in exactly 2 contexts. The rank of a cyclic system is the number of contents, or equivalently, the number of contexts.

A cyclic system of rank n is contextual if and only if CNT_1 is positive, where CNT_1 is defined as:

$$\text{CNT}_1 := s_{\text{odd}} \left(\left\{ \langle R_{i_j}^j R_{i'_j}^j \rangle \right\}_{j=1, \dots, n} \right) - \Delta - n + 2 > 0 \quad (3)$$

where $i_j \neq i'_j$ for all j and $R_{i_j}^j, R_{i'_j}^j$ are well-defined for all j . Quantities $s_{\text{odd}} : \mathbb{R}^n \rightarrow \mathbb{R}$ and Δ are defined as follows:

$$s_{\text{odd}}(\underline{x}) = \max_{\substack{\sigma \in \{\pm 1\}^k; \\ \text{p}(\underline{\sigma} = -1)}} \underline{\sigma} \cdot \underline{x}; \quad \Delta = \sum_{i=1}^n \left| \langle R_i^{i_i} \rangle - \langle R_i^{i'_i} \rangle \right| \quad (4)$$

where $\text{p}(\underline{\sigma}) = \prod_{i=1}^n \sigma_i$ (p is the parity function of $\underline{\sigma}$). The quantity Δ measures the degree of signalling in the system. Thus, a non-signalling system has $\Delta = 0$.

For a rank 4 cyclic system, i.e. the Bell-CHSH scenario, the above inequality reduces to the maximum violation of the Bell-CHSH inequalities over the choices of the four signs:

$$\text{CNT}_1 = \pm \langle R_0^0 R_1^0 \rangle \pm \langle R_1^1 R_2^1 \rangle \pm \langle R_2^2 R_3^2 \rangle \pm \langle R_3^3 R_0^3 \rangle - 2 \quad (5)$$

where the number of minus signs has to be taken odd. Therefore, the CbD criterion of contextuality coincides with the Bell-CHSH inequalities for the Bell-CHSH scenario.

2.3 Ambiguous words as observables

Ambiguities in natural language have posed a challenge to natural language processing. Lexical ambiguity, where a word has multiple meanings, is one of the most common types of ambiguity in natural language. For instance, the word *produce* has two possible meanings: *to give birth* and *to make something*.

Without any context, it is not possible to determine which of the two meanings is intended. Another type of ambiguity is *coreference ambiguity*, where a word can potentially refer to different entities. For instance, the pronoun *it* can refer to the *dog* or the *cat* in the sentence *The dog chased the cat. It barked..* In this paper, we focus on the latter type of ambiguity.

A method to formalise the notion of contextuality in natural language is by viewing an ambiguous word as an observable, with its interpretations as possible outcomes. For instance, the word *produce* has (at least) two possible interpretations: *to give birth* and *to make something*. Measuring the word *produce* amounts to selecting one of these interpretations by a reader.

We can assign probabilities to these interpretations based on the frequency of the interpretations in an experiment where a group of readers is asked to interpret the word *produce*, or a single reader is asked to assign a probability to each of the interpretations. The first approach is more costly as it requires a

large number of readers to be involved in the experiment. However, the latter approach is better suited to machine learning models since they can be trained to assign probabilities to different interpretations.

This way of treating ambiguous words as observables was first proposed by Wang et al. [36, 37]. The authors considered subject-verb and verb-object phrases where each word carries at least two possible interpretations. Measurement contexts were constructed by selecting different pairs of nouns and verbs, in a way similar to how Alice and Bob select their measurements in the Bell-CHSH scenario. The probabilities in the results were estimated from a group of crowd workers who were asked to assign a score to the different interpretations.

3 Winograd Schema Challenge

Commonsense reasoning, the inherent human capacity to logically comprehend the world around us, has long been a focal point in the field of artificial intelligence, with the aim to cultivate this ability in machines.

The Winograd Schema Challenge (WSC) emerged as a measure of this commonsense reasoning capability. The challenge was inspired by Terry Winograd’s seminal paper [38], wherein he contended that syntax alone falls short in the interpretation of natural language, necessitating commonsense or world knowledge as well. The challenge presents a collection of sentences, each with an ambiguous pronoun whose meaning can be clarified via the context. A machine is deemed to have passed the test if it can disambiguate the pronoun with an accuracy on par with human performance.

The classic example of a Winograd schema, originally constructed by Winograd himself, is the following pair of sentences:

- (1) a. The city councilmen refused the demonstrators a permit because **they** *feared* violence.
- b. The city councilmen refused the demonstrators a permit because **they** *advocated* violence.

