Quantum Contextuality in Natural Language

Kin Ian Lo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

25 March 2025

I, Kin Ian Lo, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Quantum computing has demonstrated computational advantages over classical computing, yet its applications to natural language processing (NLP) remain in their early stages. Quantum contextuality, a fundamental feature of quantum mechanics, has been identified as a resource for achieving these advantages. Meanwhile, ambiguities in human languages present challenges in building systems capable of understanding and generating natural language. Many of these ambiguities can be addressed by considering the context in which they arise. This thesis investigates the connection between quantum contextuality and ambiguities in natural language. It introduces a framework that models ambiguities in language as a form of measurement, analogous to physical experiments. The findings reveal that contextuality is present in natural language and that its degree is linked to the reasoning capabilities of large language models (LLMs).

Impact Statement

The advent of large language models (LLMs) like ChatGPT has significantly advanced the field of natural language processing (NLP), demonstrating incredible abilities in understanding and generating languages. These breakthroughs, however, raise important questions about the underlying mechanisms of these models and the nature of their intelligence.

At the same time, quantum computing has introduced a new paradigm of computing, using principles of quantum mechanics to solve certain problems more efficiently than classical computers. Quantum mechanics, a fundamental theory of physics, describes the behaviour of matter and energy at the smallest scales, where classical intuitions often fail. It has led to revolutionary advancements in various fields, including cryptography, optimization, and simulation of quantum systems. A central concept in quantum mechanics is contextuality, which challenges traditional ideas of realism and locality. Contextuality has been identified as a key resource for achieving computational advantages, enabling fault-tolerant quantum computation through techniques like magic state distillation and surface codes.

This research explores the intersection of quantum contextuality and natural language, examining how contextuality models linguistic ambiguities. This project introduces a novel framework connecting quantum contextuality with natural language, offering a theoretical basis for understanding language ambiguities through quantum principles. This work also demonstrates that contextuality can be linked to the reasoning capabilities of LLMs, providing a new perspective on evaluating their performance.

Although the impact of this work is mainly academic, it has the potential to influence the development of future AI systems and their applications in various fields. Recent studies have shown that LLMs suffer from hallucinations, which are instances where the model generates incorrect or nonsensical information. Resoning ability is a key factor in the performance of LLMs, and contextuality can provide an alternative way to evaluate the reasoning capabilities of these models and go beyond the limitations of traditional evaluation metrics. By bridging the gap between quantum mechanics and natural language processing, this research opens new avenues for understanding language and computation, paving the way for innovative approaches to AI and NLP.

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor, Prof. Mehrnoosh Sadrzadeh, for her invaluable guidance, insightful feedback, and unwavering support throughout my PhD journey. Her patience, encouragement, and expertise have been instrumental in shaping my research and academic growth.

I am also profoundly thankful to Shane Mansfield, Ehtibar Dzhafarov, and my colleagues at UCL for their guidance and thought-provoking discussions, which have greatly enriched my work. Their contributions and perspectives have been invaluable.

I would like to extend my heartfelt thanks to my office mates, Hadi Wazni, Daphne Wang, Lachlan McPheat, Mina Abbaszadeh, and Tilen Limback-Stokin, for their friendship, support, and the countless moments of laughter and encouragement. Their camaraderie has made this journey not only more enjoyable but also truly memorable.

Lastly, I am eternally grateful to my family for their unwavering love, support, and belief in me. Their encouragement has been my foundation and motivation throughout this journey.

Research Paper Declaration

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?**
   Generalised Winograd Schema and its Contextuality

   (b) **Please include a link to or doi for the work:**
   https://doi.org/10.4204/EPTCS.384.11

   (c) **Where was the work published?**
   Proceedings of the Twentieth International Conference on Quantum Physics and Logic

   (d) **Who published the work?**
   EPTCS (Electronic Proceedings in Theoretical Computer Science)

   (e) **When was the work published?**
   2023

   (f) **List the manuscript's authors in the order they appear on the publication:**
   Kin Ian Lo, Mehrnoosh Sadrzadeh and Shane Mansfield

   (g) **Was the work peer reviewed?**
   Yes

   (h) **Have you retained the copyright?**

Yes

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

https://arxiv.org/abs/2308.16498

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

   If 'Yes', please please give a link or doi:

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**

   (a) **Kin Ian Lo:** Development of the mathematical framework; Computation and data analysis; Original conceptualisation.

   (b) **Mehrnoosh Sadrzadeh:** Computational linguistics expertise; Data collection and analysis; Contribution to the original conceptualisation.

   (c) **Shane Mansfield:** Contextuality expertise; Contribution to the original conceptualisation.

4. **In which chapter(s) of your thesis can this material be found?**

   Chapter 7.

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

Date: 25 March 2025

Supervisor/Senior Author signature (where appropriate)**:**

Date: 25 March 2025

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?**

   (b) **Please include a link to or doi for the work:**

   (c) **Where was the work published?**

   (d) **Who published the work?**

   (e) **When was the work published?**

   (f) **List the manuscript's authors in the order they appear on the publication:**

   (g) **Was the work peer reviewed?**

   (h) **Have you retained the copyright?**

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   Quantum-Like Contextuality in Large Language Models

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

   If 'Yes', please please give a link or doi:

   https://doi.org/10.48550/arXiv.2412.16806

(c) **Where is the work intended to be published?**

Proceedings of the Royal Society A

(d) **List the manuscript's authors in the intended authorship order:**

Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield

(e) **Stage of publication:**

Revised version submitted.

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**

   (a) **Kin Ian Lo:** Development of the mathematical framework; Computation and data analysis; Original conceptualisation.

   (b) **Mehrnoosh Sadrzadeh:** Computational linguistics expertise; Data collection and analysis; Contribution to the original conceptualisation.

   (c) **Shane Mansfield:** Contextuality expertise; Contribution to the original conceptualisation.

4. **In which chapter(s) of your thesis can this material be found?**
   Chapter 6.

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

Date: 25 March 2025

Supervisor/Senior Author signature (where appropriate)**:**

Date: 25 March 2025

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?**
   A Model of Anaphoric Ambiguities using Sheaf Theoretic Quantum-like Contextuality and BERT

   (b) **Please include a link to or doi for the work:**
   https://doi.org/10.4204/EPTCS.366.5

   (c) **Where was the work published?**
   Proceedings End-to-End Compositional Models of Vector-Based Semantics

   (d) **Who published the work?**
   EPTCS (Electronic Proceedings in Theoretical Computer Science)

   (e) **When was the work published?**
   022

   (f) **List the manuscript's authors in the order they appear on the publication:**
   Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield

   (g) **Was the work peer reviewed?**
   Yes

   (h) **Have you retained the copyright?**
   Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**
   https://arxiv.org/abs/2208.05313

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server ’e.g. medRxiv’?**
   If ’Yes’, please please give a link or doi:

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript’s authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**

   (a) **Kin Ian Lo:** Development of the mathematical framework; Computation and data analysis; Original conceptualisation.

   (b) **Mehrnoosh Sadrzadeh:** Computational linguistics expertise; Data collection and analysis; Contribution to the original conceptualisation.

   (c) **Shane Mansfield:** Contextuality expertise; Contribution to the original conceptualisation.

4. **In which chapter(s) of your thesis can this material be found?**
   Chapter 7.

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

Date: 25 March 2025

Supervisor/Senior Author signature (where appropriate)**:**

Date: 25 March 2025

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?**
   Developments in Sheaf-Theoretic Models of Natural Language Ambiguities

   (b) **Please include a link to or doi for the work:**
   https://doi.org/10.4204/EPTCS.408.4

   (c) **Where was the work published?**
   Proceedings 13th International Workshop on Developments in Computational Models

   (d) **Who published the work?**
   EPTCS (Electronic Proceedings in Theoretical Computer Science)

   (e) **When was the work published?**
   2024

   (f) **List the manuscript's authors in the order they appear on the publication:**
   Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield

   (g) **Was the work peer reviewed?**
   Yes

(h) **Have you retained the copyright?**

Yes

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

https://arxiv.org/abs/2208.05313

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

(a) **What is the current title of the manuscript?**

(b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**
If 'Yes', please please give a link or doi:

(c) **Where is the work intended to be published?**

(d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**

(a) **Kin Ian Lo:** Development of the mathematical framework; Computation and data analysis; Original conceptualisation.

(b) **Mehrnoosh Sadrzadeh:** Computational linguistics expertise; Data collection and analysis; Contribution to the original conceptualisation.

(c) **Shane Mansfield:** Contextuality expertise; Contribution to the original conceptualisation.

4. **In which chapter(s) of your thesis can this material be found?**

   Chapter 7.

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

Date: 25 March 2025

Supervisor/Senior Author signature (where appropriate)**:**

Date: 25 March 2025

Contents

List of Figures

List of Tables

Introduction

Realism and non-contextuality are two fundamental principles that are believed to be at the core of classical physics. Realism posits that the value of a physical quantity exists independently of measurements, while non-contextuality posits that the value of a physical quantity is independent of what other quantities are being measured simultaneously. Non-contextuality is also related to locality, which means that information cannot travel faster than light. This ensures that what happens in one place does not immediately affect what happens elsewhere, so measurement results are independent of distant choices.

To better understand these principles, let us consider an illustrative example: imagine a London bus with three observable properties: its colour, the number of passengers, and its speed. Realism asserts that the values of these properties exist independently of observation. Non-contextuality, on the other hand, posits that the value of one property, such as the colour, should remain unaffected by which other properties are being observed. For instance, the colour of the bus should remain unchanged regardless of whether we observe the number of passengers or the speed.

However, in quantum physics, these two principles—realism and non-contextuality—are fundamentally incompatible. This incompatibility was first demonstrated in the 1960s by John Bell [3] and later by Simon Kochen and Ernst Specker [4]. Hypothetically, this means that in a quantum scenario, the colour of the bus we observe could appear red if we also observe the number of passengers, but it might appear blue if we instead observe the speed of the bus. Such behaviour highlights the contextual nature of quantum systems.

The study of contextuality has led to a deeper understanding of the fundamental differences between quantum and classical physics. Building on Bell's theoretical work, Clauser et al. [5] proposed an experiment using entangled photons to test Bell's inequalities. Aspect et al. [6] later performed such experiments, providing strong empirical evidence that quantum mechanics violates Bell's inequalities. These experimental violations confirm the contextual nature of quantum systems and demonstrate the incompatibility of realism and non-contextuality with quantum physics.

Recently, it has been shown that contextuality is linked to the computational advantage of quantum computers, which is an application of quantum technologies that promises speed-ups over classical computers. Anders and Browne were the first to note that correlations in quantum systems can be seen as a source of computational power in the context of measurement-based quantum computation (MBQC) [7]. Raussendorf strengthened this connection by showing that the computational power of quantum correlations is linked to the degree of contextuality in the quantum state [8]. Howard et al. [9] showed that contextuality is necessary for magic state distillation [10], which is a key step in the realisation of fault-tolerant universal quantum computation via the error-correcting scheme known as the surface code [11].

On the other hand, human languages exhibit another form of contextuality, where the meaning of a word or phrase can depend on the context in which it is used. However, this form of contextuality differs from that found in quantum physics, as the change in meaning arises from the *causal influence* of the context. For example, when a reader interprets the phrase "slippy bank", they first see the word "slippy" and then the word "bank". The meaning of "bank" is casually influenced by the preceding word "slippy", leading the reader to interpret "bank" as the side of a river rather than a financial institution. This type of contextuality is more akin to signalling in quantum physics, where information can be transmitted from one part of a system to another. Hence, disambiguation is a central challenge in natural language processing (NLP), since the meaning of a word or phrase can vary depending on its context and, at times, the common sense or background knowledge of the reader.

This linguistic contextuality can be seen as a form of quantum contextuality, where the meaning of a word or phrase is not fixed but depends on the context in which it is used. This has led to the idea of formalising ambiguity in natural language through the lens of quantum contextuality, with the hope of using this formalism to develop new methods for natural language processing.

In this thesis, we make a step towards this goal by uncovering a form of quantum contextuality in a type of ambiguity known as anaphoric ambiguity, which arises when a word or phrase refers to another word or phrase that appears earlier in the text. We construct explicit examples of ambiguous English sentences and show that they exhibit quantum contextuality. In doing so, we shed light on the quantum nature of natural language understanding and pave the way for quantum-inspired methods for natural language processing.

Structure of the thesis

Part I of this thesis introduces the necessary background knowledge used in the rest of the thesis.

- Chapter 1 introduces the basic concepts in quantum physics that are used in this thesis.
- Chapter 2 reviews the literature on natural language processing and in particular coreference resolution.
- Chapter 3 introduces the concept of contextuality and related concepts in quantum physics.

Part II presents the main findings of this thesis, which involve demonstrating quantum contextuality within natural language and initial results on making use of contextuality to evaluate language models.

Published Contributions

This thesis contains results that have been published in the following paper:

- **Title:** Generalised Winograd Schema and its Contextuality
  Authors: **Kin Ian Lo**, Mehrnoosh Sadrzadeh and Shane Mansfield

Published in: EPTCS as part of the proceedings of the 20th International Conference on Quantum Physics and Logic (QPL 2023)
Publication date: 2023
Material presented in thesis in: Chapter 7

- **Title:** A Model of Anaphoric Ambiguities using Sheaf Theoretic Quantum-like Contextuality and BERT
  Authors: **Kin Ian Lo**, Mehrnoosh Sadrzadeh and Shane Mansfield
  Published in: EPTCS as part of the Proceedings End-to-End Compositional Models of Vector-Based Semantics
  Publication date: 2022
  Material presented in thesis in: Chapter 7

- **Title:** Developments in Sheaf-Theoretic Models of Natural Language Ambiguities
  Authors: **Kin Ian Lo**, Mehrnoosh Sadrzadeh and Shane Mansfield
  Published in: EPTCS as part of the proceedings of the 13th International Workshop on Developments in Computational Models of (DCM 2023)
  Publication date: 2023
  Material presented in thesis in: Chapter 6

- **Title:** Quantum-Like Contextuality in Large Language Models
  Authors: **Kin Ian Lo**, Mehrnoosh Sadrzadeh and Shane Mansfield
  Submitted to: Proceedings of the Royal Society A
  Status: Revised version submitted
  Material presented in thesis in: Chapter 6

Part I

Background

Chapter 1

Quantum Mechanics

The study of the quantumness of physical systems dates back to the 19th century, when the physics community was puzzled by the ultraviolet catastrophe, where classical physics predicted that the electromagnetic heat radiation would scale to infinite energy at high frequencies. It was later resolved by the German physicist Max Planck in 1900, who proposed that energy is quantised, and that the energy of a photon is proportional to its frequency [12]. This was a departure from classical physics, in which every physical system was assumed to be composed of continuous, infinitely divisible quantities.

In 1905, Albert Einstein successfully explained the photoelectric effect using Plank's quantum hypothesis [13], for which he was awarded the Nobel Prize in Physics in 1921. The next big step was De Broglie's hypothesis in 1924, which stated that not only light has both wave-like and particle-like properties, but that particles also exhibit wave-like behaviour [14]. Waves, unlike particles, are able to interfere and diffract, leading to the development of quantum mechanics. In the following years, several physicists made crucial contributions to the field. Erwin Schrödinger developed wave mechanics and his famous equation describing the evolution of quantum systems [15]. Werner Heisenberg formulated matrix mechanics and the uncertainty principle [16, 17]. Paul Dirac unified these approaches and developed the bra-ket notation widely used in quantum mechanics today [18]. John von Neumann formalised these ideas into a comprehensive mathematical framework making use of Hilbert spaces and operators by the early 1930s [19].

A key concept in quantum mechanics is the description of quantum systems using vectors in a Hilbert space, and observables as operators acting on these vectors. Composite systems are described using the tensor product of the individual systems' state spaces. These basic concepts will be introduced in the next section.

In the rest of this chapter, I will give a brief overview of the basic concepts of orthodox quantum mechanics used in the text book by Nielsen and Chuang [20].

Quantum States In quantum mechanics, the state of a physical system is described by a vector in a complex vector space known as a Hilbert space. Vectors, or quantum states, are denoted by the ket notation $|\psi\rangle$, and the corresponding dual vectors are denoted by the bra notation $\langle\psi|$. The inner product of a ket and a bra is written as $\langle\psi|\phi\rangle$, which is a complex scalar number. A state is called normalised if its inner product with itself is 1, i.e. $\langle\psi|\psi\rangle = 1$.

For a simple two-level quantum system (like the spin of an electron or polarization of a photon), this state can be represented as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

where $|0\rangle$ and $|1\rangle$ are the orthonormal basis states (often referred to as computational basis in the context of quantum computing), and α and β are complex numbers that satisfy the normalization condition $|\alpha|^2 + |\beta|^2 = 1$. This state $|\psi\rangle$ is known as a quantum bit or qubit, the fundamental unit of quantum information. The fact that the state is a linear combination of basis states is what allows for the description of superposition.

Generally speaking, the state of a quantum system can be written as a linear combination of basis states in the form:

$$|\psi\rangle = \sum_i \alpha_i |i\rangle$$

where $|i\rangle$ are basis states and α_i are complex coefficients satisfying the normalization condition $\sum_i |\alpha_i|^2 = 1$.

Quantum Measurement

Measurement in quantum mechanics is described by a kind of mathematical operations known as observables, which are linear operators acting on the Hilbert space. Measurement made on a quantum system against an observable will yield one of the eigenvalues of the observable, and the system will collapse to the corresponding eigenstate. Hence, a proper observable A must satisfy the following properties:

- be diagonalizable, i.e. have a complete set of eigenvectors,
- have real eigenvalues,
- have mutually orthogonal eigenvectors.

It turns out these properties are equivalent to the operator being Hermitian: $A = A^\dagger$. Hence, we say that any valid observable must be an Hermitian operator.

The probability of measuring a particular eigenvalue a_i of an observable A from a state $|\psi\rangle$ is given by the Born rule:

$$P(a_i) := |\langle i|\psi\rangle|^2 = \langle\psi|i\rangle\langle i|\psi\rangle$$

where $|i\rangle$ is the eigenvector corresponding to the eigenvalue a_i.

Sometimes we would be interested in the expectation value of an observable A in a state $|\psi\rangle$, which is given by:

$$\langle A\rangle = \langle\psi|A|\psi\rangle.$$

Since the observable is Hermitian, we can write it as a sum of its eigenvalues and eigenvectors (the eigen-decomposition):

$$A = \sum_i a_i|i\rangle\langle i|.$$

Rewriting the expectation value in terms of the eigen-decomposition gives

$$\begin{aligned}\langle A\rangle &= \langle\psi|A|\psi\rangle \\ &= \langle\psi|\left(\sum_i a_i|i\rangle\langle i|\right)|\psi\rangle \\ &= \sum_i a_i\langle\psi|i\rangle\langle i|\psi\rangle \\ &= \sum_i a_i P(a_i),\end{aligned}$$

which is the weighted average of the eigenvalues of the observable, where the weights are the probabilities of measuring the corresponding eigenvalues.

A commonly used observable in the Pauli-Z operator, which is defined as:

$$Z = |0\rangle\langle 0| - |1\rangle\langle 1|,$$

where the eignvalue of $|0\rangle$ is 1 and the eigenvalue of $|1\rangle$ is -1. However, in the field of quantum computing, we often use a slightly variant of the Pauli-Z operator, which is defined as:

$$M = 0\,|0\rangle\langle 0| + 1\,|1\rangle\langle 1| = |1\rangle\langle 1|,$$

where the eigenvalue of $|0\rangle$ is 0 and the eigenvalue of $|1\rangle$ is 1.

As an example, consider a single qubit in the state:

$$|\psi\rangle = \sqrt{\frac{1}{3}}|0\rangle + \sqrt{\frac{2}{3}}|1\rangle.$$

To measure this qubit in the computational basis $\{|0\rangle, |1\rangle\}$, we use the observable $M = |1\rangle\langle 1|$. The probability of measuring the eigenvalue 0 and corresponding state $|0\rangle$ is:

$$P(0) = |\langle 0|\psi\rangle|^2 = \left|\frac{1}{\sqrt{3}}\right|^2 = \frac{1}{3}.$$

Similarly, the probability of measuring the eigenvalue 1 and corresponding state

$|1\rangle$ is:

$$P(1) = |\langle 1|\psi\rangle|^2 = \left|\sqrt{\frac{2}{3}}\right|^2 = \frac{2}{3}.$$

The expectation value of the observable M in the state $|\psi\rangle$ is:

$$\langle M\rangle = \langle\psi|M|\psi\rangle = \langle\psi|1\rangle\langle 1|\psi\rangle = \frac{2}{3}.$$

This way of interpreting measurement in quantum mechanics is known as the Measurement Postulate, which states that the outcome of a measurement is probabilistic, and the state of the system collapses to the eigenstate corresponding to the measured eigenvalue.

That concludes the introduction to measurement in quantum mechanics, which is critical to the understanding of the concept of contextuality, which will be discussed later in this thesis.

Composite Systems and Entanglement The state of a composite system is described by the tensor product of the individual systems' states, where a tensor product of two states $|a\rangle$ and $|b\rangle$ is defined as:

$$|a\rangle \otimes |b\rangle := |a,b\rangle := |ab\rangle$$

Superposition of composite systems like above is what allows for the phenomenon of entanglement, which is a unique quantum phenomenon where the quantum states of two or more particles become correlated in such a way that the state of each particle cannot be described independently of the state of the others, even when the particles are separated by large distances. As an example, consider the equal superposition of $|00\rangle$ and $|11\rangle$:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle),$$

which is one of the four Bell states. The rest of the Bell states are:

$$|\Phi^-\rangle = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle),$$
$$|\Psi^+\rangle = \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle),$$
$$|\Psi^-\rangle = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle).$$

The Bell states are maximally entangled states, and hence plays a crucial role in quantum information theory. A composite state is non-entangled or separable if it can be written as a product of the individual systems' states. One example of a separable state is:

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |10\rangle) = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \otimes |0\rangle = |+\rangle \otimes |0\rangle,$$

for it can be written as a product of the plus state $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and the zero state $|0\rangle$.