Note that the two sentences differ only in the words *feared* and *advocated*. In both sentences, there is an ambiguous pronoun **they** which can either refer to the *city councilmen* or the *demonstrators*. In the first sentence, it can be inferred through commonsense reasoning that the pronoun **they** refers to the *city councilmen*, as it is within our common sense that city councilmen are the ones who tend to prevent violence in demonstrations. In the second sentence, the pronoun *they* refers to the *demonstrators*, as it is within our common sense (stereotype) that demonstrators tend to advocate violence and that doing so would lead to the refusal of a permit for a demonstration.

Another classic example of a Winograd schema is the following pair of sentences:

- (2) The trophy doesn’t fit into the suitcase because it’s too [*small / large*].

Here we adopt a compact notation in which the pair of square brackets encloses the two possible word choices, each leading to a different sentence. This notation will be employed throughout the paper.

In a WSC, the participant is asked to identify the correct interpretation of the ambiguous pronoun. Success in the test is defined by the participant’s accuracy equalling or approximating human performance. The evaluation of responses to a WSC question is straightforward, either the correct referent of the ambiguous pronoun is identified or not.

In contrast, the Turing Test has been criticised for being too difficult to evaluate. Originated as the imitation game by Turing [33], the test involves a human judge interrogating a machine via a textual interface. The conversation between the judge and the machine is unrestricted. If the judge or a panel of judges cannot distinguish the machine from a human based on the conversation, the machine is deemed

to have passed the test. However, this unrestricted nature of the Turing Test opens doors to potential deception. In fact, for a machine to pass the test, it must deceive as machines lack physical bodies. If questioned about its physical attributes, like height or weight, the machine must lie to successfully pose as a human. Due to this advantage of the ease of evaluation over the Turing Test, the WSC was proposed as a replacement for the Turing Test.

Unlike the Turing Test, the WSC is a structured binary-choice test. The major issue with the WSC is that it is over-constrained - it is unexpectedly difficult to construct examples of it, due to the numerous requirements that must be satisfied. A valid Winograd schema must satisfy the following requirements:

1. A Winograd Schema comprises a pair of sentences that differ slightly from each other. The first sentence includes a *special* word which, when replaced by an *alternate* word, yields the second sentence. For instance, in the *trophy-suitcase* example, *small* is the *special* word, and *large* is its *alternate*.
2. The sentences should contain two noun phrases. In the *trophy-suitcase* example, *the trophy* and *the suitcase* serve as the two noun phrases.
3. A pronoun, which agrees with the two noun phrases in number and gender, must be present in the sentences. For example, in the *trophy-suitcase* scenario, the pronoun *it* aligns with both *the trophy* and *the suitcase* regarding number and gender.
4. The pronoun's referent should be easily identifiable from a natural reading of the sentence, and the correct referent should differ between the two sentences.
5. Each sentence in the pair should be fluid and natural to read, to the extent that they could feasibly appear in regular text sources like news articles or Wikipedia pages.

The outlined requirements ensure the preservation of both linguistic structure and the test's integrity:

1. The first requirement ensures grammatical consistency across the pair of sentences.
2. The fourth requirement necessitates a change in the correct referent of the pronoun when the special word is replaced with the alternate. This stipulation indicates that grammatical structure alone does not determine the correct pronoun referent.
3. The fifth requirement safeguards the authenticity of the language used in the test, ensuring it remains aligned with naturally occurring language.

Crafting valid examples of the Winograd schema is a complex task due to the set restrictions and requirements. The challenge of creating such schemas is evidenced by the limited number of examples in the original Winograd Schema Challenge set, which includes only 285 instances².

In 2018, the first system achieved a better-than-chance accuracy of 57.1% [16] on the original 285 examples of the WSC. In 2019, a fine-tuned RoBERTa [27] model achieved a human-like accuracy of 90.1% [31]. The WSC has suffered from the same problem that plagued the Turing Test – there are weaknesses in the test that can be exploited without having to demonstrate the desired human-level intelligence. Simply put, the WSC has been defeated [24].

It is even more so for the WSC precisely because of its ease of evaluation. Proposals to increase the difficulty of the WSC, such as requiring the test-taker to select a correct explanation for their answer from a list of options [39, 19], emerged as potential solutions. However, these suggestions further complicate the already challenging task of question set construction. An alternative could involve requiring free-form explanations from the test-taker, though this would likely introduce additional ambiguity and make the evaluation process more difficult.

²Available at <https://cs.nyu.edu/davise/papers/WinogradSchemas/WS.html>.

4 Generalised Winograd Schema

In this section, we present our approach for the generalisation of the Winograd Schema, enabling the potential observation of contextuality. We will first discuss why the original Winograd Schema is insufficiently complex to exhibit contextuality, and then propose a generalised Winograd Schema that is sophisticated enough to host contextuality.

4.1 Modelling Winograd Schemas as measurement scenarios

To study the contextuality in the Winograd Schema, we model it with a measurement scenario in the sheaf-theoretic framework. This way of treating ambiguity in language is akin to the way ambiguous phrases are treated in [36], where an ambiguous word is considered an observable in a measurement scenario.

However, the same ambiguous word, i.e. the ambiguous pronoun, is shared across the twin pair of sentences in a Winograd Schema. Thus, if we follow the approach of “words as observables” strictly, then we will end up with a trivial measurement scenario, where there is only one observable, i.e. the ambiguous pronoun. Moreover, this naive approach deviates from the spirit of the Winograd Schema, which is to disambiguate a pronoun by considering the linguistic context. Instead, We argue that there should be exactly two contexts in the measurement scenario, one for each sentence in the twin pair. Recall that in the original Winograd Schema, the twin pair of sentences are identical except for the special word and the alternate word. In a rough sense, the special word and the alternate word provide the *linguistical context* for disambiguating the pronoun. This way of defining the measurement contexts provides a concrete link between *context in language* and *contextuality in quantum mechanics*.

Following from the above discussion, we define an observable as a tuple: **(pronoun, special word)** or **(pronoun, alternate word)**, to distinguish between the two pronouns in different linguistic contexts. The possible outcomes of each of the two observables are the candidate referents of the pronoun.

Definition 1 (Winograd Schema scenario) *Given a Winograd Schema with two noun phrases A and B ; an ambiguous pronoun \mathbf{p} which refers to either A or B ; a special word (s) and an alternate word (a), the corresponding measurement scenario is defined by the data:*

- observables $X = \{(\mathbf{p}, s), (\mathbf{p}, a)\}$;
- contexts $\mathcal{M} = \{\{(\mathbf{p}, s)\}, \{(\mathbf{p}, a)\}\}$;
- outcomes $O = \{A, B\}$.

We call such a measurement scenario a Winograd Schema scenario, or a WS scenario in short.

With the *councilmen-demonstrators* example, the measurement scenario would be given by the data:

- observables $X = \{(\mathbf{they}, \textit{feared}), (\mathbf{they}, \textit{advocated})\}$;
- contexts $\mathcal{M} = \{\{(\mathbf{they}, \textit{feared})\}, \{(\mathbf{they}, \textit{advocated})\}\}$;
- outcomes $O = \{\textit{city councilmen}, \textit{demonstrators}\}$.

It becomes apparent that any Winograd Schema scenario is too simplistic to accommodate any contextual model due to the absence of overlapping contexts. One can always construct a compatible global distribution by taking the product of the local distributions.

4.2 Generalising the Winograd Schema scenario

Before proceeding to the generalisation of Winograd Schema, we point out an interpretation of the WS scenario as an analogy to an experiment in quantum physics. Consider an imaginary experimenter, Alice, who decides whether to measure the pronoun with the special word, or with the alternate word. That is, Alice chooses between the two observables: (\mathbf{p}, s) and (\mathbf{p}, a) . This is exactly analogous to Alice choosing between two projection axes in an experiment measuring a spin-1/2 particle.

A natural and obvious way to generalise the WS scenario would be to add one more experimenter, Bob. This results in the Bell-CHSH scenario, which is well-known to be able to host contextual models. That amounts to introducing one more pronoun, one more special word and its alternate word, to the original Winograd Schema. We use the subscript 1 to denote objects relating to the first pronoun and the subscript 2 to denote objects relating to the second pronoun.

Here we give a set of requirements for the generalised Winograd Schema, in the style of the original WSC:

1. A generalised schema consists of four slightly differing sentences. The first sentence contains two special words s_1 and s_2 . Similar to the original Winograd Schema, s_1 can be replaced by an alternate word a_1 and s_2 can be replaced by an a_2 . The possibility of replacing special words with alternate words creates the rest of the four sentences.
2. There are a pair of noun phrases.
3. There are two pronouns in the sentences. The first pronoun refers to one of the noun phrases in the first pair of noun phrases. The second pronoun refers to either one noun phrase in the second pair of noun phrases.
4. All four sentences should be natural to read.

In short, a generalised Winograd Schema is two Winograd Schemas put together in a single discourse.

Definition 2 (Generalised Winograd Schema scenario) *Given a Generalised Winograd Schema with two noun phrases A and B ; two ambiguous pronouns \mathbf{p}_1 and \mathbf{p}_2 can each refer to either A or B ; two special words (s_1) and (s_2) ; two alternate words (a_1) and (a_2) , the corresponding measurement scenario is defined by the data:*

- observables $X = \{(\mathbf{p}_1, s_1), (\mathbf{p}_1, a_1), (\mathbf{p}_2, s_2), (\mathbf{p}_2, a_2)\}$
- contexts $\mathcal{M} = \{(\mathbf{p}_1, s_1), (\mathbf{p}_2, s_2)\}, \{(\mathbf{p}_1, s_1), (\mathbf{p}_2, a_2)\}, \{(\mathbf{p}_1, a_1), (\mathbf{p}_2, s_2)\}, \{(\mathbf{p}_1, a_1), (\mathbf{p}_2, a_2)\}$;
- outcomes $O = \{A, B\}$.