State evolution and Quantum Gates The evolution of quantum states is determined by the Hamiltonian H of the system, which is a Hermitian operator whose eigenvalues correspond to the energy levels of the system. The time evolution of a quantum state $|\psi(t)\rangle$ is given by the Schrödinger equation:

$$i\hbar \frac{d}{dt}|\psi(t)\rangle = H|\psi(t)\rangle,$$

where $\hbar$ is the reduced Planck constant. The solution to this equation is given by the unitary operator:

$$U(t) = e^{-iHt/\hbar},$$

which is known as the time evolution operator. The time evolution of a quantum state is then given by:

$$|\psi(t)\rangle = U(t)|\psi(0)\rangle.$$

The unitarity of the time evolution operator ensures that the total probability of all possible outcomes remains 1, that is, $\langle\psi(t)|\psi(t)\rangle = 1$ for all t.

In quantum computing, the continuous time evolution is abstracted into discrete quantum operations or gates, which are unitary operators acting on the state space of the system.

The first gate we will introduce is the identity gate, which is defined as:

$$I = |0\rangle\langle 0| + |1\rangle\langle 1| = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which leaves the state of the system unchanged.

Another set of commonly used gates are the Pauli gates, which are defined as:

$$\begin{aligned} X &= |0\rangle\langle 1| + |1\rangle\langle 0| = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \\ Z &= |0\rangle\langle 0| - |1\rangle\langle 1| = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \\ Y &= iXZ = i(|0\rangle\langle 1| - |1\rangle\langle 0|) = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}. \end{aligned}$$

The Pauli-X gate, also known as the bit-flip gate, swaps the amplitudes of the basis states $|0\rangle$ and $|1\rangle$. The Pauli-Z gate, also known as the phase-flip gate, leaves the basis state $|0\rangle$ unchanged while flipping the sign of the basis state $|1\rangle$. The Pauli-Y gate is less intuitive but could be thought as a combination of the bit-flip and phase-flip gates up to a global phase factor.

The Pauli gates are both unitary and Hermitian. That means they are gates as well as Hamiltonians. For example, the evolution operator generated by the Pauli-X gate is given by:

$$U_X(t) = e^{-iXt/\hbar},$$

which can be interpreted as a rotation of the state vector around the X-axis of the

Bloch sphere.

Another commonly used gate is the Hadamard gate, which is defined as:

$$H = \frac{1}{\sqrt{2}}(X+Z) = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

One useful property of the Hadamard gate is that it maps the basis states to the plus and minus states:

$$\begin{aligned} H|0\rangle &= |+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), \\ H|1\rangle &= |-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle). \end{aligned}$$

So far we have introduced common single-qubit gates. In quantum computing, we are often interested in entangled states, which require the use of multi-qubit gates. One of the most important multi-qubit gates is the CNOT gate, which is a two-qubit gate that flips the second qubit if the first qubit is in the state $|1\rangle$. The CNOT gate is defined as:

$$\text{CNOT} = |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

To prepare the Bell state $|\Phi^+\rangle$, we can start with the state $|00\rangle$ and apply the Hadamard gate to the first qubit followed by the CNOT gate with the first qubit as the control and the second qubit as the target. Explicitly, the preparation of the Bell state $|\Phi^+\rangle$ is given by:

$$\begin{aligned}
|\Phi^+\rangle &= \text{CNOT}(H \otimes I)|00\rangle \\
&= \text{CNOT}\left(\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \otimes |0\rangle\right) \\
&= \text{CNOT}\left(\frac{1}{\sqrt{2}}(|00\rangle + |10\rangle)\right) \\
&= \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle).
\end{aligned}$$

To summaries, we have introduced the basic concepts of quantum mechanics, including quantum states as vectors in a Hilbert space, quantum measurements of observables, composite systems and entanglement, and quantum gates for state manipulation. In particular, quantum measurement is a crucial for making sense of the examples discussed in the later chapters, where we will explore the concept of contextuality in quantum physics and more broadly in fields other than physics.

I will wrap up this chapter by introducing the four postulates of quantum mechanics, which are the axioms that form the foundation of modern quantum theory. The postulates are as follows:

1. **Postulate 1:** The state of an isolated quantum system is described by a vector $|\psi\rangle$ in a complex vector space known as a Hilbert space.

2. **Postulate 2:** The evolution of the state of a closed quantum system $|\psi\rangle$ is described by a unitary operator U acting on the state space:

$$|\psi(t)\rangle = U(t)|\psi(0)\rangle.$$

3. **Postulate 3:** Quantum measurement of the observable A correspond to collapsing the state $|\psi\rangle$ to one of the eigenstates $|i\rangle$ of A with probability $P(a_i) = |\langle i|\psi\rangle|^2$ (Born rule).

4. **Postulate 4:** The state of a composite system is described by the tensor product of the individual systems' states: $|\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots |\psi_n\rangle$.

Chapter 2

Natural Language Understanding

Language is a cornerstone of human civilization, enabling communication to organise complex societies and the transfer and accumulation of knowledge across generations. Teaching machines to understand human languages allows for a more natural way for humans to interact with computers, and allows machines to tap into the vast amount of information accumulated in text written in numerous human languages. Machine understanding of human languages can also foster communication between people who speak different languages, and can help bridge the digital divide by making information more accessible to people less familiar with technology. In recent years, the field of Natural Language Processing (NLP) has made significant strides in developing transformer-based models [21] that can understand and generate human language, paving the way for a new era of human-computer interaction.

The significance of language understanding in Artificial intelligence is underscored by the Turing Test [22], proposed by Alan Turing in the 1950s, which aims to assess whether a machine can exhibit intelligent behaviour similar to, or indistinguishable from, that of a human. The field of Natural Language Processing (NLP) has a rich history that spans several decades, evolving alongside advancements in better computational resources and innovation in algorithms and machine learning techniques.

The origins of NLP can be traced back to the 1950s. One of the earliest and most famous experiments in machine translation was the Georgetown-IBM experiment in 1954 [23, 24]. By making use of 6 hand-crafted rules and the dictionary of

250 lexical items between the two languages, the program was able to translate 60 carefully chosen Russian sentences into English, demonstrating the feasibility of machine translation, but it did not scale. Figure 2.1 demonstrates the flowchart used in the translation system and a few examples of the translations produced by the system are shown below:

Russian sentence	**English Translation**
Качество угля определяется его калорийностью.	The quality of coal is determined by calory content.
Крахмал вырабатывается механическим путем из картофеля.	Starch is produced by mechanical method out of potatoes.
Величина угла определяется отношением длины дуги к радиусу.	Magnitude of angle is determined by the relation of length of arc to radius.
Мы передаем мысли посредством речи.	We transmit thoughts by means of speech.
Железо добывается из руды химическим процессом.	Iron is obtained out of ore by chemical process.
Динамит приготовляется химическим путем из нитроглицерина с примесью инертного материала.	Dynamite is prepared by chemical method out of nitroglycerine with admixture of inert material.
Военный суд приговорил сержанта к лишению гражданских прав.	Military court sentenced the sergeant to deprival of civil rights.

Throughout the 1960s and 1970s, NLP research continue to focused on rule-based systems. These systems relied on hand-crafted rules to parse and understand language. Notable projects during this era included ELIZA [25], a rule-based chatbot that could engage in dialogue by pattern matching and substitution, and SHRDLU [26], a program that could understand and execute commands in a simplified blocks world.

ELIZA was the first chatbot that could engage in a conversation with a human user that made some believe that the program was truly intelligent. It was developed by Joseph Weizenbaum at MIT in 1966 and was designed to simulate a Rogerian

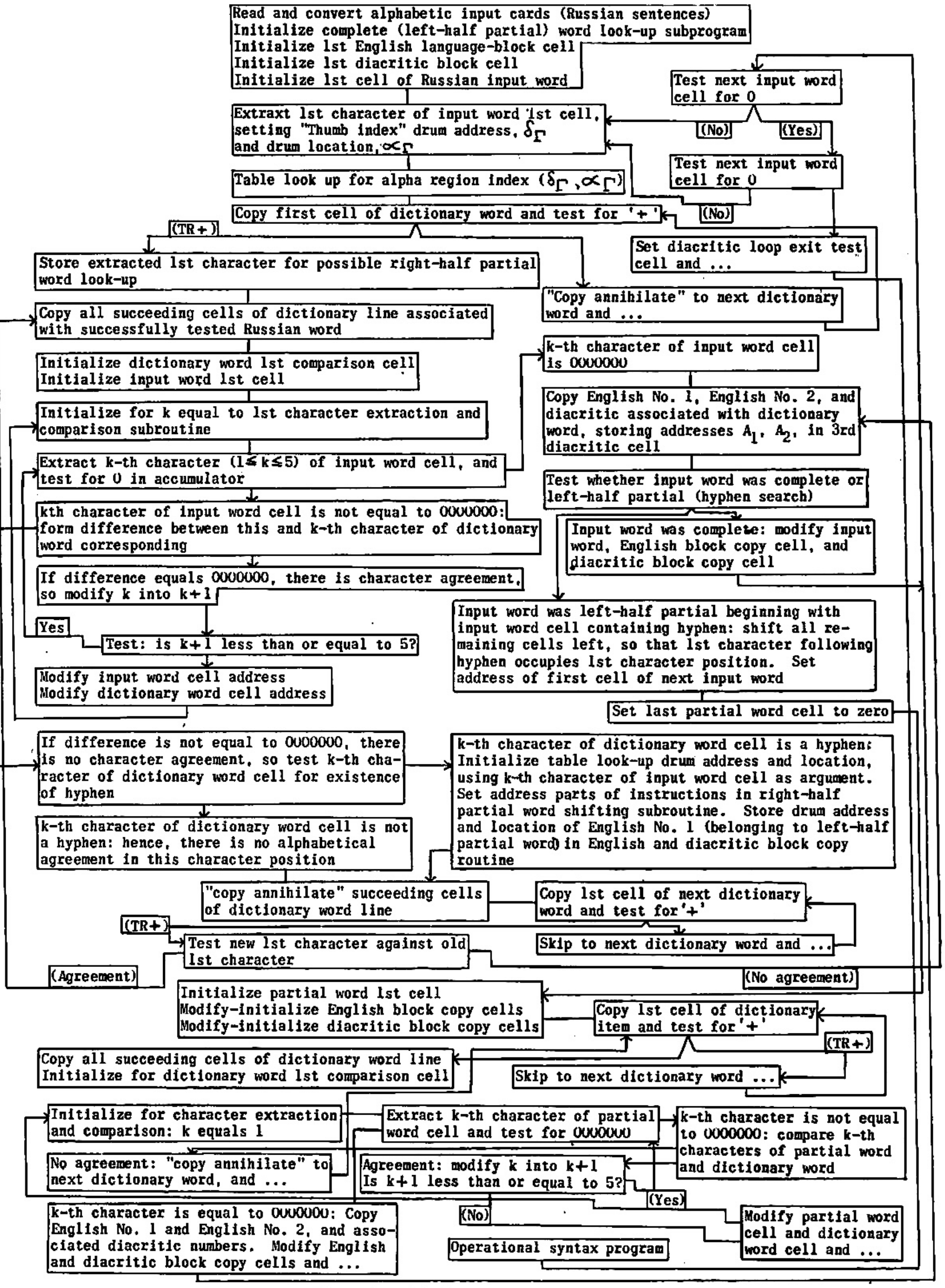


Figure 1 — Dictionary Syntax Flow Chart

Figure 2.1: The flowchart of the Georgetown-IBM translation system. Image taken from Figure 1 in [1].

psychotherapist. An ELIZA response is generated by transforming the user's input using a set of rules. For example, if the user's input is "I am sad", ELIZA will match it with the pattern "I am (.*)" and generate the response "How long have you been

\$1?", where \$1 is a placeholder for the captured text "sad". The following is my conversation with ELIZA which starts with me expressing my struggles to write this exact thesis:

ELIZA

Me: I am **struggling to write my thesis**.

ELIZA: How long have you been **struggling to write your thesis**?

Me: Far too long to be honest.

ELIZA: I think you are blocking what you really want to say.

Me: How can I be motivated?

ELIZA: I'm not convinced that you truly want to be motivated.

Me: I truly want to be motivated.

ELIZA: I'm not sure I understand you fully.

Me: It's ok.

ELIZA: OK... "How can you be motivated". Tell me more.

The bolded text indicates the matched patterns and the responses generated by ELIZA using the `(I am (.*), How long have you been $1?)` rule (where `$1` is the captured text), and the pronoun substitution rule `(your, my)`. Eventually ELIZA failed to believe that I truly want to be motivated to write this thesis.

The 1980s and 1990s saw a paradigm shift towards statistical methods in NLP. This approach, driven by the increasing availability of textual data in digital forms and computational power, relied on probabilistic models trained on large corpora. Key developments included Hidden Markov Models for speech recognition [27] and part-of-speech tagging [28], statistical machine translation models [29, 30], and N-gram language models for text prediction and generation[31].

From the 2010s onwards, neural network-based models have become increasingly prevalent in NLP. In particular, Recurrent Neural Networks (RNNs) [32] and Long Short-Term Memory (LSTM) [33] have been widely used for sequence modelling tasks like machine translation and sentiment analysis due to their ability to capture long-range dependencies in sequences. These advancements have dramatically improved performance on various NLP tasks, including machine translation, sentiment analysis, and question answering. The attention mechanism [34] was introduced in 2014 to improve the performance of neural machine translation models by allowing the model to focus on longer-range dependencies in the input sequence. In fact, the Transformer architecture [21] was introduced in 2017, which demonstrated the effectiveness of self-attention mechanisms in capturing long-range dependencies in sequences, without the need for recurrent connections. The Transformer architecture has since become the basis for many state-of-the-art NLP models, including BERT [35], GPTs [36], T5 [37], LLamas [38], and the recent state-of-the-art reasoning model DeepSeek R1 [39] which achieved on-par performance with OpenAI's 4o model on the mathematical reasoning benchmarks AIME 2024 and MATH-500. For instance, see below the same conversation as the one above but instead with DeepSeek:

Conversation with DeepSeek V3

Me: I am struggling to write my thesis.

DeepSeek V3: Writing a thesis can be a challenging and overwhelming process, but breaking it down into manageable steps can make it more achievable. Here are some tips to help you get started and stay on track:

1. **Clarify Your Topic and Structure**

Ensure your research question or thesis statement is clear and focused. Create an outline with chapters or sections (e.g., Introduction, Literature Review, Methodology, Results, Discussion, Conclusion). Break each section into smaller tasks to make progress feel less daunting.

2. **Set Realistic Goals**

Set daily or weekly writing goals (e.g., "Write 500 words today" or "Complete the Literature Review by Friday"). Use tools like a planner, calendar, or apps (e.g., Trello, Notion) to track your progress.

3. **Start Writing, Even If It's Imperfect**

Don't aim for perfection in the first draft. Just get your ideas down on paper.

Use placeholders for sections you're unsure about (e.g., "Add citation here" or "Expand on this point later").

4. **Work on Sections Out of Order**

If you're stuck on one section, move to another. For example, if the Introduction feels overwhelming, start with the Methodology or Literature Review. Writing something—anything—can help build momentum.

5. **Schedule Regular Writing Time**

Dedicate specific blocks of time to writing, even if it's just 30 minutes a day. Find a time of day when you're most productive and stick to it.

6. **Seek Feedback Early**

Share drafts with your advisor, peers, or a writing group to get constructive feedback. Don't wait until the end to seek input—early feedback can save you time and effort.

7. **Manage Perfectionism and Procrastination**

Remember that your thesis doesn't have to be perfect; it just needs to be done. Break tasks into smaller steps to avoid feeling overwhelmed. Use techniques like the Pomodoro Technique (25 minutes of focused work, 5-minute breaks) to stay productive.

8. **Take Care of Yourself**

Writing a thesis is a marathon, not a sprint. Take breaks, exercise, and get enough sleep. Avoid burnout by setting boundaries and rewarding yourself for milestones.

9. **Use Resources**

Check if your university offers writing workshops, thesis boot camps, or one-on-one consultations. Use reference management tools like Zotero, Mendeley, or EndNote to organise citations. Read other theses in your field to get a sense of structure and style.

10. **Stay Motivated**

Remind yourself why your research matters and how far you've come. Cel-

> ebrate small wins, like finishing a section or hitting a word count goal. If you're feeling stuck on a specific part of your thesis, let me know, and I can help you brainstorm or provide more targeted advice. You've got this!

Neural translation services are now widely used for communication, making it easier for people to interact across different languages. They are also used to create parallel corpora for businesses and government organizations. Popular tools include Google Translate, while platforms like Lokalise and Weglot focus on low-resource languages. Other tools like eLUNA and eTranslation are designed for use by the United Nations and the European Commission.

In summary, at the time of writing this thesis, the field of NLP has made significant progress in developing models that can understand and generate natural language text. These models have been applied to a wide range of tasks, including machine translation, sentiment analysis, question answering, code generation, and problem-solving.

The recent advancements in NLP have been driven by the availability of large-scale datasets, powerful computational resources, and innovative algorithms. However, there are still many challenges to be addressed in NLP, including the need for more explainable and interpretable models, to shed a light into the inner workings of neural learning models, and the energy efficiency of language models.

2.1 Ambiguity in Natural Language

Ambiguity is a major challenge in NLP. Words and sentences can have multiple meanings depending on the context in which they are used. For example, the word "bank" can refer to a financial institution or the side of a river. Similarly, "book" might refer to the physical object or the content within it. These examples illustrate a type of ambiguity known as *lexical ambiguity*, where a word has multiple meanings, and the intended meaning must be inferred from the context in which it is used. In particular, the first example (bank) is an example of *homonymy*, where two words have the same form but different meanings, while the second example (book) is an example of *polysemy*, where a single word has multiple related meanings, called *senses*.

Word-level ambiguity is also known as lexical ambiguity. Early attempts to resolve this issue led to the formulation of Word Sense Disambiguation (WSD), a problem first introduced by Warren Weaver [40]. In the 1950s, initial solutions were predominantly rule-based, using manually curated resources such as dictionaries and thesauri. By the 1970s, researchers began incorporating semantic roles—such as hypernymy and hyponymy—into disambiguation systems. 1980's and 90s saw the introduction of networks of semantic roles, leading to the development of resources such as WordNet [41], FrameNet [42], and ConceptNet [43] – large-scale graphs where the nodes are words and the edges are semantic relations. In particular, WordNet provided a comprehensive lexical database for English, FrameNet focused on verb-centric frames and their participants, and ConceptNet offered a commonsense knowledge graph emphasizing conceptual relationships. Statistical methods emerged in the late 20th century, using clustering techniques on large corpora to automatically classify word meanings based on context. Since 2010, these approaches have been significantly advanced by deep neural network algorithms and specifically transformer models such as BERT [35].

Coreference ambiguity is another type of ambiguity that arises when a pronoun or noun phrase refers to a previously mentioned entity. For example, the sentence "John sleeps. He snores." contains two *mentions* of the same entity, the human being "John". The first mention is simply "John" in the first sentence and the second is "He" in the second sentence. In fact, this is an example of an *anaphor*, where "He" refers *back* to "John". An anaphor is a noun phrase (usually a pronoun) that refers back to a previously mentioned noun phrase. On the other hand, a *cataphor* is a noun phrase that refers *forward* to a noun phrase that will be mentioned later in the text. For example, in the sentence "When he arrived, John was tired.", the pronoun "he" is a cataphor that refers forward to "John". Cataphora usually occurs in compound sentences where the pronoun was mentioned in the first clause and the referent is mentioned in the second clause. The task of coreference resolution is to determine which mentions in a text refer to the same entity.

Early work on coreference resolution focused on rule-based systems that used hand-crafted rules to identify coreferent mentions based on syntactic and semantic information. Hobbs [44] introduced a syntactic algorithm for pronoun resolution using parse trees and a set of semantic rules on gender, number, and animacy (the distinction between living and non-living entities). However, Hobbs' algorithm was limited to simple pronoun resolution and did not handle more complex cases such as those involving ellipsis, as "John played the piano so did Mary." It also did not encode thematic preferences such when the subject of the first sentence will also often be the reference of the subject pronoun of the second one, as in "John met Bill in a Café. He bought some coffee.". In this example, almost certainly "He" refers to "John" and not "Bill". Soon after, Lappin and Leass [45] proposed a more general algorithm for pronoun resolution that used a set of constraints based on syntactic and semantic information to identify coreferent mentions. Neither of these systems and none of the ones the improvements that followed could deal with cases where common sense reasoning and world knowledge helped. As a result, they could not resolve many of the cases that were easy for humans, as in "The trophy did not fit in the suitcase. It was too small.". We will see below that this weakness led to a famous challenge argued to be a better test of machine intelligence than Turing's test.

2.2 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, proposed in 2011 by Hector Levesque et al. [46]. The challenge was inspired by Terry Winograd's seminal paper [47], where he contended that syntax alone falls short in the interpretation of natural language, necessitating common sense or world knowledge as well. Every Winograd schema is a pair of sentences that differ only in one word or phrase, and the task is to determine the referent of an ambiguous pronoun in both sentences. The WSC is a collection of such sentence pairs, demanding common-sense reasoning abilities to disambiguate the pronouns.

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

City Councilmen Example

The city councilmen refused the demonstrators a permit because **they** *feared* violence.

The city councilmen refused **the demonstrators** a permit because **they** *advocated* violence.

Note that the two sentences differ only in the adjective *feared* and *advocated*. The ambiguous pronoun **they** 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 aligns with the common sense that city councilmen are the ones who tend to worry about violence and would refuse a permit to prevent any violence in demonstrations. In the second sentence, the pronoun *they* refers to the *demonstrators*, as it is within the common sense that demonstrators may advocate for violence which would lead to the refusal of a permit for a demonstration.

Another classic example of a Winograd schema proposed by Levesque et al. [46] is the following pair of sentences:

Trophy-Suitcase Example

The trophy doesn't fit into **the suitcase** because **it** is too *small*.

The trophy doesn't fit into the suitcase because **it** is too *large*.

The pronoun *it* in the first sentence refers to the trophy, while in the second sentence, it refers to the suitcase. Answering this question correctly requires understanding the common sense that an object larger than the container cannot fit inside it.

Since a Winograd Schema is a pair of sentences that differ slightly, we can compactly represent the pair of sentences as well as the answers in the following compact format:

Winograd Schema Compact Notation

The trophy doesn't fit into **the suitcase** because **it** is too [small/large]. What is too [small/large]?