Such a measurement scenario is called a Generalised Winograd Schema scenario, or a generalised WS scenario in short.

The generalised WS scenario is isomorphic, i.e. identical upon relabelling, to the Bell-CHSH scenario shown in Figure 1. It has long been known that the Bell-CHSH scenario can host contextual models [6, 9]. Thus a carefully designed generalised Winograd Schema would be able to demonstrate contextuality.

Here we provide a straightforward example of a generalized Winograd Schema scenario, built upon the original *trophy-suitcase* example:

- (3) The trophy doesn't fit into the suitcase because \mathbf{it}_1 is too [$s_1 = \textit{small} / a_1 = \textit{large}$]. Nonetheless, \mathbf{it}_2 is [$s_1 = \textit{light} / a_2 = \textit{heavy}$].

The corresponding generalised WS scenario is given by:

- observables $X = \{(\mathbf{it}_1, \textit{small}), (\mathbf{it}_1, \textit{large}), (\mathbf{it}_2, \textit{light}), (\mathbf{it}_2, \textit{heavy})\}$
- contexts $\mathcal{M} = \{ \{(\mathbf{it}_1, \textit{small}), (\mathbf{it}_2, \textit{light})\}, \{(\mathbf{it}_1, \textit{small}), (\mathbf{it}_2, \textit{heavy})\}, \{(\mathbf{it}_1, \textit{large}), (\mathbf{it}_2, \textit{light})\}, \{(\mathbf{it}_1, \textit{large}), (\mathbf{it}_2, \textit{heavy})\} \};$
- outcomes $O = \{\textit{trophy}, \textit{suitcase}\}$.

Interestingly, it was in the original set of Winograd Schemas (WSC285) that Davis designed a special example making use of two pronouns:

- (4) Sid explained his theory to Mark but **he** couldn't [*convince / understand*] **him**.

The author deemed this example a ‘‘Winograd schema in the broad sense’’ since using more than one pronoun violates the requirements of the original Winograd Schema. Yet, this example is not a proper generalised Winograd Schema defined in this paper, as it only employs one special word and one alternate word.

Other than the fact that its scenario is too simple, there is another reason why the original Winograd Schema is not contextual: the intended referent of the pronoun should be obvious to a human reader. That means an empirical model constructed with judgement data collected from human subjects on the original Winograd Schema would be deterministic or nearly deterministic. It is known that deterministic systems are not contextual [10]. On the other extreme, a completely random model is trivially non-contextual. Intriguingly, it seems that only a system with a moderate level of intelligence, in between that of humans and that of complete randomness, would have the possibility of being contextual.

There are two directions to where we could take the generalised Winograd Schema: (1) to continue its mission to be a test of intelligence or commonsense reasoning; (2) to become a well-structured linguistic setting under which contextual models could be found.

Recent results from large language models have demonstrated human-like accuracies in solving the Winograd Schema Challenge. The introduction of one more pronoun might increase the difficulty of the challenge, possibly stipulating advancements in the field of natural language processing. However, it is our goal to find bridges between natural language and contextuality. Therefore the second direction will be the focus of this paper.

4.3 An example of the generalised Winograd Schema

As our goal is to uncover contextual models in natural language, we need to gather judgment data from human participants to build empirical models for generalized Winograd Schema instances. Crucially, deterministic systems lack contextuality. Therefore, our generalized Winograd Schema examples should be inherently ambiguous to human readers, unlike the original Winograd Schema where humans can easily resolve the pronoun.

Due to the requirement of having two almost identical pairs of naturally-sounding sentences, it is a difficult task to come up with examples of the original Winograd Schema. The extra requirements we put forward for the generalised Winograd Schema make it even harder to come up with naturally-sounding examples. Here we report an example of the generalised Winograd Schema³:

- (5) A and B belong to the same [*cannibalistic / herbivorous*]₁ species of animal. On a hot afternoon in the south Sahara, **one of them**₁ was very hungry. They noticed each other when they were roaming in the field. After a while, **one of them**₂ is no longer [*hungry / alive*]₂.

³It was pointed out by one of the reviewers that the original version of the example contains several incorrect uses of English. Here we provide the corrected version of the example.

Note that we had to violate the requirement of having a single sentence because it is difficult to come up with a naturally-sounding sentence that contains every ingredient of the generalised Winograd Schema. We also decided to use the referring phrase **one of them** instead of the third-person pronoun **it** to improve the naturalness of the example.

We used the alphabetic symbols A and B as the two noun phrases as we wanted to make the two symmetric. That is, any empirical model of the scenario is invariant to the interchanging of A and B. It turns out that all symmetric models are non-signalling, at least for cyclic scenarios such as that Bell-CHSH scenario. Dealing with symmetric models carries two disadvantages: (1) it is more difficult to assert the contextuality of a signalling model; (2) the sheaf-theoretic criterion of contextuality applies to non-signalling models only. By considering only symmetric models, we thereby avoid the complications of dealing with non-signalling models.

4.4 Human judgements on the example

We collected human judgments on this example on the crowd-sourcing platform Amazon Mechanical Turk in the form of a questionnaire. There were four versions of the questionnaire, each corresponding to one of the four contexts in the generalised WS scenario. The respondents were asked to read the example and answer a question about the correct referents, A or B, of the two referring phrases **one of them₁** and **one of them₂**. A screenshot of the questionnaire is shown in Figure 2.

Instruction: Please read the following short story which contains some ambiguities, then select the interpretations you think are the most appropriate.

Story: A and B belong to the same $\{\text{word1}\}$ species of animals. In a hot afternoon in south Sahara, one of them was very hungry. They notice each other when they were roaming in the field. In a while, one of them is no longer $\{\text{word2}\}$.

Question: The following are 4 different interpretations of the story. Please select the **2** most appropriate interpretations.

- A was the very hungry $\{\text{word1}\}$ animal. A is no longer $\{\text{word2}\}$.
- A was the very hungry $\{\text{word1}\}$ animal. B is no longer $\{\text{word2}\}$.
- B was the very hungry $\{\text{word1}\}$ animal. A is no longer $\{\text{word2}\}$.
- B was the very hungry $\{\text{word1}\}$ animal. B is no longer $\{\text{word2}\}$.

Figure 2: A screenshot of the template of the questionnaire. The placement holders $\{\text{word1}\}$ and $\{\text{word2}\}$ are instantiated with the two special words or the alternate words of the generalised Winograd Schema. In this example, $\{\text{word1}\}$ can be either *cannibalistic* or *herbivorous* and $\{\text{word2}\}$ can be either *hungry* or *alive*. Four versions of the questionnaire were created, each corresponding to one of the four contexts in the generalised WS scenario. *Note that the story contains several incorrect uses of English. Unfortunately, we did not notice these until a reviewer pointed them out, after data collection.*

Since each referring phrase can be interpreted in two ways, there are 4 possible combinations of interpretations, (A, A), (A, B), (B, A), (B, B), of the two referring phrases. The symmetry between A and B in the example ensures that the combinations (A, A) and (B, B) are equally plausible and (A, B) and (B, A) are also equally plausible. Therefore we asked the respondents to pick two out of the four combinations. This design choice also allows the detection of invalid answers, that is, those that do not

(a)	(A, A)	(A, B)	(B, A)	(B, B)	(b)	(A, A)	(A, B)	(B, A)	(B, B)
(<i>canni</i> , <i>hungry</i>)	0.402	0.097	0.097	0.402	...	1/2	0	0	1/2
(<i>canni</i> , <i>alive</i>)	0.044	0.455	0.455	0.044	...	0	1/2	1/2	0
(<i>herbi</i> , <i>hungry</i>)	0.345	0.154	0.154	0.345	...	1/2	0	0	1/2
(<i>herbi</i> , <i>alive</i>)	0.344	0.155	0.155	0.344	...	1/2	0	0	1/2

Table 1: (a) The empirical model constructed with the 410 human judgments collected from Amazon Mechanical Turk. The violation of Bell’s inequality of the model is 0.192 ± 0.176 . For brevity, the special word *cannibalistic* is shortened to *canni* and the alternate word *herbivorous* is shortened to *herbi*. The model generally resembled the PR model shown in Table (b) on the right.

respect the symmetry between A and B.

A total of 410 responses were collected on Amazon Mechanical Turk separately on two dates: 20th Oct 2022 and 23rd Nov 2022. Out of the 410 responses, 110 were to the context (*cannibalistic*, *hungry*) and 100 each were to the rest of the three contexts. Out of all the responses, 348 were valid, i.e. their responses respected the symmetry between A and B. The respondents were each financially rewarded USD 1.00, regardless of the validity of their responses.

The collected valid data were used to build an estimated probability distribution for each of the four contexts. The resulting empirical model is shown in Table 1. The model violates the Bell-CHSH inequality by 0.192 with a standard deviation of 0.176. Since the model is symmetric in the outcomes by construction, it is non-signalling and thus the measure of contextuality CNT_1 in the Cbd framework coincides with the degree of violation [25]. The symmetry in the outcomes also allows the violation to saturate the bound defined by CF in sheaf-theoretic framework [1], i.e. the following equality is attained

$$\max \left\{ 0, \frac{1}{2} \text{ violation of Bell-CHSH inequality} \right\} = \text{CF}. \quad (6)$$

Thus, our model is considered contextual in both the sheaf-theoretic framework and the Cbd framework.