Answer: The suitcase/the trophy.

Here the pair of square brackets encloses the two possible word choices, each leading to a different sentence. Usually we refer to the first option as the *special* word and the second option as the *alternate* word. The answers to both options are provided in the same order as the options in the square brackets. This notation will be employed throughout the paper.

The correctness of the answer to a Winograd schema question is unambiguous and can be judged by determining whether the correct referent of the ambiguous pronoun is identified. A machine that achieves accuracies on-par with humans on the WSC is considered to have human-level commonsense reasoning abilities. In contrast, the Turing Test has been criticised for its subjective and ambiguous evaluation criteria. Originally proposed as the imitation game by Turing [22], the test involves a human judge engaging in a textual conversation with a machine and a human. The goal is for the judge to determine which participant is the machine. If the judge or a panel of judges cannot reliably distinguish the machine from the human, the machine is considered to have passed the test. However, the unrestricted nature of the Turing Test allows for potential exploitation, as machines can rely on deceptive tactics or avoid answering certain questions to obscure their non-human nature. In fact, for a machine to pass the test, it must deceive about its identity, as machines inherently lack biological attributes. If questioned about physical traits, such as skin colour or heart rate, the machine is compelled to fabricate responses to convincingly pose as a human. Due to its more straightforward evaluation process compared to the Turing Test, the WSC was proposed as a more practical and objective alternative for assessing machine intelligence.

A major issue with the WSC is that it is over-constrained - it is surprisingly difficult to construct examples of it, due to the numerous requirements that must be satisfied. To see why this is the case, let us consider the requirements of a valid Winograd schema:

WSC 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 the 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 the 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 that it remains aligned with naturally occurring language.

Crafting valid examples of the Winograd schema is a complex and time-consuming task due to the set restrictions and requirements. In the original paper of WSC, only 100 expert-crafted schemas were provided. The set of schemas was later expanded to 273 examples (known as WSC273) and further to 285 examples (WSC285), though creating valid schemas remained challenging. Below are some examples from the original 100 schemas:

Schemas from the original WSC paper

1. The city councilmen refused the demonstrators a permit because they [feared/advocated] violence. Who [feared/advocated] violence?
   Answer: The city councilmen/the demonstrators.

2. The trophy doesn't fit into the brown suitcase because it's too [small/large]. What is too [small/large]?
   Answer: The suitcase/the trophy.

3. Joan made sure to thank Susan for all the help she had [given/received]. Who had [given/received] help?
   Answer: Susan/Joan.

4. Paul tried to call George on the phone, but he wasn't [successful/available]. Who was not [successful/available]?
   Answer: Paul/George.

5. The lawyer asked the witness a question, but he was reluctant to [answer/repeat] it. Who was reluctant to [answer/repeat] the question?
   Answer: The witness/the lawyer.

6. This book introduced Shakespeare to [Ovid/Goethe]; it was a fine selection of his writing. A fine selection of whose writing?

   Answer: Ovid/Shakespeare

7. Alice looked for her friend Jade in the crowd. Since she always [has good luck/wears a red turban], Alice spotted her quickly. Who always [has good luck/wears a red turban]?

   Answer: Alice/Jade

8. During a game of tag, Ethan [chased/ran from] Luke because he was "it". Who was "it"?

   Answer: Ethan/Luke

9. At the Loebner competition the judges couldn't figure out which respondents were the chatbots because they were so [advanced/stupid]. Who were so [advanced/stupid]?

   Answer: the chatbots/the judges.

10. 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]? **Answer:** the password "GrWQWu8JyC"/the password "willow-towered Canopy Huntertropic wrestles".

After the proposal of the WSC, attempts to solve it were made using various methods, including feature-based approaches, neural networks, and language models. Sharma et al. [48] first focused on two particular categories of commonsense reasoning: event-event causality and causal attributive. In the former, an event is the cause of another event, while in the latter, an attribute is the cause of an event. For example, in the sentence “Sid explained his theory to Mark but he could not convince him”, the pronoun “he” refers to “Sid” because it is more likely that *someone who explains something to a person* also *could not convince that person.* This is an example of event-event causality. In the sentence “Pete envies Martin because he is very successful", the pronoun "he" refers to "Martin" because it is more likely that *someone who is very successful* is *envied by others*. This is an example of causal attributive reasoning. The authors developed a semantic parser that could extract the two categories of reasoning from an input sentence and used a knowledge hunting method to identify commonsense relations in a large corpus. They achieved an accuracy of 69% on a subset of 71 schemas of WSC282 which belonged to the two said categories of reasoning.

Initial progress on the tackling the full WSC was made by Emami et al. in 2018 [49], who achieved a better-than-chance accuracy of 57.1% on WSC273 using a more robust knowledge hunting framework. The advent of transformer models has revolutionised the field of natural language processing and has become the state-of-the-art approach for various tasks, including WSC. Sakaguchi et al. [50] achieved a human-like accuracy of 90.1% using a fine-tuned RoBERTa model [51].

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 [52].

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 [53, 54], 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.

Chapter 3

Contextuality

In the following, I will first give an overview of the history of contextuality in quantum mechanics, and then discuss its manifestations in other domains. Two different frameworks to study contextuality will be presented: the sheaf-theoretic framework and the contextuality-by-default framework.

3.1 Contextuality in quantum mechanics

The study of contextuality originated in the early developments of quantum mechanics, where the probabilistic behaviour of quantum systems seemed to defy classical intuition. In the orthodox description of quantum mechanics, measurements on quantum systems are intrinsically probabilistic. This raised the question whether the probabilistic behaviour of quantum systems is due to our lack of knowledge of the system's hidden variables, or if it is an intrinsic property of the system itself. The debate was initiated by the Einstein-Podolsky-Rosen (EPR) paradox, which proposed a thought experiment involving a pair of entangled particles and demonstrated that measurements on one particle can be used to predict the outcome of measurements on the other particle deterministically [55]. The paradox comes from the Heisenberg uncertainty principle, which states that the position and momentum of a particle cannot be measured simultaneously with arbitrary precision. However, in the EPR thought experiment, once one of the particle's properties is measured, the other particle's property is determined. Hence, Heisenberg's uncertainty principle seems to be violated. EPR concluded that quantum mechanics was not complete and there should be some underlying hidden variables that determine the outcomes of measurements on entangled particles. In modern language, the classical explanations EPR sought were called *realistic hidden variable theories*. A theory is considered to be realistic if it assigns definite values to observable features (observables) of a system, regardless of acts of measurement. Classical physics had been built on this natural assumption of realism but quantum mechanics seemed to challenge it. EPR proposed such a realistic hidden variable theory for a pair of entangled photons but concluded that the hidden variables must be non-local, that is, the particles must communicate with each other instantaneously over arbitrarily large distances. This phenomenon was later famously coined "spooky action at a distance" to argue for the incompleteness of quantum mechanics.

Bohm proposed a simplified version of the EPR setup [56], making use of spins instead of continuous properties like position and momentum. Bell used Bohm's version of the EPR setup of two entangled and spatially separated particles to derive his famous theorem, which states that no local hidden variable theory can reproduce all the predictions of quantum mechanics [3]. Kochen and Specker later proved another no-go theorem for realistic hidden variable theories in a more general setting, where they carefully designed an example which requires the hidden variable to be dependent on the choice of the set of compatible measurements [4]. Such set of compatible measurements are also known as a *context*. Contextuality refers to the need for the hidden variable to depend on the context in which the system is measured. Quantum physics is considered to be contextual precisely because any realistic hidden variable theory that describes quantum systems must be contextual. Contrary to a common misconception, the orthodox quantum mechanics [20] is not contextual, as the probability of measuring a particular outcome is independent of the context. Quantum mechanics is however non-realistic, as outcomes are not deterministic but probabilistic according to the Born rule [57]. Therefore, a theory of quantum physics can be either non-contextual or realistic, but not both at the same time. There have been many attempts to construct realistic theories of quantum physics, with the most famous one being the pilot-wave theory proposed by Louis de

Broglie [58], which was later mathematically formalised by David Bohm [59].

3.1.1 Bell's inequalities

Bell's inequalities are a type of inequalities in terms of measurement probabilities that are satisfied by local realistic hidden variable theories but violated by quantum mechanics. In this subsection, we will review CHSH inequality, which was derived by Clauser, Horne, Shimony and Holt [5], which has been experimentally tested by several groups [60, 61, 62, 63, 64].

The setting of the Bell-CHSH experiment is as follows: Two spatially separated observers, Alice and Bob, each perform measurement on a shared pair of entangled particles, which are commonly implemented using photons. The state of the pair of particles is maximally entangled. An example of such a state is given by the following Bell state:

$$\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$$

Alice and Bob can each choose to measure their part of the state with one of two incompatible observables, a_1 and a_2 for Alice and b_1 and b_2 for Bob. To be precise, a_1 and a_2 are incompatible and so are b_1 and b_2, but a_1 and b_1 are compatible and so are a_2 and b_2. In the traditional settings, the measurement outcomes are either -1 or $+1$ for each observable. Therefore, there are 4 possible measurement contexts: $\{a_1, b_1\}$, $\{a_1, b_2\}$, $\{a_2, b_1\}$, and $\{a_2, b_2\}$. The Bell-CHSH inequalities are satisfied by all local and real systems. There are 8 Bell-CHSH inequalities in total:

$$\begin{aligned}
-2 \le S_{+++-} &:= +\langle a_1 b_1\rangle + \langle a_1 b_2\rangle + \langle a_2 b_1\rangle - \langle a_2 b_2\rangle \le 2 \\
-2 \le S_{++-+} &:= +\langle a_1 b_1\rangle + \langle a_1 b_2\rangle - \langle a_2 b_1\rangle + \langle a_2 b_2\rangle \le 2 \\
-2 \le S_{+-++} &:= +\langle a_1 b_1\rangle - \langle a_1 b_2\rangle + \langle a_2 b_1\rangle + \langle a_2 b_2\rangle \le 2 \\
-2 \le S_{-+++} &:= -\langle a_1 b_1\rangle + \langle a_1 b_2\rangle + \langle a_2 b_1\rangle + \langle a_2 b_2\rangle \le 2
\end{aligned}$$

The Bell-CHSH inequalities form a complete characterization of all local observations, i.e. a theory satisfies the Bell-CHSH inequalities if and only if the theory is local.

3.1.2 Leggett-Garg inequalities

Leggett-Garg inequalities [65] are a set of inequalities designed to test the principles of macroscopic realism and non-invasive measurability in quantum mechanics. These inequalities serve as temporal analogues to Bell's inequalities, describing a single system measured at different times rather than spatially separated systems.

These inequalities reveal a non-classical aspect of quantum systems known as *macroscopic realism*, which posits that a macroscopic system with two or more distinct states is always in one of those states at any given time.

Expressed in terms of correlation functions of measurements performed at different times, the Leggett-Garg inequalities consider a system with two possible states. Let $Q(t_i)$ represent the measurement outcome at time t_i, where $Q(t_i) \in \{-1, 1\}$. The correlation function between measurements at times t_i and t_j is given by $\langle Q(t_i)Q(t_j)\rangle$.

A typical form of the Leggett-Garg inequality is:

$$+\langle Q(t_1)Q(t_2)\rangle + \langle Q(t_2)Q(t_3)\rangle - \langle Q(t_1)Q(t_3)\rangle \leq 1.$$

Similar to the Bell-CHSH inequalities, by flipping the signs of the observables, we can derive a total of 4 Leggett-Garg inequalities:

$$\begin{aligned}
+\langle Q(t_1)Q(t_2)\rangle + \langle Q(t_2)Q(t_3)\rangle - \langle Q(t_1)Q(t_3)\rangle &\leq 1,\\
+\langle Q(t_1)Q(t_2)\rangle - \langle Q(t_2)Q(t_3)\rangle + \langle Q(t_1)Q(t_3)\rangle &\leq 1,\\
-\langle Q(t_1)Q(t_2)\rangle + \langle Q(t_2)Q(t_3)\rangle + \langle Q(t_1)Q(t_3)\rangle &\leq 1,\\
-\langle Q(t_1)Q(t_2)\rangle - \langle Q(t_2)Q(t_3)\rangle - \langle Q(t_1)Q(t_3)\rangle &\leq 1.
\end{aligned}$$

In quantum mechanics, violations of the Leggett-Garg inequalities have been observed, indicating that either macroscopic realism or non-invasive measurability, or both, do not hold. This suggests that quantum systems cannot be fully described by classical intuition, even at macroscopic scales.

In recent years, several experimental studies have provided compelling evidence for these violations. For example, experiments using superconducting qubits have observed clear deviations from classical predictions [66], highlighting the intrinsic quantum coherence in these systems. Similarly, tests performed with nuclear magnetic resonance (NMR) setups have shown that even nuclear spin systems exhibit behaviour inconsistent with macroscopic realism [67]. Together, these results demonstrated the fundamental role of quantum mechanics in describing the behaviour of systems across different scales.

3.1.3 KCBS inequality

The KCBS inequality, named after Klyachko, Can, Binicioğlu, and Shumovsky [68], is a simple way to demonstrate the contextuality of qutrit systems, such as a spin-1 system, without the need for entanglement among space-like separated systems.

The 5-cyclic scenario is defined as follows:

- Observables: $\mathcal{X} = \{A_1, A_2, A_3, A_4, A_5\}$.
- Contexts: $\mathcal{M} = \{\{A_1, A_2\}, \{A_2, A_3\}, \{A_3, A_4\}, \{A_4, A_5\}, \{A_5, A_1\}\}$.
- Outcomes: $\mathcal{O} = \{-1, +1\}$.

The KCBS inequality is expressed in terms of the expectation values of the observables in each context:

$$+\langle A_1 A_2 \rangle + \langle A_2 A_3 \rangle + \langle A_3 A_4 \rangle + \langle A_4 A_5 \rangle + \langle A_5 A_1 \rangle \geq -3,$$

Here, $\langle A_i A_{i\oplus 1} \rangle$ denotes the expectation value of the product of the outcomes of the observables A_i and $A_{i\oplus 1}$ in the context $\{A_i, A_{i\oplus 1}\}$. The above inequality can be violated by a qutrit system with five diachromatic observables $A_i = 2P_i - I$, where P_i is a projector. The projects are arranged in a pentagonal configuration, where each pair of neighbouring projectors are chosen to be orthogonal to ensure compatibility. The state of the system is chosen to the eigenstate of $\sum_{i=1}^{5} A_i A_{i\oplus 1}$ with the lowest eigenvalue, which turns out to be $5 - 4\sqrt{5} < -3$, violating the KCBS inequality.

Similar to the Bell-CHSH inequalities, more KCBS inequalities can be derived by considering flipping the signs of observables. Flipping the sign of one observable results in the sign switching of two terms in the sum. Hence the number of negative signs in the sum must be even. The following are the full set of signs pattern of the KCBS inequalities:

$$
\begin{array}{lllll}
--+++ & -+-++ & -++-+ & -+++- & \\
+--++ & +-+-+ & +-++- & & \\
++--+ & ++-+- & & & \\
+++-- & & & & \\
----+ & ---+- & --+-- & -+--- & +----
\end{array}
$$

For example, the signs pattern $-+-++$ corresponds to the following KCBS inequality:

$$-\langle A_1A_2\rangle + \langle A_2A_3\rangle - \langle A_3A_4\rangle + \langle A_4A_5\rangle + \langle A_5A_1\rangle \geq -3.$$

Lapkiewicz et al. [69] experimentally demonstrated the violation of the KCBS inequality using a single photonic qutrit system. The KCBS inequality was also tested and used a certification of quantum randomness in an ion-trap system [70].

3.2 Contextuality in other domains

We have seen that contextuality is not only limited to quantum mechanics, but also exhibits in other domains such as psychology experiments. Indeed, the Bell-CHSH inequality is not only violated by quantum mechanics, but also by other systems with the PR box as a well-known example of a non-quantum system that violates Bell's inequalities.

3.2.1 Quantum cognition and psychology

Quantum cognition is a field that emerged in the early 2000s, using quantum mechanics' mathematical principles — like superposition, entanglement, and interference — to model human cognition, particularly in decision-making, perception, and memory. Traditional cognitive models often rely on classical probability but struggle with phenomena such as the conjunction fallacy and order effects, which defy classical expectations. Busemeyer and Pothos proposed quantum probability models to address these anomalies [71]. Earlier attempts to apply quantum contextuality to cognition, such as the work by Aerts et al. [72], claimed violations of the CHSH inequality in concept combinations. However, these results were later shown to be flawed due to the presence of signalling, which was not properly accounted for.

Quantum models have proven effective across cognitive domains. In decision-making, they account for preference reversals and inconsistent patterns [73]. In memory and language, quantum interference and contextuality have modelled associative recall and meaning ambiguities [74]. Contemporary studies also explore potential links between quantum theory and neural processes, with future research aiming at quantum cognitive neuroscience and broader applications in behavioural science [75].

An important advancement in applying contextuality to psychology was the development of the Contextuality-by-Default (CbD) theory by Dzhafarov and Kujala [76, 77]. This framework provided a systematic way to analyse experiments in which signalling is prevalent. In the CbD framework, signalling (also referred to as *direct influences*) is explicitly treated, while contextuality is defined as the remaining context-dependence that cannot be explained by these direct influences alone.

3.2.2 Linguistics

Wang et al. [78, 79] pioneered the application of contextuality to the study of lexical ambiguity in natural language. By examining combinations of ambiguous subject-verb or verb-object phrases, they constructed the Bell-CHSH scenario within the domain of linguistics.

For example, the verb-object system $\{\text{tap, box}\} \times \{\text{pitcher, cabinet}\}$ induces four distinct contexts: $\{\text{tap, pitcher}\}$, $\{\text{tap, cabinet}\}$, $\{\text{box, pitcher}\}$, and $\{\text{box, cabinet}\}$. Each of these words can be interpreted in two different ways:

Encoding	tap	box	pitcher	cabinet
0	touch	put in boxes	jug	government
1	record	fight	baseball player	furniture

One can view this system as a Bell-CHSH scenario, where the two verbs are seen as Alice's measurements and the two nouns as Bob's measurements. The outcomes are encoded as 0 and 1, representing the two possible meanings of the words shown in the table above. The authors demonstrated that such a system can exhibit contextuality. They also considered different contexts formed by the ordering of the words in the phrases, such as "throw pitcher" and "pitcher throws", and showed that CbD contexutlity can be observed in these systems as well.

Chapter 4

Contextuality frameworks

There is a need of frameworks to study contextuality systematically that goes beyond describing contextuality using the language of quantum mechanics. The objects of study in these frameworks are only the empirically observed statistics of compatible measurements and nothing else. It allows for a unified approach to study contextuality in a way that is independent of the physical system or the context in which the measurements are performed. One such framework is the sheaf-theoretic framework proposed by Abramsky and Brandenburger [80], which makes use of the mathematical language of sheaves to formalise the global compatibility of locally compatible observations. The other one is the contextuality-by-default framework proposed by Dzhafarov and Kujala [77], which takes on an even more general approach by considering random variables.

4.1 Sheaf-theoretic framework

Abramsky and Brandenburger introduced the sheaf-theoretic framework to unify the study of non-locality and contextuality within a single mathematical formalism [80]. Part of their motivation was the realization that both phenomena can be understood as obstructions to forming a single *global* probability distribution consistent across the *local* probability distributions on sets of compatible measurements, which are called contexts. Such idea could be traced back to the work of Fine [81] where he showed that the following statements are equivalent:

1. A local hidden variable theory exists.

2. There exists a global joint probability distribution that marginalises to the local distributions.

The mathematical structure of sheaves naturally captures the notion of globally compatible local data and the authors in [80] made use of sheaf theory to formalise contextuality and non-locality in a unified way.

In the following, I will give a brief overview of sheaf theory and how it is used to study contextuality.

4.1.1 Sheaf theory

Sheaf theory is a branch of mathematics that provides a formal framework for studying local-to-global data living on some kind of space. The core idea of sheaf theory is best understood through an example. A well-known example of the phenomenon that local consistency does not imply global consistency is Penrose's staircase, which is a hypothetical staircase that appears to have steps that are all going down (when viewed in the clockwise direction) but still manages to return to the same height after a full turn.

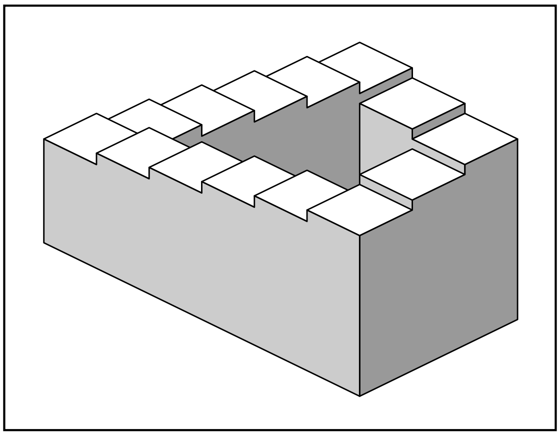
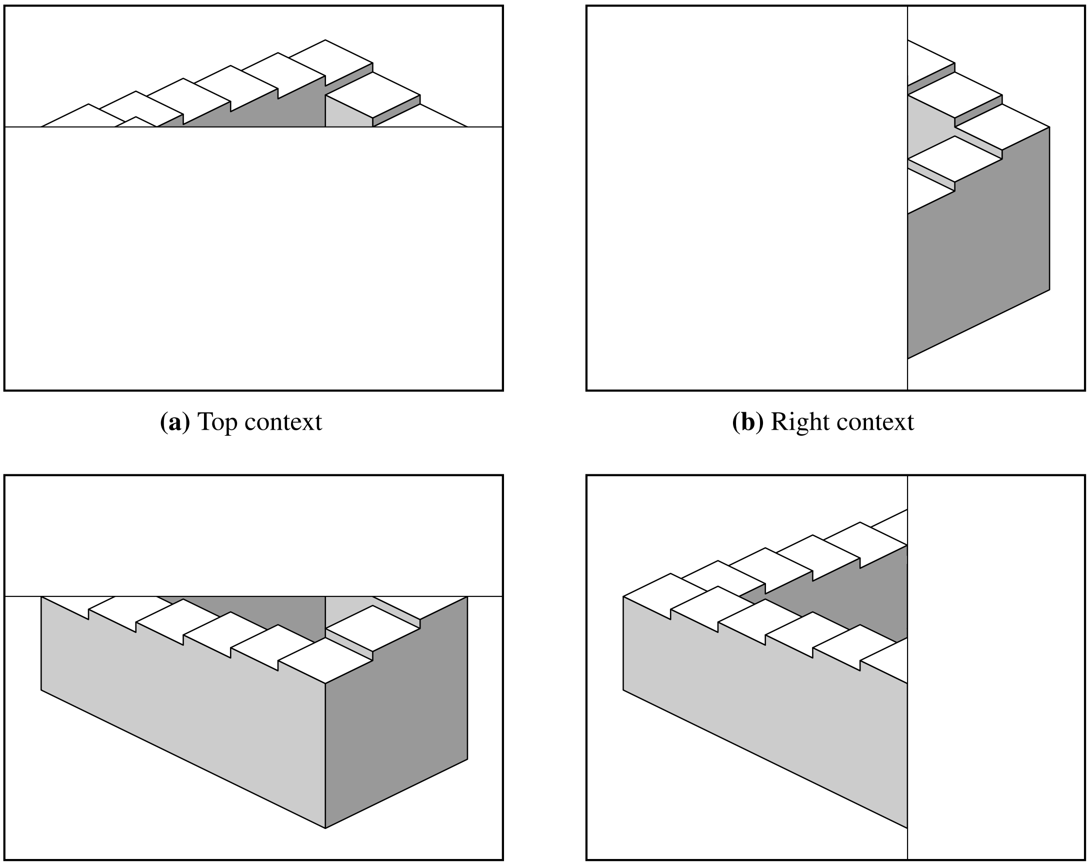


Figure 4.1: A global view of the Penrose staircase.