To establish the significance of the contextuality result, we conducted bootstrap resampling to estimate the spread of the violation to the Bell-CHSH inequality. Simulated datasets were generated by random sampling with replacement from the original dataset. The resulting distribution of violations is depicted in Figure 3. Among the resampled datasets, 87% of them exhibited a positive violation, indicating that our experimental model demonstrates contextuality with a significance level of 87%.

5 Conclusions and Future Work

In this work, we employed the sheaf-theoretic framework for contextuality to model the Winograd Schema, originally formulated as an ambiguous coreference resolution task. Our findings revealed that the original Winograd Schema scenario lacked the necessary complexity to exhibit contextuality. To address this limitation, we introduced an additional ambiguous pronoun and a new pair of special and alternate words, creating a generalized Winograd Schema reminiscent of the Bell-CHSH scenario. Through crowdsourcing, we collected human judgments on an example of the generalized Winograd Schema and observed a contextual empirical model with a significance level of 87

An intriguing direction for future research involves constructing a comprehensive set of examples based on the proposed generalized Winograd Schema, thereby establishing it as a new challenge in the field of natural language processing. One potential approach is to leverage state-of-the-art generative

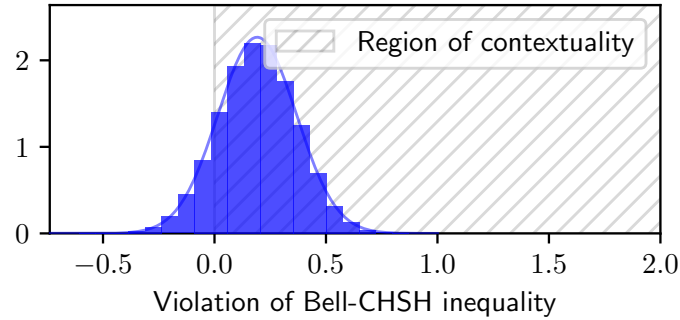


Figure 3: A normalised histogram of the Bell-CHSH inequality violation for 100,000 bootstrap samples from the model shown in Table 1. A positive violation, indicative of contextuality, is observed in 87% of the resampled models. The standard deviation of the distribution is 0.176.

language models such as GPT-4 to systematically generate examples of the schema with minimal human intervention. Careful prompt engineering would be needed to ensure that the generated examples are of high quality.

As collecting human judgments is costly and time-consuming, another alternative approach for constructing empirical models of the generalized Winograd Schema involves utilizing generative language models to generate responses to examples. This approach also offers an opportunity to explore the extent to which the responses generated by language models align with human responses. By comparing and analysing the correspondence between model-generated responses and human responses, one could gain insights into the capabilities and limitations of language models in capturing the way human beings understand language.

This paper presents an approach that consists of deliberately constructing sentences that exhibit contextuality. This strategy of “detecting contextuality in natural language” may invite criticism for its contrived nature.

An alternative approach could involve the application of mathematical frameworks designed for contextuality to analyze pre-existing natural language data, moving away from the intentional construction of examples with distinct features [35]. The aim of this strategy would not be to pursue contextuality within natural language. Instead, it would focus on developing novel methods for modelling natural language phenomena from a different perspective.

Acknowledgements

We are grateful to Daphne Wang for insightful discussions and the anonymous reviewers for their constructive comments. KL is supported by the Engineering and Physical Sciences Research Council [grant number EP/S021582/1]. MS is supported by the Royal Academy of Engineering research chair RCSR2122-14-152 on Engineered Mathematics for Modelling Typed Structures.