The staircase is said to be locally consistent in the sense that a local view of the staircase at any point is consistent with a physically possible staircase. Figure 4.2 shows four different local views of the staircase. Although each local view appears consistent with a physically possible staircase, the global view constructed by gluing them together creates an impossible staircase.



(a) Top context **(b)** Right context

(c) Bottom context **(d)** Left context

Figure 4.2: Different local views of the Penrose staircase. Although each local view is with a physically possible staircase, the global view pieced together from these local views is not.

Sheaf theory provides a mathematical framework to study such phenomena. Here I will briefly introduce the basic concepts of sheaf theory while referring to the example of the Penrose staircase. First, we need to define what a presheaf is.

Definition 1 (Presheaf)**.** *Given a topological space X, a presheaf $\mathcal{F}$ on X is defined such that*

1. *For each open subset $U \subseteq X$, there is a set $\mathcal{F}(U)$, whose elements are called* sections over U.
2. *For each pair of open sets U and V such that $V \subseteq U$, there is a restriction map $res_{U,V} : \mathcal{F}(U) \to \mathcal{F}(V)$, which satisfies the following properties:*

(a) For every open set U, $res_{U,U}$ is the identity map on $\mathcal{F}(U)$.

(b) For every triple of open sets $W \subseteq V \subseteq U$, $res_{V,W} \circ res_{U,V} = res_{U,W}$.

Using terminology from category theory, the presheaf $\mathcal{F} : \mathcal{P}(X)^{op} \to Set$ is a contravariant functor from the category of open sets of X to the category of sets, where $\mathcal{P}(X)$ is the poset of open sets in X and Set is the category of sets.

A presheaf starts with a topological space X, which in the case of the Penrose staircase is a rectangular canvas on which the staircase is drawn. The topological space X is by definition a collection of open sets, which can be structured into a partially ordered set by set inclusion. The four visible regions of the Penrose staircase in Figure 4.2 are examples of open sets in the space X. For each open set U in X, we assign a set $\mathcal{F}(U)$ which is the set of all possible data that live on the open set U, and elements in $\mathcal{F}(U)$ are called *sections over* U. In the case of the Penrose staircase, the set $\mathcal{F}(U)$ is the set of all possible ways to draw a physically possible staircase consistent within the local region U.

Now we have defined how data can be attached to open sets in the space X. The next step is to define how a section on an open set U can be restricted to a smaller open set $V \subseteq U$. This is done by specifying a restriction map $\text{res}_{U,V} : \mathcal{F}(U) \to \mathcal{F}(V)$ for each pair of open sets U and V such that $V \subseteq U$. The restriction map $\text{res}_{U,V}$ is a map that takes a section $s \in \mathcal{F}(U)$ and returns a restricted section in $\mathcal{F}(V)$. The restriction maps should also compose naturally, i.e., given three open sets $W \subseteq V \subseteq U$, the restriction map $\text{res}_{V,W} \circ \text{res}_{U,V} = \text{res}_{U,W}$ should hold. In the case of the Penrose staircase, the restriction map $\text{res}_{U,V}$ takes a drawing on the open set U and just erases the parts that are not in V. In summary, a presheaf defines what data can be attached to open sets via $\mathcal{F}$ and how the data can be restricted to smaller open sets via res.

A sheaf is a presheaf that allows for gluing of local data to form a global consistent data, which is formally defined as follows.

Definition 2 (Sheaf)**.** *A sheaf is a presheaf that satisfies the gluing property: for every open cover $\{U_i\}_i$ of an open set U: if $s_i \in \mathcal{F}(U_i)$ are sections such* *that* $res_{U_i,U_i\cap U_j}(s_i) = res_{U_j,U_i\cap U_j}(s_j)$ *for all* i, j*, then there exists a* unique *section* $s \in \mathcal{F}(U)$ *such that* $res_{U,U_i}(s) = s_i$ *for all* i*.*

Here an open cover $\{U_i\}_i$ of an open set U is a collection of open sets such that their union $\bigcup_i U_i$ is equal to U. The gluing property requires that if we have local sections defined on each open set in a cover that agree with one another on all pairwise intersections, then there exists a unique global section over the entire space that restricts to each of these local sections. Recall that the presheaf $\mathcal{F}$ for the Penrose staircase allows only drawings of physically possible staircases as sections. Figure 4.2 shows the local sections on the four open sets that form an open cover of the entire space X. If we glue the local sections together to form a global drawing like Figure 4.1, we will end up with an impossible staircase which is not a physically possible drawing. Therefore, the presheaf $\mathcal{F}$ for the Penrose staircase is not a sheaf.

4.1.2 Framework for contextuality based on sheaf theory

In the sheaf-theoretic framework of contextuality of Abramsky and Brandenburger [80], the presheaf of interest is a presheaf of joint probability distributions over measurement outcomes on compatible sets of observables in a measurement scenario.

Definition 3 (Measurement scenario)**.** *A measurement scenario is a tuple* $\langle \mathcal{X}, \mathcal{M}, \mathcal{O} \rangle$ *where*

1. *(observables)* $\mathcal{X}$ *is a set of observables.*
2. *(contexts)* $\mathcal{M}$ *is a collection of subsets of* $\mathcal{X}$ *(called contexts), where each context represents a set of compatible measurements.* $\mathcal{M}$ *forms a simplicial complex on* $\mathcal{X}$*, that is, for any context* $C \in \mathcal{M}$*, any subset* $C' \subseteq C$ *is also in* $\mathcal{M}$*.*
3. *(outcomes)* $\mathcal{O}$ *is a set of possible outcomes for each observable in* $\mathcal{X}$*.*

An observable $X \in \mathcal{X}$ is a quantity that can be measured to produce an outcome $O \in \mathcal{O}$. For example, a Hermitian operator in quantum mechanics is an observable.

A context is a set of compatible observables that can be measured simultaneously. In quantum mechanics, two observables are compatible if they commute, i.e., $[X,Y]=0$.

It is important to note quantum mechanics is not the only system that can be modelled as a measurement scenario. Any theory or system that can be measured and there are incompatible measurements can be modelled as a measurement scenario.

Definition 4 (Maximal context)**.** *A maximal context is a context $C \in \mathcal{M}$ such that there is no other context $C' \in \mathcal{M}$ such that $C \subset C'$.*

Definition 5 (Empirical model)**.** *An empirical model is a collection $\{P_C\}_{C\in\mathcal{M}}$ where each P_C is a joint probability distribution on $\mathcal{O}^{|C|}$, assigning probabilities to each possible combination of measurement outcomes for the observables in maximal context $C \in \mathcal{M}$.*

Concretely, given a maximal context $C = \{X_1,\ldots,X_n\}$, the joint probability distribution P_C is a function $P_C : \mathcal{O}^n \to [0,1]$ such that $\sum_{o_1,\ldots,o_n} P_C(o_1,\ldots,o_n) = 1$.

It is important to note that every joint probability distribution P_C where $C \in \mathcal{M}$ can be estimated from repeated measurements on the observables in C, or be calculated exactly using an underlying theory of the concerned system, e.g. using Born's rule in quantum mechanics for a quantum system. Whereas the global joint probability distribution P over all observables in $\mathcal{X}$, if it exists at all, cannot be estimated from repeated measurements unless in the trivial case where all observables in $\mathcal{X}$ are compatible with each other, i.e. $\mathcal{M} = \{\mathcal{X}\}$.

We can now define the presheaf of joint probability distributions on a measurement scenario.

Definition 6 (Distribution presheaf)**.** *Given a measurement scenario $\langle \mathcal{X}, \mathcal{M}, \mathcal{O} \rangle$, the corresponding distribution presheaf $\mathcal{F}$ is defined as follows:*

1. *For each context $C \in \mathcal{M}$, $\mathcal{F}(C)$ is the set of all joint probability distributions on $\mathcal{O}^{|C|}$.*

2. *For each pair of contexts $C, C' \in \mathcal{M}$ such that $C' \subseteq C$, the restriction map $res_{C,C'} : \mathcal{F}(C) \to \mathcal{F}(C')$ is defined by marginalizing the joint distribution on C to the smaller context C'.*

The marginalization of a joint distribution P_C on C to a smaller context C' is defined as usual:

$$res_{C,C'}(P_C)(\{x'\}_{X \in C'}) = \sum_{x \in C \setminus C'} P_C(\{x\}_{X \in C})$$

Definition 7 (Non-signalling model)**.** *A non-signalling model is an empirical model $\{P_C\}_{C \in \mathcal{M}}$ such that for every pair of contexts $C_1, C_2 \in \mathcal{M}$, the marginalization of P_{C_1} to $C_1 \cap C_2$ is equal to the marginalization of P_{C_2} to $C_1 \cap C_2$, that is*

$$res_{C_1, C_1 \cap C_2}(P_{C_1}) = res_{C_2, C_1 \cap C_2}(P_{C_2})$$

Classical physics has been built on the assumption that all measurements are revealing deterministic pre-existing values of the observables, in the sense that any probabilistic behaviour of the measurement outcomes is due to our ignorance of the underlying state of the system. Fine's theorem [81] states that if a local hidden variable theory exists for a particular system, then there is a global joint distribution over all the observables in the measurement scenario that marginalises to every local joint distribution in the empirical model of the system.

An empirical model is said to be contextual if it cannot be explained by a local hidden variable theory, which means there is not a global section over the entire space $\mathcal{X}$ that marginalises to every local section in the empirical model.

Definition 8 (Non-contextual model)**.** *An empirical model $\{P_C\}_{C \in \mathcal{M}}$ is said to be* non-contextual *if there exists a global joint distribution P_X on $\mathcal{X}$ that marginalises to every P_C for all contexts $C \in \mathcal{M}$. Conversely, an empirical* *model that is not non-contextual is said to be* contextual.

Note that the global distribution P_X in Definition 8, if it exists, is actually a global section when given local sections P_C for all contexts $C \in \mathcal{M}$. Therefore, the existence of a contextual model proves that the distribution presheaf $\mathcal{F}$ on the measurement scenario is not a sheaf.

Definition 9 (Non-contextual measurement scenario)**.** *A measurement scenario $\langle \mathcal{X}, \mathcal{M}, \mathcal{O} \rangle$ is said to be non-contextual if all the empirical models on the scenario are non-contextual. That is, the distribution presheaf $\mathcal{F}$ on the measurement scenario is a sheaf.*

We say that a measurement scenario supports contextuality *if it is not non-contextual.*

As an example, the Bell-CHSH scenario involves two experimenters, Alice and Bob, who share between them a two-qubit quantum state. Alice is allowed to measure her part of the state with one of two incompatible observables, a_1 and a_2, which gives either 0 or 1 as the outcome. Similarly, Bob can choose to measure his part with observables b_1 and b_2. Therefore, the Bell-CHSH measurement scenario is fully described with the following data: $\mathcal{X} = \{a_1, b_1, a_2, b_2\}$, $\mathcal{M} = \{\{a_1, b_1\}, \{a_1, b_2\}, \{a_2, b_1\}, \{a_2, b_2\}\}$, and $\mathcal{O} = \{0, 1\}$. Notice that $\{a_1, a_2\}$ and $\{b_1, b_2\}$ are not in $\mathcal{M}$ as they cannot be measured simultaneously due to their quantum mechanical incompatibility.

So far we have specified what measurements are allowed and what outcomes are possible. Suppose now Alice and Bob repeat the experiment many times and have gathered sufficient data to estimate the joint probability distributions for each context in $\mathcal{M}$ to a sufficient degree of accuracy. Their results can be summarised in a table referred to as an *empirical table*, see Figure 4.3, where each row in the table represent a joint distribution on the context shown in the leftmost column. For instance, the bottom right entry in the table ($1/8$) is the probability of both Alice and Bob getting 1 as their measurement outcomes when Alice chooses to measure a_2 and Bob chooses to measure b_2. Note that the empirical model of the system is

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
(a_1,b_1)	$1/2$	0	0	$1/2$
(a_1,b_2)	$3/8$	$1/2$	$1/2$	$3/8$
(a_2,b_1)	$3/8$	$1/8$	$1/8$	$3/8$
(a_2,b_2)	$1/8$	$3/8$	$3/8$	$1/8$

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
(a_1,b_1)	1	0	0	1
(a_1,b_2)	1	1	1	1
(a_2,b_1)	1	1	1	1
(a_2,b_2)	1	1	1	1

Figure 4.3: Empirical tables of measurement scenarios: Bell-CHSH (left), possibilistic Bell-CHSH (right)

entirely described by the empirical table.

One can show that, using elementary linear algebra, there exists no global distribution over $\{a_1,a_2,b_1,b_2\}$ that marginalises to the 4 local distribution shown in the above empirical table [80]. Therefore, the empirical model considered here is indeed contextual.

4.1.3 Possibilistic models

Instead of probability distributions, one can also consider possibility, i.e. whether an outcome is possible or not. If we use Boolean values to represent possibility, 0 for *impossible* and 1 for *possible*, the passage from probability to possibility is just a mapping of all zero probabilities to 0 and all non-zero probabilities to 1. The following is a summary of the mapping from probability to possibility:

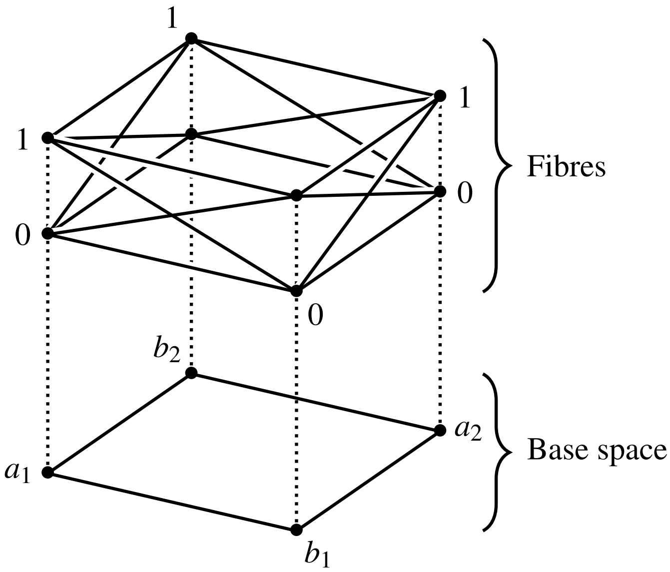
	Probability	Possibility
values	$[0,1]$	$\{0,1\}$
addition	$+$	$\vee$
multiplication	$\cdot$	$\wedge$

This irreversible mapping is called a *possibilistic collapse* of the empirical model. Similarly, we can define the possibilistic version of the distribution presheaf $\mathcal{F}$, which is a presheaf of joint possibility distributions over measurement outcomes on compatible sets of observables in a measurement scenario.

Definition 10 (Possibility distribution presheaf)**.** *Defined analogously to the distribution presheaf (Definition 6), except that the probability distributions are replaced with possibility distributions. Addition is replaced with logical* *disjunction $\vee$ and multiplication is replaced with logical conjunction $\wedge$.*

For the empirical table of the possibilistic version of Bell-CHSH see Figure 4.3.

Given a measurement scenario $\langle \mathcal{X}, \mathcal{M}, \mathcal{O} \rangle$, one can visualise its possibility distribution presheaf in the form of a *bundle diagram*, which consists of a base space and a fibre over each point in the base space. The base space is the measurement cover $\mathcal{M}$, which is a simplicial complex on the set of observables $\mathcal{X}$, and the fibre over each context $C \in \mathcal{M}$ is the all possible joint measurement outcomes on the observables in C.



The bundle diagram of a possibilistic empirical model is drawn similarly to the one of a distribution presheaf, except that the impossible sections are not shown in the diagram. Figure 4.4 shows the bundle diagrams of three different empirical models.

Definition 11 (Logical contextuality)**.** *An empirical model is said to be* logically contextual *if there is no global possibilistic distribution that marginalises to the local distributions of the possibilistic collapse of the model.*

There is a topological interpretation of logical contextuality in terms of the bundle diagram of an empirical model – an empirical model is not logically contextual if and only if each local outcome is a restriction of a global joint outcome. This is because a global possibilistic distribution can be thought of as simply a collection of

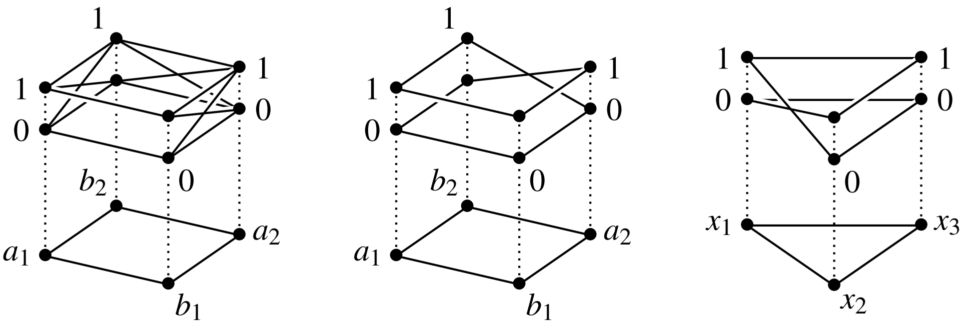
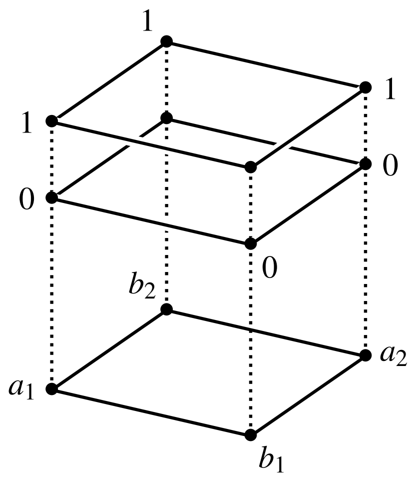


Figure 4.4: Bundle diagrams of possibilistic empirical models: (left) the Bell-CHSH test; (middle) a PR box; (right) a PR prism.

global joint outcomes. Hence, the fibre part of the bundle diagram of such global possibilistic distribution is simply a collection of global joint outcomes.

Using the Bell-CHSH scenario as an example, consider a global possibilistic distribution $P(a_1, b_1, a_2, b_2)$ which assigns the value 1 to only two joint outcomes, $(0,0,0,0)$ and $(1,1,1,1)$, which manifest on the bundle diagram (shown on the left below) as the two cycles that wrap around the base space.



The bundle diagram on the right shows an example of a logically contextual model, because the edge coloured in red is not extendable to a cycle that wraps around the base once.

Definition 12 (Strong contextuality)**.** *An empirical model is said to be* strongly contextual *if there is no global joint outcome that restricts to any of the local joint outcomes of the model.*

An example of a strongly contextual model on the Bell-CHSH scenario is the PR box, whose bundle diagram is shown in the middle of Figure 4.4. Note that every edge belongs to a cycle that wraps around the base *twice*.

4.1.4 Cyclic scenarios

In the following, we introduce a special family of measurement scenarios, the k-cyclic scenarios [82, 83], and argue that they are minimal in terms of supporting contextuality for any $k \geq 3$. In the formal language of sheaf, that is to say, the corresponding presheaf of a k-cyclic scenario is not a sheaf, provided $k \geq 3$. The k-cyclic scenarios encompass the Leggett-Garg experiment ($k = 3$) [65], the Bell-CHSH test ($k = 4$) [3, 5], and the KCBS inequality ($k = 5$) [68].

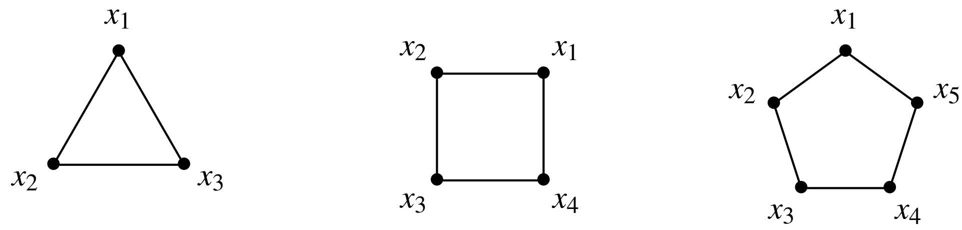
Definition 13 (k-cyclic scenario with binary outcomes)**.** *A k-cyclic scenario with binary outcomes is a measurement scenario $\langle \mathcal{X}, \mathcal{M}, \mathcal{O} \rangle$ such that*

1. $|\mathcal{X}| = |\mathcal{M}| = k$ *(there are k observables and k contexts),*
2. $\forall C \in \mathcal{M}, |C| = 2$ *(each context contains exactly two observables),*
3. $|\mathcal{O}| = 2$ *(there are two possible outcomes).*

The data of the measurement scenario can always be written in the following canonical form:

1. $\mathcal{X} = \{x_1,\, x_2,\, \ldots,\, x_k\}$,
2. $\mathcal{M} = \{\{x_1,\, x_2\},\, \{x_2,\, x_3\},\, \ldots,\, \{x_k,\, x_1\}\}$,
3. $\mathcal{O} = \{0,\, 1\}$.

For each context $\{x_i,\ x_{i\oplus 1}\} \in \mathcal{M}$*, where* $i \oplus 1 = i$ *if* $i < k$ *and* $i \oplus 1 = 1$ *if* $i = k$*, we denote the local distribution as* $P_{(i,\ i\oplus 1)}(x_i,\ x_{i\oplus 1})$*. We denote the local joint distribution as* $P_{(i,\ i\oplus 1)}(x_i,\ x_{i\oplus 1})$*.*



Leggett-Garg ($k = 3$) Bell-CHSH ($k = 4$) KCBS ($k = 5$)

Figure 4.5: The measurement covers $\mathcal{M}$ of the k-cyclic scenarios with $k = 3$, $k = 4$ and $k = 5$. The vertices represent the observables and the edges represent the contexts.

In the sheaf-theoretic framework, a 2-cyclic scenario does not admit contextual since it has only one context $\{x_1, x_2\}$. This is due to the fact that $\{x_1, x_2\}$ and $\{x_2, x_1\}$ are the same set. However, in the Contextuality-by-Default framework, which will be introduced in Section 4.2, a 2-cyclic scenario can admit contextual models as the framework allows for the same set of observables to be treated as different contexts and a different joint probability distribution can be assigned to each of them.

The following proposition, stated without proof in [80] and explicitly proven in [82], establishes that PR boxes are the only strongly contextual empirical models on a k-cyclic scenario.

Proposition 1. *The strongly contextual empirical models on a k-cyclic scenario, where* $k \geq 3$*, can be fully characterised by the following requirements:*

1. *For each* $i \in \{1, 2, \ldots, k\}$*, the local joint distribution* $P_{(i,\ i\oplus 1)}(x_i,\ x_{i\oplus 1})$ *is either*

   (a) *balanced perfect correlated, i.e.*

$$P_{(i,\ i\oplus 1)}(0,0) = P_{(i,\ i\oplus 1)}(1,1) = 1/2,$$

(b) or balanced perfect anti-correlated, i.e.

$$P_{(i,\ i\oplus 1)}(0,1) = P_{(i,\ i\oplus 1)}(1,0) = 1/2.$$

2. *The number of contexts with anti-correlated local joint distributions is odd.*

Such a model is called a PR box *on the k-cyclic scenario.*

In particular, we refer to the PR box on a 3-cyclic scenario as the *PR prism*, see Figure 4.4 for its bundle diagram. The parallel edges over the contexts $\{x_2, x_3\}$ and $\{x_3, x_1\}$ correspond to perfect correlation, while the crossed edges over context $\{x_1, x_2\}$ correspond to perfect anti-correlation.

4.1.5 The polytopes of empirical models

The space of probabilistic empirical models can be organised into polytopes, an idea originating from Pitowsky's pioneering work [84, 85, 86] on correlation polytopes. For a visualisation of the polytopes, see Figure 4.6. To formalise this, we define a convex combination of two empirical models e_1 and e_2 on the same measurement scenario $\langle \mathcal{X}, \mathcal{M}, \mathcal{O} \rangle$ as the following linear combination:

$$e = (1-\lambda)e_1 + \lambda e_2, \tag{4.1}$$

where $0 \leq \lambda \leq 1$. Scalar multiplication and addition on empirical models are defined as the same operations on the joint distribution of the empirical models. Let $e_1 = \{P_C\}_{C \in \mathcal{M}}$ and $e_2 = \{Q_C\}_{C \in \mathcal{M}}$ be two empirical models, where P_C and Q_C are the joint distributions of the empirical models. Then, the scalar multiplication and addition are defined as

$$e_1 + e_2 = \{P_C + Q_C\}_{C \in \mathcal{M}}, \tag{4.2}$$

$$\alpha e_1 = \{\alpha P_C\}_{C \in \mathcal{M}}. \tag{4.3}$$

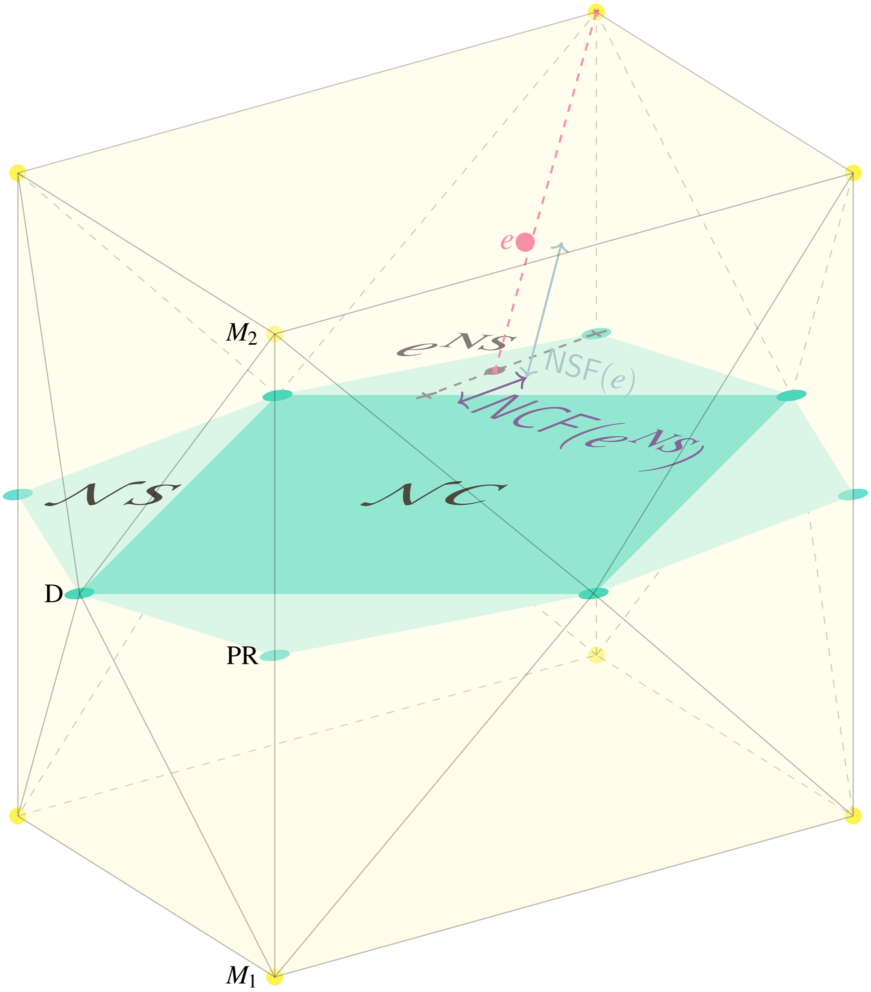


Figure 4.6: A 3-dimensional slice of the polytope of empirical models on the Bell-CHSH scenario. The polytope $\mathcal{NS}$ represents all non-signalling models, while $\mathcal{NC}$ represents all non-contextual models. The vertices of the $\mathcal{NC}$ polytope correspond to deterministic models that admit a global joint distribution, such as D. In contrast, the vertices of the $\mathcal{NS}$ polytope that are not part of $\mathcal{NC}$ are the PR boxes, which exhibit maximal contextuality. The vertices highlighted in yellow, such as M_1 and M_2, represent deterministic models that are signalling. The figure is adopted from [2].

The convex hull of all empirical models, including the ones that are signalling, is called the *polytope of empirical models*. The next polytope to introduce is $\mathcal{NS}$, which contains all non-signalling models, which are closed under convex combinations. To show that non-signalling models are indeed closed under convex combinations, we make use of the fact that the restriction map is a linear map, i.e.

$$\mathrm{res}_{C,C'}((1-\lambda)e_1+\lambda e_2) = (1-\lambda)\mathrm{res}_{C,C'}(e_1)+\lambda\mathrm{res}_{C,C'}(e_2), \tag{4.4}$$

where $\mathrm{res}_{C,C'}$ is the restriction map from the joint distribution of the empirical model to the joint distribution of the local model, and $C' \subseteq C$ is a subcontext of C. The proof of this property is straightforward and follows from the definition of marginalization of probability distributions.

Another important polytope is the $\mathcal{NC}$, which contains all non-contextual models. To show that non-contextual models are indeed closed under convex combinations, consider two non-contextual models e_1 and e_2 on the same measurement scenario $\langle \mathcal{X}, \mathcal{M}, \mathcal{O} \rangle$, which admit global distributions P_1 and P_2 respectively. The global distribution of the convex combination $e = (1-\lambda)e_1 + \lambda e_2$ is given by

$$P = (1-\lambda)P_1 + \lambda P_2. \tag{4.5}$$

To verify that P marginalises to the local distributions in e_1 and e_2, one can make use of the linearity property of the marginalization map.

The polytopes offer a geometric perspective on the relationships between different types of models, providing insights into their structure and boundaries. They also serve as a foundation for visualizing and quantifying key measures such as the contextual fraction (CF) and the signalling fraction (SF), which will be rigorously defined and explored in the subsequent sections.

4.1.6 Contextual fraction

The contextual fraction CF [87] measures the degree of contextuality of a given no-signalling model. Given an empirical model e, the CF of e is defined as the minimum λ such that the following convex decomposition of e works:

$$e = (1-\lambda)e^{NC} + \lambda e^{C}, \tag{4.6}$$

where e^{NC} is a non-contextual (and no-signalling) empirical model and e^{C} is a model allowed to be contextual. More formally,

$$\mathsf{CF} = \min_{\lambda} \left\{ \lambda \mid e = (1-\lambda)e^{NC} + \lambda e^{C} \text{ holds} \right\}. \tag{4.7}$$

The non-contextual fraction (NCF) is simply defined as $1-\mathsf{CF}$.

For no-signalling models, the criterion of contextuality is just

$$\mathsf{CF} > 0. \tag{4.8}$$

As e^{NC} is not allowed to be signalling, the CF of a signalling model must be greater than zero. Thus, interpreting CF as a measure of contextuality for signalling models would lead to erroneous conclusions. However, most models, including the ones considered in this paper, are signalling.

4.1.7 Signalling fraction

One can try to define a signalling fraction (SF), in the same way CF is defined, to quantity the degree of signalling. Given a model e, the SF of e is defined as the minimum μ such that the following convex decomposition of e works:

$$e = (1-\mu)e^{NS} + \mu e^{S}, \tag{4.9}$$

where e^{NS} is a no-signalling empirical model and e^{S} is a model allowed to be signalling. More concretely,

$$\mathsf{SF} = \min_{\mu} \left\{ \mu \mid e = (1-\mu)e^{NS} + \mu e^{S} \text{ holds} \right\}. \tag{4.10}$$

The non-signalling fraction (NSF) is simply defined as $1-\mathsf{SF}$.

4.1.8 Contextuality in the presence of signalling

In [2], the authors considered how much of the contextual fraction can be explained by the experimental imperfections that lead to signalling. The main idea is to quantify the change in the CF when the ideal empirical model is perturbed by noise that introduces signalling. To this end, the authors proved a continuity property of the CF with respect to the distance between empirical models:

Theorem 1 (Continuity of CF [2]). *Let e and e' be two empirical models on the same measurement scenario $\langle \mathcal{X}, \mathcal{M}, \mathcal{O} \rangle$. If the total variation distance between e and e' is $V(e, e') \leq \varepsilon$, then*

$$|\mathsf{CF}(e) - \mathsf{CF}(e')| \leq 2|\mathcal{M}|\,\varepsilon, \tag{4.11}$$

where $|\mathcal{M}|$ is the number of contexts in the measurement scenario. Here, the total variation distance between two empirical models $e = \{P_C\}_{C \in \mathcal{M}}$ and $e' = \{Q_C\}_{C \in \mathcal{M}}$ is defined as

$$V(e, e') = \max_{C \in \mathcal{M}} V(P_C, Q_C), \tag{4.12}$$

where $V(P_C, Q_C)$ is the total variation distance between two probability distributions.

The Theorem allows us to have a quantifiable flexibility in the choice of an ontological model to explain the given empirical model.

In the following, we extend CF to signalling models by allowing e^C in the convex decomposition (4.6) to be signalling.

Definition 14 (Contextual fraction for signalling models). *More formally, the* CF *of a signalling model e is defined as*

$$\mathsf{CF} = \min_{\lambda} \left\{ \lambda \mid e = (1-\lambda)e^{NC} + \lambda e^{CS} \text{ holds} \right\}, \tag{4.13}$$

with e^{CS} a model that is not both non-contextual and non-signalling.

Consider a given signalling empirical model e with $\mathsf{CF} > 0$. One can try to explain the empirical e using an ontological model h^S which is close to a non-signalling and non-contextual ontological model h^{NC}, such that

$$V(h^S, h^{NC}) \leq \mathsf{SF}. \tag{4.14}$$

Essentially this is to say that we allow the same amount of signalling in the ontological model as in the empirical model.

Using the continuity property of CF, we have

$$|\mathsf{CF}(h^S) - \mathsf{CF}(h^{NC})| \leq 2|\mathcal{M}|\,\mathsf{SF}. \tag{4.15}$$

Which implies

$$\mathsf{CF}(h^S) \leq 2|\mathcal{M}|\,\mathsf{SF}, \tag{4.16}$$

as $\mathsf{CF}(h^{NC}) = 0$. This means that even if we allow some amount of signalling in the ontological model, the (fictitious) CF in the ontological model cannot exceed $2|\mathcal{M}|\,\mathsf{SF}$. Now, if the CF of the empirical model exceeds $2|\mathcal{M}|\,\mathsf{SF}$, then no such ontological model h^S exists which explains the empirical model e. Hence the criterion for contextuality in the presence of signalling is given by

$$\mathsf{CF} > 2|\mathcal{M}|\,\mathsf{SF}, \tag{4.17}$$

where $|\mathcal{M}|$ denotes the number of measurement contexts. Again, to compute the CF for use in the above criterion, e^C is allowed to be signalling. Notice how criterion (4.17) reduces to the usual criterion for contextuality (4.8) when $\mathsf{SF} = 0$. Finally, it is worth mentioning that the criterion (4.17) is not a necessary condition for contextuality as Equation (4.16) only provides an upper bound on the CF of the ontological model. A more precise estimate of the CF of the ontological model would yield a tighter criterion for contextuality.

4.2 Contextuality-by-default framework

The Contextuality-by-Default (CbD) framework [76, 77] is a framework within which contextuality can be defined even in the presence of signalling. The primary objects of study in the CbD framework are random variables that represent the outcomes of measurements. The starting point of the CbD framework is to assign different random variables to the same measurable quantity when it is measured under different contexts, and that there is no joint distribution between random variables across different contexts. More formally, the outcome of measuring a content q under the context c is represented by a random variable R_q^c. In this approach, the same property measured under different contexts is represented by distinct random variables, making contextuality the default assumption unless proven otherwise.

In the sheaf-theoretic framework, a measurement scenario describes the set of all the possible joint measurements that can be performed. In the CbD framework, the analogous object is a *context-content matrix*, which is a matrix whose rows are indexed by the contexts and columns are indexed by the contents. The entries of the matrix are the random variables R_q^c and whenever a content is not measured in a context, the corresponding entry is left blank.

For example, the Bell-CHSH scenario has 4 contents $q_1 = a_1$, $q_2 = a_2$, $q_3 = b_1$, $q_4 = b_2$ and 4 contexts $c_1 = \{a_1, b_1\}$, $c_2 = \{a_1, b_2\}$, $c_3 = \{a_2, b_1\}$ and $c_4 = \{a_2, b_2\}$. The context-content matrix of the Bell-CHSH scenario is given by the following 4-by-4 matrix:

$$\begin{array}{|c|c|c|c||c|}
\hline
R_1^1 & R_2^1 & & & c^1 \\
\hline
 & R_2^2 & R_3^2 & & c^2 \\
\hline
 & & R_3^3 & R_4^3 & c^3 \\
\hline
R_1^4 & & & R_4^4 & c^4 \\
\hline
\hline
q_1 & q_2 & q_3 & q_4 & \mathcal{R}_4 \\
\hline
\end{array} \tag{4.18}$$

All random variables in the same context are considered *jointly distributed*, meaning that there is a well-defined joint distribution for them. While the random variables for the same content in different contexts are considered *stochastically unrelated*, meaning that there is no joint distribution for them. A system is *consistently connected* if for any content q, the set of random variables $\{R_q^c\}_{\{c|q\in c\}}$ follows the same distribution. For cyclic scenarios, consistent connectedness coincides with the no-signalling condition, which states that the marginal distribution of a *set of contents* does not depend on the context in which they are measured.

A crucial tool used in the CbD framework is the *coupling* of random variables.

Given a set of **stochastically unrelated** random variables $\{X_i\}_i$, a coupling is a set of jointly distributed random variables $\{S_i\}_i$ such that for each i, the random variable S_i has the same distribution as X_i.

In the following, we describe how contextuality is treated for cyclic systems in the CbD framework. Non-contextuality of a system is defined via the existence of a *maximally connected coupling*. We refer to a coupling for the set of all random variables in the context-content matrix as a *global coupling*, while a coupling for the random variables of a single content q in the context-content matrix is called a *local coupling* for content q. For a global coupling S and a content q, we define the quantity

$$\mathrm{eq}(S|_q) = \Pr[S_q^{c_1} = S_q^{c_2}] \tag{4.19}$$

which is the probability that the two random variables $S_q^{c_1}, S_q^{c_2}$ that are measuring the content q are equal according to the global coupling S. We can also define the quantity for a local coupling T_q for the content q as

$$\mathrm{eq}(T_q) = \Pr[T_q^{c_1} = T_q^{c_2}]. \tag{4.20}$$

A global coupling S is *maximally connected* if, for each content q, $\mathrm{eq}(S|_q)$ is maximal, i.e.

$$\mathrm{eq}(S|_q) = \max_{T_q} \mathrm{eq}(T_q). \tag{4.21}$$

Maximality here means that no local coupling T_q can do better than the global coupling S in terms of the probability of the random variables being equal.

Definition 15 (Non-contextuality in the CbD framework)**.** *A system is said to be non-contextual in the CbD framework if there exists a maximally connected coupling for the system.*

Hence, a system is contextual if for every global coupling S, there is at least one content q such that

$$\mathrm{eq}(S|_q) < \max_{T_q} \mathrm{eq}(T_q). \tag{4.22}$$

This hints at ways to develop quantitative measures of contextuality via the differences between the local and global couplings in terms of the probability of the random variables being equal. In particular, one can take an unweighted sum over all contents q of the differences between the local and global couplings, i.e.

$$\sum_q \max_{T_q} \mathrm{eq}(T_q) - \sum_q \mathrm{eq}(S|_q), \tag{4.23}$$

which measures the deviation of the coupling S from being maximally connected. We can then define a measure of contextuality as the maximum of the quantity in (4.23) over all global couplings S:

$$\mathsf{CNT} = \max_S \left(\sum_q \max_{T_q} \mathrm{eq}(T_q) - \sum_q \mathrm{eq}(S|_q) \right). \tag{4.24}$$

For more quantitative measures of contextuality in the CbD framework, we refer the reader to [88, 83].

4.2.1 Cyclic systems

A well-studied class of CbD systems are the cyclic systems [77], 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. For example, a cyclic system of rank 5 has the following context-content matrix:

$$\begin{array}{|c|c|c|c|c||c|}
\hline
R_1^1 & R_2^1 & & & & c^1 \\
\hline
 & R_2^2 & R_3^2 & & & c^2 \\
\hline
 & & R_3^3 & R_4^3 & & c^3 \\
\hline
 & & & R_4^4 & R_5^4 & c^4 \\
\hline
R_1^5 & & & & R_5^5 & c^5 \\
\hline
\hline
q_1 & q_2 & q_3 & q_4 & q_5 & \mathcal{R}_5 \\
\hline
\end{array} \tag{4.25}$$

It was conjectured in [89] and proven in [90] that for a cyclic system of rank n, the measure CNT can be expressed as the following form:

$$\mathsf{CNT} = s_{odd}\left(\left\{\langle R_j^j R_{j\oplus 1}^j\rangle\right\}_{j=1,\ldots,n}\right) - \Delta - n + 2 > 0 \tag{4.26}$$

where $j \oplus 1 = j$ if $j < n$ and $j \oplus 1 = 1$ if $j = n$. The function s_{odd} takes a set of n numbers $\{x_1, x_2, \ldots, x_n\}$ and returns the maximum of the odd parity sum of the numbers, i.e.

$$s_{odd}(x_1, x_2, \ldots, x_n) = \max(\pm x_1 \pm x_2 \pm \ldots \pm x_n) \tag{4.27}$$

such that the number of $-$ signs used is odd. The function s_{odd} can be computed efficiently as follows:

$$s_{odd}(x_1, x_2, \ldots, x_n) = \begin{cases} \sum_{i=1}^n |x_i| & \text{if number of negative } |\{k | x_k < 0\}| \text{ is odd} \\ \sum_{i=1}^n |x_i| - 2\min(|x_i|) & \text{otherwise} \end{cases} \tag{4.28}$$

The quantity Δ is called *direct influence* which measures the degree of signalling in the system. A no-signalling system has $\Delta = 0$.