References

- [1] Samson Abramsky, Rui Soares Barbosa & Shane Mansfield (2017): *Contextual Fraction as a Measure of Contextuality*. *Physical Review Letter* 119, p. 050504, doi:10.1103/PhysRevLett.119.050504.
- [2] Samson Abramsky & Adam Brandenburger (2011): *The sheaf-theoretic structure of non-locality and contextuality*. *New Journal of Physics* 13(11), p. 113036, doi:10.1088/1367-2630/13/11/113036.
- [3] Samson Abramsky, Shane Mansfield & Rui Soares Barbosa (2012): *The Cohomology of Non-Locality and Contextuality*. *Electronic Proceedings in Theoretical Computer Science* 95, pp. 1–14, doi:10.4204/EPTCS.95.1.
- [4] Janet Anders & Dan E. Browne (2009): *Computational Power of Correlations*. *Physical Review Letter* 102, p. 050502, doi:10.1103/PhysRevLett.102.050502.
- [5] Alain Aspect, Jean Dalibard & Gérard Roger (1982): *Experimental Test of Bell's Inequalities Using Time-Varying Analyzers*. *Phys. Rev. Lett.* 49, pp. 1804–1807, doi:10.1103/PhysRevLett.49.1804.
- [6] J. S. Bell (1964): *On the Einstein Podolsky Rosen paradox*. *Physics Physique Fizika* 1(3), pp. 195–200, doi:10.1103/PhysicsPhysiqueFizika.1.195.
- [7] John S. Bell (1966): *On the Problem of Hidden Variables in Quantum Mechanics*. *Reviews of Modern Physics* 38(3), pp. 447–452, doi:10.1103/RevModPhys.38.447.
- [8] Adán Cabello, Simone Severini & Andreas Winter (2014): *Graph-theoretic approach to quantum correlations*. *Physical Review Letters* 112(4), pp. 1–5, doi:10.1103/PhysRevLett.112.040401.
- [9] John F. Clauser, Michael A. Horne, Abner Shimony & Richard A. Holt (1969): *Proposed Experiment to Test Local Hidden-Variable Theories*. *Physical Review Letters* 23(15), pp. 880–884, doi:10.1103/PhysRevLett.23.880.
- [10] Ehtibar N. Dzhafarov (2019): *The Contextuality-by-Default View of the Sheaf-Theoretic Approach to Contextuality*. doi:10.48550/arXiv.1906.02718.
- [11] Ehtibar N. Dzhafarov & Janne V. Kujala (2013): *All-Possible-Couplings Approach to Measuring Probabilistic Context*. *PLoS ONE* 8(5), p. e61712, doi:10.1371/journal.pone.0061712.
- [12] Ehtibar N. Dzhafarov & Janne V. Kujala (2016): *Contextuality-by-Default 2.0: Systems with Binary Random Variables*. doi:10.48550/arXiv.1604.04799.
- [13] Ehtibar N. Dzhafarov, Janne V. Kujala & Victor H. Cervantes (2015): *Contextuality-by-Default: A Brief Overview of Ideas, Concepts, and Terminology*. *Lecture Notes in Computer Science* 9535, 12–23, 2016, doi:10.1007/978-3-319-28675-4-2.
- [14] Ehtibar N. Dzhafarov, Janne V. Kujala, Víctor H. Cervantes, Ru Zhang & Matt Jones (2016): *On contextuality in behavioural data*. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374(2068), p. 20150234, doi:10.1098/rsta.2015.0234.
- [15] Albert Einstein, Boris Podolsky & Nathan Rosen (1935): *Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?* *Phys. Rev.* 47, pp. 777–780, doi:10.1103/PhysRev.47.777.
- [16] Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman & Jackie Chi Kit Cheung (2018): *A Knowledge Hunting Framework for Common Sense Reasoning*. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, pp. 1949–1958, doi:10.18653/v1/D18-1220.
- [17] Arthur Fine (1982): *Hidden Variables, Joint Probability, and the Bell Inequalities*. *Phys. Rev. Lett.* 48, pp. 291–295, doi:10.1103/PhysRevLett.48.291.
- [18] Marissa Giustina, Marijn A.M. Versteegh, Sören Wengerowsky, Johannes Handsteiner, Armin Hochrainer, Kevin Phelan, Fabian Steinlechner, Johannes Kofler, Jan-Åke Larsson, Carlos Abellán, Waldimar Amaya, Valerio Pruneri, Morgan W. Mitchell, Jörn Beyer, Thomas Gerrits, Adriana E. Lita, Lynden K. Shalm,