A feature of the CNT measure is that it generalises the Bell-CHSH inequality. To see this, consider a cyclic system of rank 4, i.e. the Bell-CHSH scenario. The CNT quantity becomes:

$$\begin{aligned} \mathsf{CNT} = & s_{\text{odd}}\left(\langle R_1^1\, R_2^1\rangle, \langle R_2^2\, R_3^2\rangle, \langle R_3^3\, R_4^3\rangle, \langle R_4^4\, R_1^4\rangle\right) \\ & - \left|\langle R_2^1\rangle - \langle R_2^2\rangle\right| - \left|\langle R_3^2\rangle - \langle R_3^3\rangle\right| - \left|\langle R_4^3\rangle - \langle R_4^4\rangle\right| - \left|\langle R_1^4\rangle - \langle R_1^1\rangle\right| - 2. \end{aligned} \tag{4.29}$$

In the usual setting of the Bell-CHSH inequality, no signalling is allowed, i.e. $\Delta = 0$. In this case, the CNT measure becomes:

$$\mathsf{CNT} = s_{\text{odd}}\left(\langle R_1^1\, R_2^1\rangle, \langle R_2^2\, R_3^2\rangle, \langle R_3^3\, R_4^3\rangle, \langle R_4^4\, R_1^4\rangle\right) - 2 \tag{4.30}$$

Recall that the system is contextual if and only if $\mathsf{CNT} > 0$. Note that the s_{odd} function encompasses the full set of 8 CHSH inequalities. To see this, consider the all possible combinations of odd number of minus signs in the s_{odd} function:

$$
\begin{array}{cccc}
+++- & ++-+ & +-++ & -+++ \\
---+ & --+- & -+-- & +---
\end{array}
$$

where the first row corresponds to the following 4 inequalities:

$$
\begin{aligned}
+\langle R_1^1\, R_2^1\rangle + \langle R_2^2\, R_3^2\rangle + \langle R_3^3\, R_4^3\rangle - \langle R_4^4\, R_1^4\rangle &\leq 2, \\
+\langle R_1^1\, R_2^1\rangle + \langle R_2^2\, R_3^2\rangle - \langle R_3^3\, R_4^3\rangle + \langle R_4^4\, R_1^4\rangle &\leq 2, \\
+\langle R_1^1\, R_2^1\rangle - \langle R_2^2\, R_3^2\rangle + \langle R_3^3\, R_4^3\rangle + \langle R_4^4\, R_1^4\rangle &\leq 2, \\
-\langle R_1^1\, R_2^1\rangle + \langle R_2^2\, R_3^2\rangle + \langle R_3^3\, R_4^3\rangle + \langle R_4^4\, R_1^4\rangle &\leq 2.
\end{aligned}
$$

The second row corresponds to the following 4 inequalities:

$$
\begin{aligned}
-\langle R_1^1\, R_2^1\rangle - \langle R_2^2\, R_3^2\rangle - \langle R_3^3\, R_4^3\rangle + \langle R_4^4\, R_1^4\rangle &\leq 2, \\
-\langle R_1^1\, R_2^1\rangle - \langle R_2^2\, R_3^2\rangle + \langle R_3^3\, R_4^3\rangle - \langle R_4^4\, R_1^4\rangle &\leq 2, \\
-\langle R_1^1\, R_2^1\rangle + \langle R_2^2\, R_3^2\rangle - \langle R_3^3\, R_4^3\rangle - \langle R_4^4\, R_1^4\rangle &\leq 2, \\
+\langle R_1^1\, R_2^1\rangle - \langle R_2^2\, R_3^2\rangle - \langle R_3^3\, R_4^3\rangle - \langle R_4^4\, R_1^4\rangle &\leq 2,
\end{aligned}
$$

which can be rewritten to the usual form of the Bell-CHSH inequalities by multiplying both sides by -1.

Part II

Contextuality in coreference ambiguities

Chapter 5

Special classes of empirical models

In the rest of the thesis, we will be making use of two special classes of models, the **PR-like models** (Section 5.2) and the **outcome-symmetric models** (Section 5.3). In this Chapter, these classes of models will be introduced, and their properties will be investigated. The motivations to study these classes of models might seem unclear at this point. Readers are recommended to skip this chapter until its material is properly motivated in Chapter 6 and Chapter **??**.

5.1 PR-box

The PR-box [91] is a model on the Bell-CHSH measurement scenario with the following empirical table:

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
(a_1,b_1)	$1/2$	0	0	$1/2$
(a_1,b_2)	$1/2$	0	0	$1/2$
(a_2,b_1)	$1/2$	0	0	$1/2$
(a_2,b_2)	0	$1/2$	$1/2$	0

The PR-box is strongly contextual in the sheaf-theoretic framework with the maximum possible $\mathsf{CF} = 1$. The first three contexts are correlated, and the last one is anti-correlated. All other strongly contextual models on the Bell-CHSH measurement scenario can be produced by swapping the outcomes of the observables. For example, by swapping the outcomes of a_1, the empirical table becomes:

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
(a_1,b_1)	0	1/2	1/2	0
(a_1,b_2)	0	1/2	1/2	0
(a_2,b_1)	1/2	0	0	1/2
(a_2,b_2)	0	1/2	1/2	0

Note that the first two rows are now anti-correlated. In fact, swapping the outcomes of any observable would lead to two rows to switch between correlated and anti-correlated. Hence, the number of anti-correlated rows is always odd. In the case of the Bell-CHSH measurement scenario, the number of anti-correlated rows is either 1 or 3. Each of these cases produces 4 distinct models, resulting in a total of 8 distinct PR models. These 8 PR boxes span the local polytope.

In the rest of the thesis, the term PR-box will refer to any of the 8 PR models in the Bell-CHSH measurement scenario.

The notion of PR-box can be generalised to k-cyclic scenarios. A PR-box in a k-cyclic scenario is a model with an odd number of anti-correlated rows. One such example is as follows:

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
(x_1,x_2)	1/2	0	0	1/2
(x_2,x_3)	1/2	0	0	1/2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(x_{k-1},x_k)	1/2	0	0	1/2
(x_k,x_1)	0	1/2	1/2	0

The rest of the PR-boxes will be generated by swapping the outcomes of the observables. We can choose independently the parity (correlated or anti-correlated) for $k-1$. Hence, the total number of PR-boxes is 2^{k-1}.

5.2 PR-like models

We define a class of models called the *PR-like* models which are models that share the same support as a PR box. That is, the set of possible outcomes which have

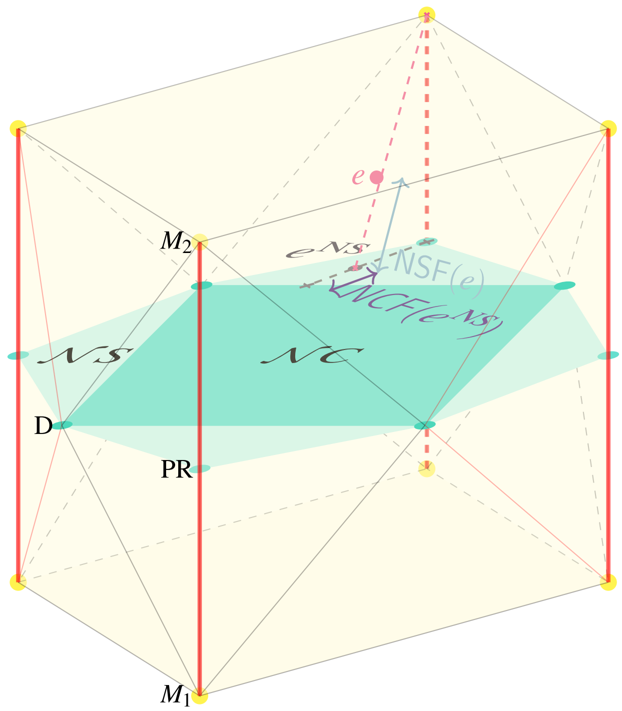


Figure 5.1: The polytope of empirical models with the PR-like models highlighted in red. The spaces of PR-like models are perpendicular to the $\mathcal{NS}$ space, and intersect with $\mathcal{NS}$ at the PR-boxes.

non-zero probabilities are the same as one of the PR boxes. The PR-like models are a relaxation of the PR boxes, where the probabilities of the outcomes are allowed to deviate from the PR box. The geometric visualisation of the PR-like models is given in Figure 5.1.

For example, on the 3-cyclic measurement scenario, a PR-like model has the following empirical table parameterised by $\varepsilon_i \in [-1,1]$:

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
(x_1,x_2)	$\frac{1+\varepsilon_1}{2}$	0	0	$\frac{1-\varepsilon_1}{2}$
(x_2,x_3)	$\frac{1+\varepsilon_2}{2}$	0	0	$\frac{1-\varepsilon_2}{2}$
(x_3,x_1)	0	$\frac{1+\varepsilon_3}{2}$	$\frac{1-\varepsilon_3}{2}$	0

The parameters ε_i can be thought of as the deviation of the probabilities from a PR box. Any PR box is also a PR-like model, with all parameters $\varepsilon_i = 0$. Below is the formal definition of a PR-like model.

Definition 16 (PR-like model)**.** *The PR-like model in the standard form on the k-cyclic measurement scenario is an empirical model with the following empirical table:*

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
(x_1,x_2)	$\frac{1+\varepsilon_1}{2}$	0	0	$\frac{1-\varepsilon_1}{2}$
(x_2,x_3)	$\frac{1+\varepsilon_2}{2}$	0	0	$\frac{1-\varepsilon_2}{2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(x_{k-1},x_k)	$\frac{1+\varepsilon_{k-1}}{2}$	0	0	$\frac{1-\varepsilon_{k-1}}{2}$
(x_k,x_1)	0	$\frac{1+\varepsilon_k}{2}$	$\frac{1-\varepsilon_k}{2}$	0

While the other PR-boxes can be obtained by symmetry of the observables and that of the outcomes.

With the below Proposition 2 and Proposition 3, we show that the contextual fraction CF and the signalling fraction SF of a PR-like model can be computed analytically without solving the linear programs as we do in the general case.

Lemma 1. *The only non-signalling PR-like models are the PR-boxes, that is,* $\varepsilon_i = 0$ *for all* $1 \leq i \leq k$*.*

Proof. The no-signalling condition requires that the marginal probability of any observable is independent of the context in which it is measured. Let's consider the observable x_i. It is measured in two contexts: (x_{i-1},x_i) and (x_i,x_{i+1}) (with indices taken modulo k). From the context (x_{i-1},x_i), the marginal probability of $x_i = 0$ is given by:

$$P(x_i = 0) = P(x_{i-1} = 0, x_i = 0) + P(x_{i-1} = 1, x_i = 0) = \frac{1+\varepsilon_{i-1}}{2}.$$

From the context (x_i,x_{i+1}), the marginal probability of $x_i = 0$ is given by:

$$P(x_i = 0) = P(x_i = 0, x_{i+1} = 0) + P(x_i = 0, x_{i+1} = 1) = \frac{1+\varepsilon_i}{2}.$$

For the model to be non-signalling, these two probabilities must be equal for all i:

$$\frac{1+\varepsilon_{i-1}}{2} = \frac{1+\varepsilon_i}{2} \implies \varepsilon_{i-1} = \varepsilon_i.$$

This implies that all ε_i must be equal. Let's call this common value ε. Now consider the special case of the context (x_k, x_1), which is anti-correlated. The marginal probability of $x_1 = 0$ from context (x_1, x_2) is $\frac{1+\varepsilon_1}{2} = \frac{1+\varepsilon}{2}$. The marginal probability of $x_1 = 0$ from context (x_k, x_1) is:

$$P(x_1 = 0) = P(x_k = 0, x_1 = 0) + P(x_k = 1, x_1 = 0) = 0 + \frac{1-\varepsilon_k}{2} = \frac{1-\varepsilon}{2}.$$

Equating the two marginals for x_1:

$$\frac{1+\varepsilon}{2} = \frac{1-\varepsilon}{2} \implies 1+\varepsilon = 1-\varepsilon \implies 2\varepsilon = 0 \implies \varepsilon = 0.$$

Therefore, for a PR-like model to be non-signalling, we must have $\varepsilon_i = 0$ for all i. This corresponds to the definition of a PR-box. □

Proposition 2. *The contextual fraction of a PR-like model is always* 1.

Proof. Consider a PR-like model e with a convex decomposition as in Definition 14:

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

where e^{NC} is a non-contextual and no-signalling model and e^{C} is a model that is not both non-contextual and non-signalling. Recall that the contextual fraction of e is the minimum λ such that the above decomposition remains valid. As the coefficients λ and $(1-\lambda)$ are non-negative, the models e^{NC} and e^{C} are also PR-like models, otherwise the decomposition would not be valid. Lemma 1 states that the only non-signalling PR-like model is the PR box. As the PR box is strongly contextual, there does not exist a valid e^{NC} for the decomposition. Therefore, the minimum λ is 1 and the CF of a PR-like model is always 1. □

Proposition 3. *The signalling fraction of a PR-like model is given by max*$|\varepsilon_i|$.

Proof. Consider a PR-like model e with a convex decomposition as in Equation 4.9:

$$e = (1-\mu)e^{NS} + \mu e^{S},$$

where e^{NS} is a non-signalling model and e^{S} is an arbitrary model allowed to be signalling. The signalling fraction of e is the minimum μ such that the above decomposition remains valid. As per the proof of Proposition 2, the only non-signalling PR-like model is the PR box itself. Therefore, e^{NS} must be a PR box e^{PR} with the same support as e. While minimising μ (that is, maximising $1-\mu$), we have to make sure that e^{S} is a valid model with non-negative probabilities. Hence, we require that

$$e \geq (1-\mu)e^{NS}.$$

The minimum μ is achieved when the inequality is tight for at least one entry in the matrices, i.e. when

$$(1-\mu_{\min})\frac{1}{2} = \min_{i\in\{1,2,3\}}\left(\frac{1\pm\varepsilon_i}{2}\right),$$

where the right-hand side is just the minimum non-zero probability of e. The $\frac{1}{2}$ is due to the fact that a PR box has probabilities of $\frac{1}{2}$ for the possible outcomes. Rearranging the equation gives us the signalling fraction of e:

$$\mathsf{SF} := \mu_{\min} = 1 - \min_{i\in\{1,2,3\}}(1\pm\varepsilon_i) = \max_{i\in\{1,2,3\}}|\varepsilon_i|.$$

Finally, the decomposition corresponding to the minimum μ is given by

$$e = (1-\mathsf{SF})e^{PR} + \mathsf{SF}e^{S},$$

where the model e^{S} can be uniquely determined as $e^{S} = \frac{1}{\mathsf{SF}}(e-(1-\mathsf{SF})e^{PR})$.

□

The two propositions above showed that the values of CF and SF of a PR-like model can be computed analytically without solving the linear programming. Additionally, the criteria for contextuality for signalling model in the sheaf-theoretic framework (4.17) can be specialised to PR-like models as follows:

$$\mathsf{SF} \leq \frac{1}{2M}, \tag{5.1}$$

where M is the number of measurement contexts.

Recall that the measure of contextuality in the CbD framework is

$$\begin{aligned}\mathsf{CNT} =& s_{\text{odd}}\left(\langle R_0^0\, R_1^0\rangle, \langle R_1^1\, R_2^1\rangle, \langle R_2^2\, R_3^2\rangle, \langle R_3^3\, R_0^3\rangle\right) \\ &- \left|\langle R_1^0\rangle - \langle R_1^1\rangle\right| - \left|\langle R_2^1\rangle - \langle R_2^2\rangle\right| - \left|\langle R_3^2\rangle - \langle R_3^3\rangle\right| - \left|\langle R_0^3\rangle - \langle R_0^0\rangle\right| - 2.\end{aligned} \tag{5.2}$$

For a PR-like model, the correlation term $\left\langle R_i^j R_{i\oplus 1}^j \right\rangle$ is either $+1$ for the correlated contexts or -1 for the anti-correlated contexts. By definition, a PR-like model has an odd number of anti-correlated contexts, thus the value of the s_{odd} term is always n.

The quantity Δ is the sum of the absolute differences between the correlations of the same observable in different contexts. For a PR-like model, Δ has the following form:

$$\Delta = |\varepsilon_1 - \varepsilon_2| + |\varepsilon_2 - \varepsilon_3| + \cdots + |\varepsilon_{n-1} - \varepsilon_n| + |\varepsilon_n + \varepsilon_1|$$

Recall that for PR-like models we have $\mathsf{SF} = \max_i |\varepsilon_i|$, which can be used to bound the value of Δ. One can show that for a PR-like model with n contexts, Δ is bounded by:

$$2\mathsf{SF} \leq \Delta \leq \begin{cases} 2n\mathsf{SF} & n \text{ odd} \\ 2(n-1)\mathsf{SF} & n \text{ even} \end{cases}.$$

The derivation of the above inequality can be found in Appendix A.2. The intuition behind the lower bound is that the value of SF makes sure that there must be an ε_i that is at least SF away from 0. The relationship between Δ and SF is illustrated in Figure 6.6. One can readily see that the sheaf contextual region is a strict subset of the CbD contextual region (considering only the subspace where empirical models are allowed). That means any PR-like model that is contextual in the sheaf-theoretic framework is also contextual in the CbD framework, but not vice versa.

5.3 Outcome-symmetric models

Outcome-symmetry captures the idea that measurement outcomes themselves are born equal. An empirical model that respects such symmetry shall be invariant to outcome permutation. In this section, we will give a formal definition of outcome-symmetric models and prove the important fact that all outcome-symmetric models are also non-signalling. The rest of this section make uses of certain aspects of group theory which are introduced in Appendix A.1.

Definition 17 (G-symmetric distribution)**.** *Let G be a permutation group that acts transitively on the set of outcomes O. For the rest of this section, we will abuse the notation and use $g \in G$ to denote both the group element and the group action which maps an outcome to another outcome. A G-symmetric distribution is a joint probability distribution $Pr(x_1, x_2, ..., x_n)$ which is invariant under the action of G on the set of outcomes O. That is, for any $g \in G$,*

$$Pr(x_1 = o_1, x_2 = o_2, ..., x_n = o_n) = Pr(x_1 = g(o_1), x_2 = g(o_2), ..., x_n = g(o_n)), \tag{5.3}$$

where $o_1, o_2, ..., o_n \in O$.

Example 1. *Consider the context (x_1, x_2) with outcomes $O = \{o_1, o_2\}$. The most general form of G-symmetric distribution is given by*

$$Pr(x_1 = o_1, x_2 = o_1) = Pr(x_1 = o_2, x_2 = o_2) = \frac{\alpha}{2}, \tag{5.4}$$

$$Pr(x_1 = o_1, x_2 = o_2) = Pr(x_1 = o_2, x_2 = o_1) = \frac{1-\alpha}{2}, \tag{5.5}$$

where $\alpha \in [0, 1]$. The distribution would manifest as the following row in an *empirical model:*

		(o_1,o_1)	(o_1,o_2)	(o_2,o_1)	(o_2,o_2)
x_1	x_2	$\frac{\alpha}{2}$	$\frac{1-\alpha}{2}$	$\frac{1-\alpha}{2}$	$\frac{\alpha}{2}$

Lemma 2. *The marginal probability distribution of a single observable x from a G-symmetric joint distribution is uniform.*

Proof. It is routine to show that the G-symmetry is preserved under marginalisation. Hence, we should have

$$Pr(x = o) = Pr(x = f_g(o)). \tag{5.6}$$

As G acts transitively on the set of outcomes O, given any two outcomes o_1 and o_2, there is a $g \in G$ such that $g(o_1) = o_2$. Thus, we can assert that $Pr(x = o_1) = Pr(x = o_2)$. As o_1 and o_2 can be arbitrarily chosen, this implies that $Pr(x)$ is a uniform distribution and $Pr(x) = 1/|O|$.

The only thing left to show is that the G-symmetry is preserved under marginalisation. Without loss of generality, consider the marginal probability distribution of the first observable x_1:

$$Pr(x_1 = o_1) = \sum_{o_2,\ldots,o_n \in O} Pr(x_1 = o_1, x_2 = o_2, \ldots, x_n = o_n). \tag{5.7}$$

The symmetry of $g \in G$ gives us the following invariance of the joint probability distribution under the group action f_g:

$$Pr(x_1 = o_1, x_2 = o_2, \ldots, x_n = o_n) = Pr(x_1 = f_g(o_1), x_2 = f_g(o_2), \ldots, x_n = f_g(o_n)). \tag{5.8}$$

Therefore, the right-hand side of (5.7) can be written in terms of the transformed outcomes:

$$Pr(x_1 = o_1) = \sum_{o_2,...,o_n \in O} Pr(x_1 = f_g(o_1), x_2 = f_g(o_2), ..., x_n = f_g(o_n)) \tag{5.9}$$

Since the summation is over all possible values of $o_2, ..., o_n$, we can relabel the summation variables to get

$$Pr(x_1 = o_1) = \sum_{o_2,...,o_n \in O} Pr(x_1 = f_g(o_1), x_2 = o_2, ..., x_n = o_n) \tag{5.10}$$

$$Pr(x_1 = o_1) = Pr(x_1 = f_g(o_1)). \tag{5.11}$$

□

Example 2 ($\mathbb{Z}_2$-Symmetric diachromatic cyclic scenarios)**.** *In a diachromatic k-cyclic scenario, each context consists of exactly two observables and there are exactly two outcomes. The symmetry group in this case is $\mathbb{Z}_2 = \{\mathbb{I}, \sigma\}$, where σ maps an outcome to the other one. The only non-trivial permutation group is $\mathbb{Z}_2$ itself, and it obviously acts transitively on the two outcomes.*

Therefore, the marginal probability distribution of any single observable is uniform over the set of outcomes, as a result of Lemma 2.

Definition 18 (G-symmetric model)**.** *A G-symmetric model is one that contains only G-symmetric local joint distributions.*

Theorem 2. *Any G-symmetry model is non-signalling.*

Proof. Lemma 2 tells us that the marginal distribution of an observable x is always uniform to $1/|O|$, where $|O|$ is the number of outcomes. This is true regardless of which joint distribution is being marginalised. All local joint distributions from every context marginalise to a uniform distribution. Therefore, the model is non-signalling. □

One interesting future work would be to study G-symmetric models where the transitivity on the group action is lifted. Such relaxation would lead to multiple orbits, that is, multiple partitions of the set of outcomes. The marginal distribution would be uniform over the outcomes within an orbit. This means there are non-zero degrees of freedom, as opposed to the zero degree of freedom in the transitive case. No-signalling could no longer be guaranteed if transitivity is lifted.

Chapter 6

PR-anaphora schema

Recall that the motivation for this work is to formalise the idea that contextuality also arises in natural language through resolving ambiguities. We were interested in solving the Winograd schema challenge, so we construct natural language examples containing coreference ambiguities. The examples were carefully constructed so that the resulting logical empirical models exhibit a structure analogous to the PR-box, but generalised to the 3-cyclic measurement scenario.

To construct examples that exhibit contextuality, we follow a two-step process:

1. Select an appropriate measurement scenario that can support contextual empirical models.

2. Carefully craft a set of sentences where ambiguous words can be modelled by the chosen measurement scenario.

For simplicity, we use the 3-cyclic scenario, as the underlying measurement scenario and design the schema in such a way that the resulting empirical model is as close to it as possible. Note that while a 2-cyclic scenario is even simpler, it cannot support contextual empirical models for non-signalling systems. Therefore, we use the 3-cyclic scenario. The measurement scenario of the 3-cyclic scenario reads as follows:

1. Observables $X = \{X_1, X_2, X_3\}$;

2. Contexts $\mathcal{M} = \Big\{\{X_1, X_2\},\, \{X_2, X_3\},\, \{X_3, X_1\}\Big\}$;

3. Outcomes $O = \{O_1, O_2\}$.

We make the following correspondence between the formal description of a measurement scenario and natural language:

1. An outcome is a noun phrase.

2. An observable is an anaphoric phrase that can refer to the outcomes.

3. A context is a sentence containing a subset of the observables.

Definition 19. *The PR-anaphora schema is defined as follows:*

There is an O_1 and an O_2.

1. *One of them is X_1, and the same one is X_2.*

2. *One of them is X_2, and the same one is X_3.*

3. *One of them is X_3, and the other one is X_1.*

Here O_1 and O_2 are noun phrases; X_1, X_2 and X_3 are modifiers used to modify the noun phrases O_1 and O_2.

In definition (19), the modifiers X_1, X_2 and X_3 are the observables in the measurement scenario, while the noun phrases O_1 and O_2 are the outcomes. The contexts are the three different sentences which contain the pair $\{X_1, X_2\}$, $\{X_2, X_3\}$ and $\{X_3, X_1\}$ respectively. Note that the observables should have been the anaphoric phrases *one of them*, *the same one* and *the other one*, but we use the modifiers instead for readability.

The modifiers can be of different types. For example:

1. Adjectival modifiers: *red*, *round*, *sweet*;

2. Prepositional modifiers: *on the table*, *in a dish*, *in the fridge*;

3. Participial modifiers: *being steamed*, *being cooked*, *being chilled*.

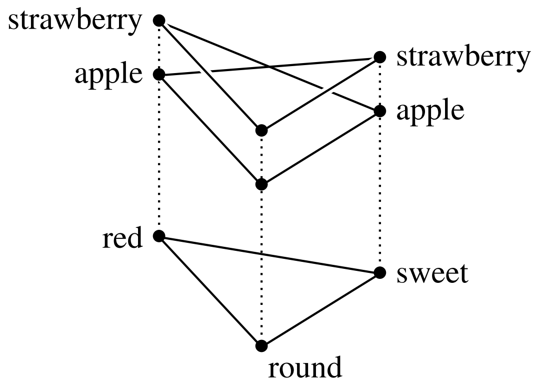


Figure 6.1: Bundle diagram of the PR-anaphora schema with adjectival phrase modifiers shown in Figure 6.2.

(Adjectival) There is an *apple* and a *strawberry*.

1. One of them is *red* and the same one is *round*.
2. One of them is *round* and the same one is *sweet*.
3. One of them is *sweet* and the other one of them is *red*.

Figure 6.2: Example of the PR-anaphora schema with predicate adjective modifiers.

It follows that the schema can be possibilistically modelled by a PR-prism and is logically contextual. The bundle diagram of the PR-anaphora schema is shown in Figure 6.1. However, the schema is not naturally occurring—it is grammatical but somewhat artificial; that is, it is unlikely to appear in natural human conversation. Figures 6.2 and 6.3 show some natural instantiations of it to nouns and their adjectival, verb, and preposition modifiers. Other modifiers can be dealt with similarly.

6.1 Probabilistic PR-anaphora schema

To construct probabilistic models for the PR-anaphora schema, we define a probability distribution over the possible referents of the anaphors using the language model BERT. BERT is pretrained using a Masked Language Model (MLM) objective, where a portion of the input words are randomly masked, and the model is trained to predict the masked words. In practice, masked words are replaced with the special token `[MASK]`. The model then predicts a probability distribution over its entire vocabulary for each masked word. For example, given a sentence such as: `The goal`

(Prepositional) There is an *apple* and a *strawberry*.

1. One of them is *on the table* and the same one is *in a dish*.
2. One of them is *in a dish* and the same one is *in the fridge*.
3. One of them is *in the fridge* and the other one is *on the table*.

(Participial) There is an *apple* and a *strawberry*.

1. One of them is being *steamed* and the same one is being *cooked*.
2. One of them is being *cooked* and the same one is being *chilled*.
3. One of them is being *chilled* and the other one is being *steamed*.

Figure 6.3: Examples of the PR-anaphora schema with participial phrase modifiers and prepositional phrase modifiers.

`of life is [MASK].`, BERT produces a probability distribution over each word in the vocabulary:

token	`life`	`survival`	`love`	`freedom`	`simplicity`	···
prob.	0.1093	0.0394	0.0329	0.0300	0.0249	···

In order to construct probabilistic models for the PR-anaphora schema in definition 19, we go through these two steps: first, as the ambiguities lies in the anaphors, we replace the anaphor "One of them" with the special token `[MASK]`, then the prediction of BERT is interpreted as the probability distribution over the possible referents of the anaphor. As an example consider the following 3 sentences. We feed them separately to BERT:

1. `There is an apple and a strawberry. The [MASK] is red and the same one is round.`
2. `There is an apple and a strawberry. The [MASK] is round and the same one is sweet.`
3. `There is an apple and a strawberry. The [MASK] is sweet and the other one is red.`

BERT will produce probabilities $P_i(\texttt{apple})$ and $P_i(\texttt{strawberry})$ for the i-th sentence shown above. Since BERT assigns probability scores to every word in its vocabulary such that they sum to one, it is generally the case that $P_i(\texttt{apple}) + P_i(\texttt{strawberry}) \neq 1$. We therefore normalise them using the following map [1]:

$$P_i(\texttt{apple}) \mapsto \frac{P_i(\texttt{apple})}{P_i(\texttt{apple}) + P_i(\texttt{strawberry})}$$
$$P_i(\texttt{strawberry}) \mapsto \frac{P_i(\texttt{strawberry})}{P_i(\texttt{apple}) + P_i(\texttt{strawberry})}$$

We use the normalised probabilities to construct a PR-like model (see Section 5.2) on a 3-cyclic measurement scenario with the following empirical table:

	(app., app.)	(app., str.)	(str., app.)	(str., str.)
(red, round)	$P_1(\texttt{apple})$	0	0	$P_1(\texttt{strawberry})$
(round, sweet)	$P_2(\texttt{apple})$	0	0	$P_2(\texttt{strawberry})$
(sweet, red)	0	$P_3(\texttt{apple})$	$P_3(\texttt{strawberry})$	0

All empirical models constructed in this way for the PR-anaphora schema are PR-like models as defined in Section 5.2. BERT will produce different probabilities if the order of the nouns changes, i.e. if we change the schema to 'There is an O_2 and an O_1, different probabilities are assigned to the masks. In order to take these differences into account, we considered both orders of the schema.

Notice that such an empirical model is non-signalling only if $P_i(\texttt{apple}) = P_i(\texttt{strawberry}) = 0.5$ for all i. It is therefore very unlikely that the model is non-signalling. To determine whether a signalling model is contextual, we use the inequality criterion of equation (4.17). According to Proposition 2, the contextual fraction CF of a model that has the same support as the PR prism is always 1. Also, all the examples we considered in this paper have 3 contexts, i.e. $|\mathcal{M}| = 3$. Thus, to tell if such a model is sheaf-contextual, we just need to check if $\mathsf{SF} < 1/6$. Since the value of SF is equal to the maximum absolute value of the ε_i's, an empirical

[1]The normalisation here is equivalent to limiting the vocabulary to just `apple` and `strawberry` when BERT computes the probability scores.

model is sheaf-contextual if BERT is uncertain about the masked word. In the extreme case where BERT is completely uncertain about the masked word, i.e. $\varepsilon_i = 0$ for all i, then $\mathsf{SF} = 0$ and the model reduces to a PR box, which is both strongly contextual and non-signalling. To achieve this, in previous work [92, 93], we chose the referents to be semantically similar, e.g. boy and girl, cat and dog and discovered 350 sheaf-contextual and 9,312 CbD-contextual models out of 11,052 examples.

Apart from the ability to predict masked words, BERT also produces vector representations for each word in its vocabulary. The idea behind using vectors to represent words comes from the distributional model of semantics [94, 95]. In this model, words come with a *distributional* property, whereby their collocational contexts play a role in identifying their meanings. This was motivated by the observation that words that have similar meanings, i.e. are synonymous, are collocated in the context of the same words. A famous example was the words “oculist” and “eye doctor”, both of which occur in the context of “eye”, “glasses”, and “doctor”. Later, it was discovered that this property can be used to reason about phenomena that go beyond synonymy, since words that have any semantic relationship with each other do also occur in the context of same words. Some examples here are “tea” and “coffee”, “cat” and “dog”, and “boy” and “girl”. It was also observed that words that are not semantically related, e.g. “cat” and “coffee” do not co-occur in the same contexts.

Distributional semantics was implemented by the first wave of Natural Language Processing researchers, including Rubenstein and Goodenough [96]. They first embedded a large corpus of natural language data into a matrix. The columns of this matrix were canonical forms from a dictionary and were referred to as “context.” The rows denoted “target” words and were all the words that had occurred in the corpus. Each target word was then represented by its row vector in a vector space generated by the context words as bases. As a result, words that were semantically similar were represented by vectors that were *close* to each other in this space. As an example, see Figure 6.4. We outline the construction below:

1. Fix a set of target and context words. In principle, these can be the set of all

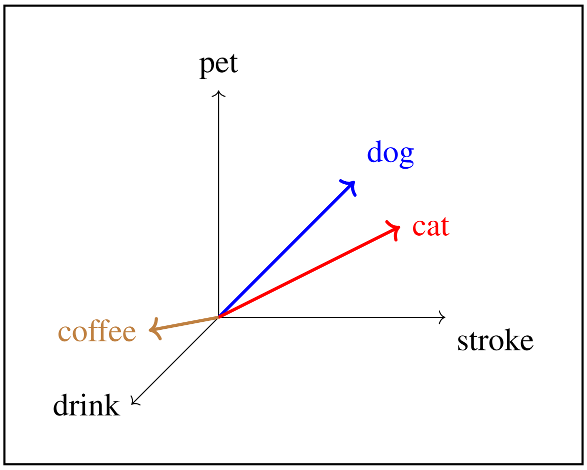


Figure 6.4: An example vector space and its word vectors. The words “dog” and “cat” are close to each other in the vector space, while the words “coffee” and “cat” are far apart.

words in the vocabulary of a language. In practice, one works with the set of canonical forms of major words as context words and a variable set of words as targets. The latter often comes from the vocabulary used in a specific dataset.

2. Form a vector space spanned by the set of context words.

3. Count and normalise the number of times a target word occurs in the neighbourhood (usually a window of 5-10 words) around a context word.

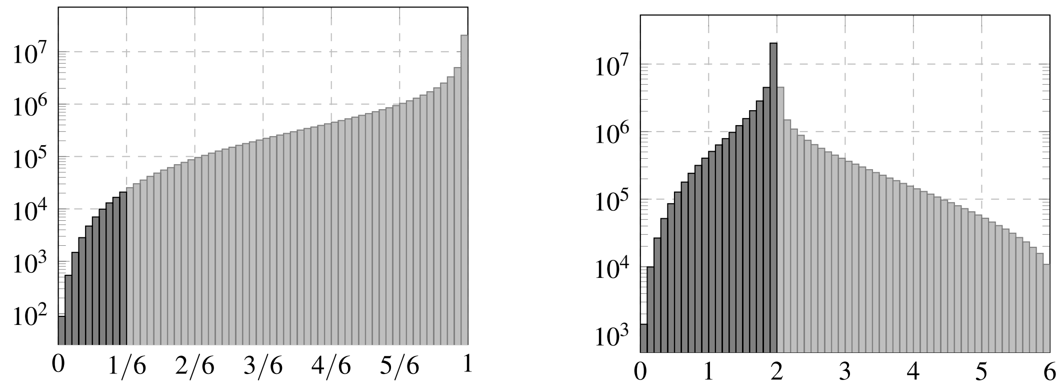
4. Form a vector for each list of normalised counts. These are the vector representations of your target words.

The above construction was initially implemented using information-theoretic measures, such as mutual information and local information [97]. Subsequent advancements introduced neural network-based implementations, with the first notable approach being a single-layer neural network proposed by researchers at Google [98]. This method, known as word2vec, revolutionised word embedding by efficiently capturing semantic relationships in large corpora. More recent implementations leverage the Transformer architecture, which employs multiple hidden layers and the attention mechanism, leading to the development of large language models such as BERT and GPT-3. Unlike traditional distributional models, where word embeddings reside in high-dimensional vector spaces spanned by a large set of context words (often tens of thousands of dimensions), neural network-based algorithms produce embeddings in lower-dimensional spaces. These spaces are spanned by a set of learned abstract features, typically with dimensionality in the range of hundreds. This shift enables more efficient and compact representations while preserving semantic relationships. For instance, the word embeddings of the BERT model we use in this paper are 768-dimensional vectors.

In both models, identifying similar pairs of words has been a cornerstone of their evaluation. Numerous word similarity datasets have been developed to benchmark this capability. Examples include various extracts of the TOEFL test [99], MEN [100], WordSim-353 [101], SimLex-999 [102], and SimVerb3500 [103]. These datasets consist of word pairs, either in or out of context, annotated for their degree of similarity or dissimilarity. Human similarity judgments are collected for these pairs, and the average judgments along with inter-annotator agreements are computed. Over the years, various vector distance measures have been evaluated to determine which measure best aligns with human judgments. It has been established that the cosine of the angle between word vectors is one of the most effective measures for approximating semantic similarity [104, 105, 106, 98]. However, distinguishing finer-grained lexical relationships such as antonymy, synonymy, hypernymy, hyponymy, or co-hyponymy remains a challenging task and has proven to be less straightforward. In line with this approach, we compute the cosine similarity between the noun vectors to quantify the degree of semantic similarity for each pair (O_1, O_2) in the schema.

6.2 Dataset

We adopt a systematic approach to construct a much larger dataset of empirical models in order to investigate the prevalence of contextuality in natural language data on a large scale. To this end, we considered the entire Simple English Wikipedia corpus using a March 2022 snapshot of it which is made available to researchers. This snapshot contains 205,328 articles with 40 million tokens in total, averaging 197 tokens per article. The Simple English Wikipedia is a version of the English Wikipedia that uses a limited vocabulary and simpler grammar. It is designed for



(a) Signalling fraction

(b) Direct influence

Figure 6.5: Histograms of (a) the signalling fraction and (b) the direct influence of the 51,966,480 models constructed for the full dataset Highlighted are the contextual models with (a) $\mathsf{SF} < 1/6$ or (b) $\Delta < 2$. The fraction of sheaf-contextual models is 0.148% and the fraction of CbD-contextual models is 71.1%.

people less proficient in English, such as children and non-native speakers. We chose this corpus due to its relatively small size while still containing a diverse range of topics. We extracted all the adjective-noun phrases from the dataset and used them to construct examples of the PR-anaphora schema. The data set was subjected to the following standard pre-processing steps used in previous work [107]:

1. Each article was tokenised using the `word_tokenize` function in NLTK [108].

2. Each tokenised article was then divided into sentences using the `sent_tokenize` function in NLTK.

3. Each sentence was tagged with the Penn Treebank tag set [109] using the `pos_tag_sents` function in NLTK to obtain the part-of-speech tags for each token. We used the default tagger offered by NLTK, which was a Greedy Averaged Perceptron tagger.

4. Adjective-noun phrases were extracted from the whole tokenised dataset by scanning through the part-of-speech tags for each sentence. Neighbouring tokens with the tags `JJ` and `NN` were extracted as adjective-noun phrases.

After filtering out noise words, e.g. one-letter words and numbers, or nouns that were not in the BERT vocabulary, we obtained 219,633 adjective-noun phrases, 9,521

Nouns		Adjectives			SF	Δ
scholar	opposition	British	great	Russian	0.963	2.217
apprentice	side	last	former	new	0.992	2.626
camera	telescope	special	main	old	0.824	2.737
tea	tree	Japanese	small	Australian	0.787	3.260
prey	day	regular	important	primary	1.000	2.035
fox	passenger	last	American	female	0.748	**1.496**
dwarf	bass	new	Russian	black	0.985	**1.979**
dancer	scene	main	nude	German	0.969	**1.938**
videos	text	short	full	sexual	1.000	**2.000**
series	sheep	regular	single	famous	1.000	**1.999**
fire	relief	American	direct	poor	**0.165**	**0.376**
person	photographer	British	American	French	**0.155**	**0.310**
track	rule	American	new	British	**0.165**	**0.779**
memory	saint	great	important	certain	**0.141**	**0.283**
island	architect	new	British	Japanese	**0.142**	**0.292**

Table 6.1: Randomly selected samples of instances of the PR-anaphora schema with signalling fraction and direct influence values, highlighted when contextual.

nouns, and 21,152 adjectives. To construct examples of the PR-anaphora schema, we chose the 5 most frequent common adjectives for the noun pair. The number was taken to be 5 since this provided us with a good level of overlap and at the same time, a large amount of data. This resulted in 866,108 noun pairs, with which we constructed 51,966,480 examples of the PR-anaphora schema. Samples of the noun pairs and their corresponding adjectives are shown in Table 6.1.

In summary, the pipeline for constructing the dataset of PR-anaphora schema instances (where each instance consists of a noun pair and a triple of adjectives) is as follows:

1. Extract all adjective-noun phrases from the Simple English Wikipedia corpus.

2. Filter out noise words and nouns not present in the BERT vocabulary.

3. Identify all noun pairs that co-occur with at least five common adjectives in the corpus, considering both possible noun orderings separately.

4. For each noun pair, select the five most frequent shared adjectives and generate all possible permutations of three adjectives from this set, resulting in $5!/(5-3)! = 60$ instances per noun pair. Note that both the signalling fraction SF and

the direct influence Δ are invariant under permutation of the first two contexts, so only 30 unique models are obtained per noun pair.

Finally, we constructed empirical models for all the instances using BERT as described in Section 6.1 and computed their signalling fractions SF and direct influences Δ using the results set out in Section 5.2.

6.3 Results

6.3.1 Contextuality in the dataset

We constructed empirical models for all the 51,966,480 examples of the PR-anaphora schema using BERT, in the manner described in Section 6.1. Out of these, 77,118 (0.148%) were found to be sheaf-contextual; 36,938,948 (71.1%) were found to be CbD-contextual. Here sheaf-contextual means that the model has $\mathsf{SF} < 1/6$ and CbD-contextual means that the model has $\Delta < 2$. The low fraction of sheaf-contextual models could be attributed to the strictness of the criterion $\mathsf{SF} < 1/6$, which is sufficient but not necessary.

Figure 6.5 shows the distribution of signalling fraction SF and direct influence Δ of the examples. The distribution of signalling fraction SF can be seen heavily skewed towards 1 and sharply peaking at 1, while the distribution of direct influence Δ sharply peaks at 2. Our hypothesis is that in the PR-anaphora schema examples, BERT often predicts the same word for the masked token in all the contexts with high probability, resulting all the ε values to be either close to -1 or 1. In order to see why, suppose that $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (1,1,1)$ or $(-1,-1,-1)$; this would result in $\mathsf{SF} = \max_i |\varepsilon_i| = 1$ and $\Delta = |\varepsilon_1 - \varepsilon_2| + |\varepsilon_2 - \varepsilon_3| + |\varepsilon_3 + \varepsilon_1| = 2$. Figure 6.6 shows the distribution of the examples in the space of Δ and SF. The majority of the examples are concentrated at the point $(\mathsf{SF}, \Delta) = (1,2)$, which is the point where the ε values are all equal to 1 or -1. One can observe that the sheaf-contextual region is entirely contained within the CbD-contextual region; it follows that the set of sheaf-contextual models is a strict subset of the set of CbD-contextual models. In other words, every sheaf-contextual model is also CbD-contextual, but not every CbD-contextual model is sheaf-contextual.