- Sae Woo Nam, Thomas Scheidl, Rupert Ursin, Bernhard Wittmann & Anton Zeilinger (2015): *Significant-Loophole-Free Test of Bell's Theorem with Entangled Photons*. *Physical Review Letters* 115(25), p. 250401, doi:10.1103/PhysRevLett.115.250401.
- [19] Weinan He, Canming Huang, Yongmei Liu & Xiaodan Zhu (2021): *WinoLogic: A Zero-Shot Logic-based Diagnostic Dataset for Winograd Schema Challenge*. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 3779–3789, doi:10.18653/v1/2021.emnlp-main.307.
- [20] B. Hensen, H. Bernien, A. E. Dreaú, A. Reiserer, N. Kalb, M. S. Blok, J. Ruitenberg, R. F. L. Vermeulen, R. N. Schouten, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, M. Markham, D. J. Twitchen, D. Elkouss, S. Wehner, T. H. Taminiu & R. Hanson (2015): *Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres*. *Nature* 526(7575), pp. 682–686, doi:10.1038/nature15759.
- [21] Mark Howard, Joel Wallman, Victor Veitch & Joseph Emerson (2014): *Contextuality supplies the 'magic' for quantum computation*. *Nature* 510(7505), pp. 351–355, doi:10.1038/nature13460.
- [22] Alexei Y. Kitaev (2003): *Fault-tolerant quantum computation by anyons*. *Annals of Physics* 303(1), pp. 2–30, doi:10.1016/s0003-4916(02)00018-0.
- [23] Simon Kochen & Ernst Specker (1968): *The Problem of Hidden Variables in Quantum Mechanics*. *Indiana Univ. Math. J.* 17(1), pp. 59–87, doi:10.1512/iumj.1968.17.17004.
- [24] Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus & Leora Morgenstern (2022): *The Defeat of the Winograd Schema Challenge*. doi:10.48550/arXiv.2201.02387.
- [25] Janne V. Kujala & Ehtibar N. Dzhafarov (2019): *Measures of Contextuality and Noncontextuality*. *Philosophical Transactions of the Royal Society A* 377:20190149, 2019 377(2157), p. 20190149, doi:10.1098/rsta.2019.0149.
- [26] Hector J. Levesque, Ernest Davis & Leora Morgenstern (2012): *The Winograd Schema Challenge*. In: *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'12*, AAAI Press, pp. 552–561, doi:10.5555/3031843.3031909.
- [27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov (2019): *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. doi:10.48550/arXiv.1907.11692.
- [28] Kin Ian Lo, Mehrnoosh Sadrzadeh & Shane Mansfield (2022): *A Model of Anaphoric Ambiguities using Sheaf Theoretic Quantum-like Contextuality and BERT*. *EPTCS* 366, 2022, pp. 23-34, doi:10.4204/EPTCS.366.5.
- [29] Cihan Okay, Aziz Kharoof & Selman Ipek (2022): *Simplicial quantum contextuality*. doi:10.48550/arXiv.2204.06648.
- [30] Robert Raussendorf (2013): *Contextuality in measurement-based quantum computation*. *Physical Review A - Atomic, Molecular, and Optical Physics* 88(2), pp. 1–7, doi:10.1103/PhysRevA.88.022322.
- [31] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula & Yejin Choi (2021): *WinoGrande: An Adversarial Winograd Schema Challenge at Scale*. *Commun. ACM* 64(9), pp. 99–106, doi:10.1145/3474381.
- [32] Lynden K. Shalm, Evan Meyer-Scott, Bradley G. Christensen, Peter Bierhorst, Michael A. Wayne, Martin J. Stevens, Thomas Gerrits, Scott Glancy, Deny R. Hamel, Michael S. Allman, Kevin J. Coakley, Shellie D. Dyer, Carson Hodge, Adriana E. Lita, Varun B. Verma, Camilla Lambrocco, Edward Tortorici, Alan L. Migdall, Yanbao Zhang, Daniel R. Kumor, William H. Farr, Francesco Marsili, Matthew D. Shaw, Jeffrey A. Stern, Carlos Abellán, Waldimar Amaya, Valerio Pruneri, Thomas Jennewein, Morgan W. Mitchell, Paul G. Kwiat, Joshua C. Bienfang, Richard P. Mirin, Emanuel Knill & Sae Woo Nam (2015): *Strong Loophole-Free Test of Local Realism*. *Phys. Rev. Lett.* 115, p. 250402, doi:10.1103/PhysRevLett.115.250402.
- [33] Alan M. Turing (1950): *Computing Machinery and Intelligence*. *Mind* 59(236), pp. 433–460, doi:10.1093/mind/LIX.236.433.
- [34] Daphne Wang (2018): *Distributional Models of Meaning The Contextuality of a Text*. MSc project report.

- [35] Daphne Wang & Mehrnoosh Sadrzadeh (2023): *The Causal Structure of Semantic Ambiguities*. doi:10.48550/arXiv.2206.06807.
- [36] Daphne Wang, Mehrnoosh Sadrzadeh, Samson Abramsky & Victor Cervantes (2021): *On the Quantum-like Contextuality of Ambiguous Phrases*. In: *Proceedings of the 2021 Workshop on Semantic Spaces at the Intersection of NLP, Physics, and Cognitive Science (SemSpace)*, Association for Computational Linguistics, Groningen, The Netherlands, pp. 42–52, doi:10.48550/arXiv.2107.14589.
- [37] Daphne Wang, Mehrnoosh Sadrzadeh, Samson Abramsky & Víctor H. Cervantes (2021): *Analysing Ambiguous Nouns and Verbs with Quantum Contextuality Tools*. *Journal of Cognitive Science* 22(3), pp. 391–420, doi:10.17791/jcs.2021.22.3.391.
- [38] Terry Winograd (1972): *Understanding natural language*. *Cognitive Psychology* 3(1), pp. 1–191, doi:10.1016/0010-0285(72)90002-3.
- [39] Hongming Zhang, Xinran Zhao & Yangqiu Song (2020): *WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge*. doi:10.48550/arXiv.2005.05763.