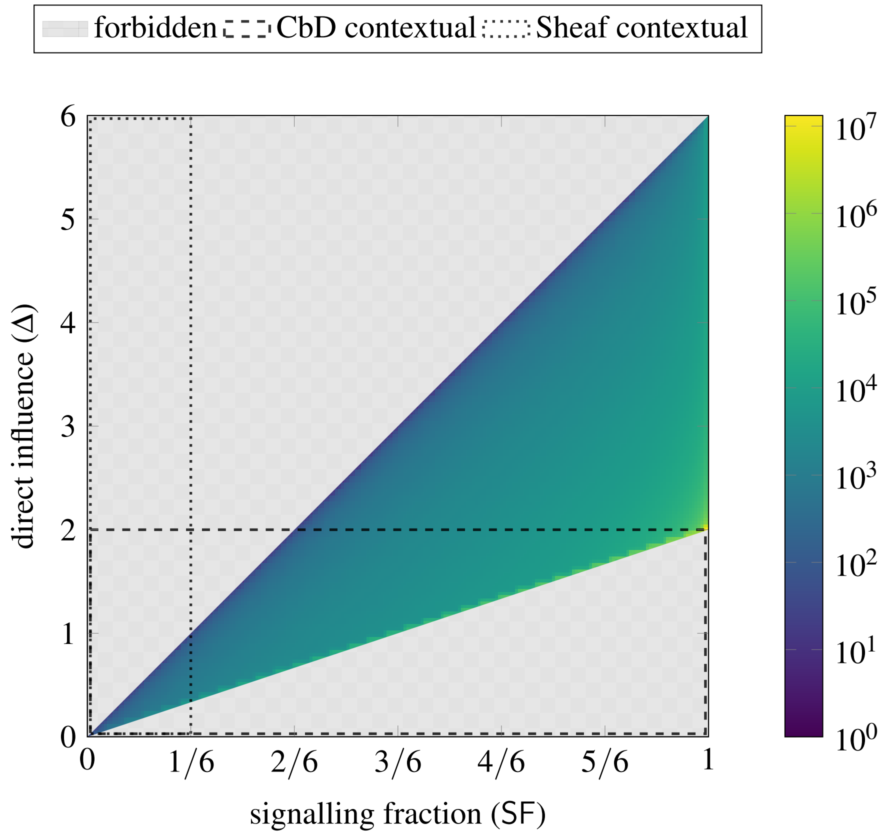
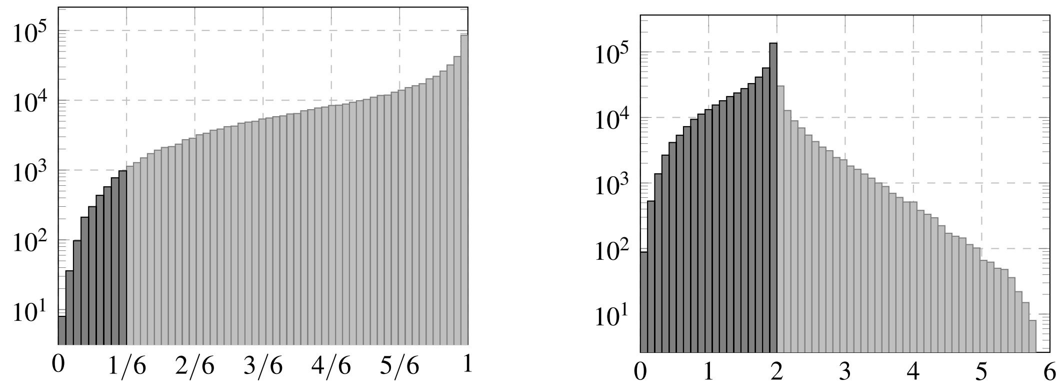


Figure 6.6: The distribution of the instances in the space of direct influence and signalling fraction, which is equally divided into 200 times 200 bins. The colour of each bin represents the log of the number of instances that fall into that bin. As determined by equation (5.2), certain regions of the space are not accessible to the instances, which is shown as *forbidden* in the figure. The regions where the instances are either CbD-contextual or sheaf-contextual are outlined in the figure.

Table 6.1 shows samples of the instances of the PR-anaphora schema with their signalling fraction SF and direct influence Δ.

6.3.2 Similar-Noun Subset of Dataset

The examples of the schema restricted the set of adjectives $\{X_1, X_2, X_3\}$ to the 5 most frequent adjectives of each noun pair (O_1, O_2). It however did not impose any restrictions on the noun pairs themselves. As a result, we come across noun pairs that are very unlikely to have occurred together in the same context. Some of these noun pairs even lead to contextual examples, for instance the pair *(memory, saint)*, from table 6.1, which is both CbD-contextual and sheaf-contextual. Such pairs of nouns can still share adjectives; as one can see both *memory* and *saint*



(a) Signalling fraction **(b)** Direct influence

Figure 6.7: Histograms of (a) the signalling fraction and (b) the direct influence of the 519,660 models constructed for the similar-noun subset of the dataset.

are commonly modified by any of the three adjectives *great, important, certain*. In order to filter out these instances, we only consider pairs of nouns that were semantically similar and formed a similar-noun subset of the dataset by restricting it to with top 1% most semantically similar noun pairs. A selection of these and their degrees of contextuality are presented in table 6.2. This resulted in an increase in the percentage of the contextual instances. The percentage of the sheaf-contextual examples increased to 0.50% from 0.0148% and that of the CbD-contextual ones increased to 81.83% from 71.1%. See figure 6.7 for the histograms of signalling fraction and direct influence of the similar-noun subset of the dataset.

Nouns		Adjectives			SF	Δ
television	tv	nationwide	web	live	0.74	**1.47**
grandmother	grandfather	paternal	great	maternal	0.37	**0.73**
painting	sculpture	modern	great	famous	0.86	2.09
artist	facility	medical	new	national	1.00	5.25
supplier	producer	local	single	main	0.73	**1.62**
railroad	railway	national	new	main	**0.11**	**0.22**
journalist	reporter	black	Italian	American	**0.12**	**0.24**
station	hospital	small	main	large	**0.14**	**0.54**
creature	snake	common	giant	wooden	**0.13**	**0.26**
assassin	journalist	Japanese	American	French	0.58	**1.15**

Table 6.2: A selection of most similar noun pairs and their adjectives.

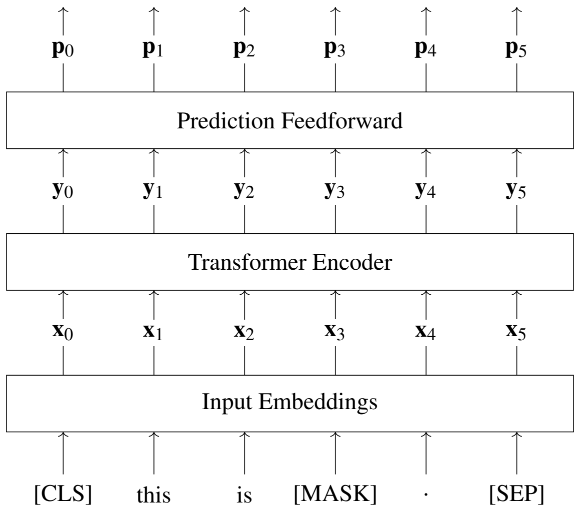


Figure 6.8: A flow chart illustrating how the embedding vectors are transformed into the output vectors in a BERT model. Extra tokens [CLS] and [SEP] are added to the input sequence to indicate the start and end of the sequence, while the [MASK] token is used to indicate the mask.

6.4 Analysis of the Results

BERT (Bidirectional Encoder Representations from Transformers) [35] is a language encoder that is based on the Transformer architecture [21]. Given a sequence of tokens $(x_1, x_2, \ldots, x_n)$, BERT encodes each token with a vector, resulting in a sequence of embedding vectors $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. The embedding vectors are then fed into a transformer-encoder which is a stack of multi-head self-attention layers to produce a sequence of embedding vectors $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$. The self-attention layers allow information to flow between any two positions in the input sequence, thereby modifying the embedding vectors to capture the context of the input sequence. Thus, the embedding vectors are considered to be *contextualised*, rather than *static* as in word2vec [98]. See figure 6.8 for a high level overview of the BERT architecture. In this section we present a geometric interpretation of the predictions of BERT, so that we can relate the factors involved in these predictions to the parameters that affect contextuality.

6.4.1 BERT logit score and the ε parameter of empirical tables

One of the two tasks that BERT was trained on was masked language modelling. In this task, a fraction of the input tokens are masked, and the model is trained to predict the masked tokens. For this purpose, a further feedforward layer was added on top of the stack of self-attention layers to produce a sequence of output vectors $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n)$, one for each token in the input sequence. Suppose that the i-th token is masked. To obtain the predicted distribution of tokens on the i-th token, the corresponding output vector $\mathbf{p}_i$ is compared against the embedding vector of all candidate tokens and a *logit* score is produced for each token. A softmax function is then applied to the logit scores to obtain the probability distribution over the vocabulary. More precisely, the logit score l_j and probability P_j of the j-th candidate token are given by:

$$l_j = \mathbf{p}_i \cdot \mathbf{e}_j + b_j \qquad (6.1) \qquad \text{and} \qquad P_j = \frac{\exp l_j}{\sum_{k=1}^{|V|} \exp l_k}, \qquad (6.2)$$

where b_j is a token-specific bias, $\mathbf{e}_j$ is the embedding vector of the j-th candidate token, and $\mathbf{p}_i$ is the output vector of the masked i-th token. Formally, the logit score of the j-th token in the vocabulary is given by Equation 6.1, where b_j is a token-specific bias term and $\mathbf{e}_j$ is the embedding vector of the j-th token in the vocabulary. The logit scores are normalised by the softmax function to produce the probability distribution in Equation 6.2

In the above, $|V|$ is the size of the vocabulary. In our case, the entire vocabulary comprises our two outcomes, i.e. the two nouns in the PR-anaphora schema. Using equations 6.1 and 6.2, below in Prop. 4 we prove a result which connects the BERT logit scores to the empirical table of the PR-like model describing the PR-anaphora schema:

	(O_1,O_1)	(O_1,O_2)	(O_2,O_1)	(O_2,O_2)
(X_1,X_2)	$\mathbf{p}_{X_1} \cdot \mathbf{e}_{O_1} + b_{O_1}$	0	0	$\mathbf{p}_{X_1} \cdot \mathbf{e}_{O_2} + b_{O_2}$
(X_2,X_3)	$\mathbf{p}_{X_2} \cdot \mathbf{e}_{O_1} + b_{O_1}$	0	0	$\mathbf{p}_{X_2} \cdot \mathbf{e}_{O_2} + b_{O_2}$
(X_3,X_1)	0	$\mathbf{p}_{X_3} \cdot \mathbf{e}_{O_1} + b_{O_1}$	$\mathbf{p}_{X_3} \cdot \mathbf{e}_{O_2} + b_{O_2}$	0

Here, the logit scores are shown instead of probabilities for clarity. The probabilities are obtained by feeding the logit scores into the softmax function per row.

Proposition 4. *The logit scores of the masked token given by BERT relates to the ε parametrisation of the PR-like model as follows:*

$$\varepsilon = \tanh\left(\frac{1}{2}(\mathbf{p}\cdot\Delta\mathbf{x} + \Delta b)\right). \tag{6.3}$$

where $\mathbf{p}$ *is the output vector of the masked token,* $\Delta\mathbf{x} = \mathbf{e}_{O_1} - \mathbf{e}_{O_2}$ *is the difference between the embedding vectors of the two nouns* O_1 *and* O_2*, and* $\Delta b = b_{O_1} - b_{O_2}$ *is the difference between the bias terms of the two nouns in the masked modelling prediction head of BERT.*

Proof. Recall the logit scores for the two outcomes are given by:

$$l_{O_1} = \mathbf{p}\cdot\mathbf{e}_{O_1} + b_{O_1}, \quad l_{O_2} = \mathbf{p}\cdot\mathbf{e}_{O_2} + b_{O_2}. \tag{6.4}$$

Since the probabilities are obtained by applying the softmax function to the logit

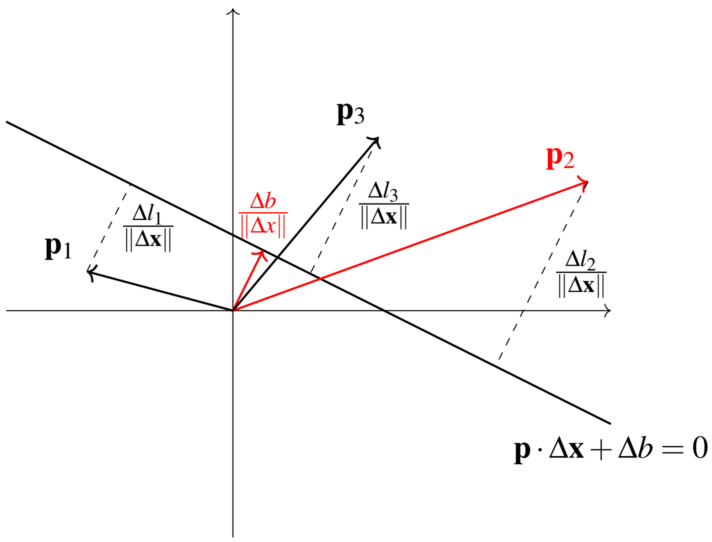


Figure 6.9: A 2-dimensional sketch of a geometric interpretation of the mask predictions from BERT for the PR-anaphora schema. The vectors $\mathbf{p}_i$ are the output vectors of the masked token for the i-th context in the schema. The distance from a predictor vector $\mathbf{p}_i$ to the hyperplane defined by the equation $\mathbf{p}\cdot\Delta\mathbf{x} + \Delta b = 0$ coincides with $\Delta l_i/\|\Delta\mathbf{x}\|$. As ε_i relates to Δl_i monotonically, specifically $\varepsilon_i = \tanh(\Delta l_i/2)$, the signalling fraction $\mathsf{SF} = \max|\varepsilon_i|$ depends only on the prediction vectors the furthest away from the hyperplane. In the figure, the prediction vector $\mathbf{p}_2$ (coloured red) is the furthest away from the hyperplane.

scores, the ratio of the probabilities is given by:

$$\frac{P_{O_1}}{P_{O_2}} = \frac{e^{l_{O_1}}}{e^{l_{O_2}}} = e^{l_{O_1} - l_{O_2}}. \tag{6.5}$$

By definition, $P_{O_1} = \frac{1+\varepsilon}{2}$ and $P_{O_2} = \frac{1-\varepsilon}{2}$. Substituting these into the ratio, we have:

$$\log \frac{1+\varepsilon}{1-\varepsilon} = l_{O_1} - l_{O_2} = \mathbf{p} \cdot \Delta\mathbf{x} + \Delta b. \tag{6.6}$$

Using the identity $\tanh^{-1}(x) = \frac{1}{2} \log \frac{1+x}{1-x}$, we can express ε as:

$$\varepsilon = \tanh\left(\frac{1}{2}(\mathbf{p} \cdot \Delta\mathbf{x} + \Delta b)\right). \tag{6.7}$$

This completes the proof. $\square$

Note that the tanh function is monotonically increasing. Therefore, we can use the difference in logit scores given below as a proxy for the value of ε.

$$\Delta l := \mathbf{p} \cdot \Delta\mathbf{x} + \Delta b \tag{6.8}$$

The value Δl can be interpreted as $||\Delta\mathbf{x}||$ times the distance from the prediction vector $\mathbf{p}$ to the hyperplane defined by the equation $\mathbf{p} \cdot \Delta\mathbf{x} + \Delta b = 0$. A visualisation of this geometric interpretation is shown in Figure 6.9. Assuming an isotropic distribution of the prediction vectors $\mathbf{p}$, equation (6.8) suggests that the value of Δl is directly proportional to the Euclidean distance between the embedding vectors of the two nouns $||\Delta\mathbf{x}||$, which in turn non-linearly scales the value of ε through equation (6.3). Since a higher ε value implies less contextuality in both the sheaf and CbD frameworks, we expect that the value of $||\Delta\mathbf{x}||$ plays an important role in determining whether a model is contextual. Equation (6.8) can be thought as a hyperplane in the word embedding space of BERT that allows a geometric interpretation of the predictions of BERT for the PR-anaphora schema which is shown in Figure 6.9. The bias difference Δb serves to offset the hyperplane from the origin, while the difference in embedding vectors $\Delta\mathbf{x}$ determines the orientation of the hyperplane.

Feature	Sheaf-theoretic Kendall	Sheaf-theoretic Spearman	Sheaf-theoretic Pearson	CbD Kendall	CbD Spearman	CbD Pearson
nouns_entropy	-0.0202	-0.0302	-0.0224	-0.0125	-0.0188	-0.0140
adjectives_entropy	-0.0214	-0.0322	-0.0257	-0.0132	-0.0198	-0.0167
bert_euclidean_dist	**0.0587**	**0.0877**	**0.0809**	**0.0450**	**0.0674**	**0.0590**
bert_bias_diff	0.0334	0.0500	0.0123	0.0115	0.0173	-0.0054

Table 6.3: Correlation coefficients between the features and the contextuality of the instances of the PR-anaphora schema, in the full dataset.

6.4.2 Factors affecting contextuality

In the previous section, we showed that the differences in BERT's logit scores, i.e. Δl can be used as a proxy for the values of ε, which are used to compute the entries of the empirical tables. Further, in Equation 6.8 we showed that Δl is directly proportional to Euclidean distance between the vectors of the two nouns in the instances of the PR-Anaphora schema. Although this finding relates Euclidean distance to contextuality, it does not rule out other features of either the vectors of the nouns or the nouns themselves that might affect it too. For instance, Equation 6.8 also hosts the variable Δb, which is the difference between the bias terms of the vectors of the two nouns. In this section, we are interested in finding out which one of these two is most correlated with contextuality. Specifically, we compute the degree of correlation between three features of BERT's predicted noun vectors, as well as two other independent features of the nouns. Our goal is to investigate which one of these features correlate best with contextuality. For correlation, we compute Spearman, Pearson and Kendall degrees. For BERT features, we consider Euclidean

Feature	Signalling fraction (SF) linear	Signalling fraction (SF) cubic	Direct influence (Δ) linear	Direct influence (Δ) cubic
nouns_entropy	0.0005	0.0006	0.0002	0.0002
adjectives_entropy	0.0007	0.0007	0.0003	0.0004
bert_euclidean_dist	**0.0065**	**0.0092**	**0.0035**	**0.0050**
bert_bias_diff	0.0001	0.0003	0.0000	0.0001

Table 6.4: Comparison of the R^2 values of the linear and cubic regression models, in the full dataset.

Feature	Sheaf-theoretic (SF < 1/6) Kendall	Spearman	Pearson	CbD (Δ < 2) Kendall	Spearman	Pearson
nouns_entropy	-0.0436	-0.0654	-0.0625	-0.0394	-0.0592	-0.0541
adjectives_entropy	-0.0227	-0.0341	-0.0362	-0.0193	-0.0288	-0.0246
bert_euclidean_dist	**0.1234**	**0.1837**	**0.1963**	**0.1155**	**0.1722**	**0.1787**
bert_bias_diff	-0.0821	-0.1241	-0.0996	-0.0670	-0.1005	-0.0679

Table 6.5: Correlation coefficients between the features and the contextuality of the instances of the PR-anaphora schema, in the similar-noun subset of the dataset.

distances and the difference between the biases of vectors. In order to compute these distances, we use the pretrained BERT model `bert-base-uncased` implemented in the HuggingFace Transformers library [110].

Computing the differences between features of word vectors are not the only ways of measuring and comparing the statistical information encoded in them. In fact, the general rule governing BERT is that it chooses the word that has occurred most in the corpus. This is too rough of a feature to be used in our schema instances, since we need BERT to choose between the words with equal probability. Here entropy can come to help. Entropy is an often used method when it comes to computing the imbalances between word frequencies. If two words have similar frequencies, entropy will peak. On the other hand, if one has a low and the other a high frequency, we will have a low entropy. In order to find out whether entropy is related to contextuality, we compute degrees of correlation between both of our contextuality measures and the entropy of nouns and adjectives.

Let us first consider the full dataset. In this dataset, the correlation scores

Feature	Signalling Fraction (SF) linear	cubic	Delta linear	cubic
nouns_entropy	0.0001	0.0003	0.0003	0.0004
adjectives_entropy	0.0012	0.0018	0.0024	0.0027
bert_euclidean_dist	**0.0779**	**0.0803**	**0.0573**	**0.0581**
bert_bias_diff	0.0036	0.0109	0.0020	0.0059

Table 6.6: Comparison of the R^2 values of the linear and cubic regression models, in the similar-noun subset of the dataset.

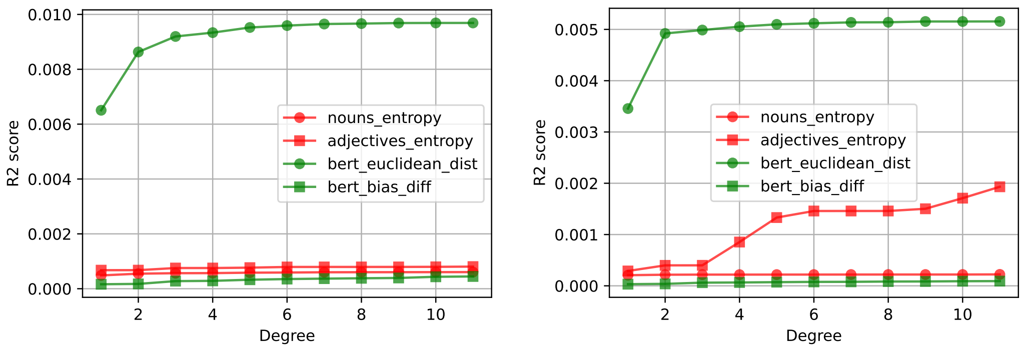


Figure 6.10: The R^2 scores of the polynomial regression models at different polynomial degrees predicting (left) the signalling fraction and (right) the direct influence.

between each of the above features and both of our contextuality measures, i.e. CbD's delta and sheaf theory's SF, are shown in Table 6.3. First, we observe that the correlations are very low in general, which is in the order of a few percent at best. This is surprising, but not entirely unexpected as the factors affecting predictions of BERT are quite complex. Equation 6.3 shows that $\Delta\mathbf{x}$ (and through this, the Euclidean distance $||\Delta\mathbf{x}||$) is not the only factor affecting ε. The bias difference Δb and the final-layer output vector $\mathbf{p}$—which encapsulates the influence of all learned parameters and contextual information processed by the transformer—also contribute to the value of ε and thus to contextuality.

The highest correlations for any of the correlation scores are with Euclidean distance. BERT's bias differences provided the second-best set of correlations, although they were much lower than Euclidean distance. Finally, the nouns entropy and the adjectives entropy both anti-correlated with both SF and Δ.

The above correlations are all statistically significant (p-values are all below 0.01), but on the low side (below 0.005). This indicates the presence of a correlation. To test this, we trained a polynomial logistic regression model, on a range of degrees from 2 to 10. We chose the cubic degree polynomial as a cut-off point. The R^2 values for linear vs cubic correlations are shown in Table 6.4. The results for all the 10 degrees are plotted in Figure 6.10. Clearly, there is a 2-3 times increase in the correlations of the cubic models in comparison to the linear one. Again, the highest correlation was with Euclidean distance. Since R^2 is the square of Pearson's correlation, all the values are positive. Naturally, with the R^2, the entropies provided better correlations than BERT bias differences, but both of these were still quite low. This shows that Euclidean distance plays a more important role when it comes to predicting contextuality, albeit still not a fairly weak one.

The similar-noun subset led to very similar results, see Table 6.5. Here, again we observe (1) an increase in the cubic regression correlations in comparison to the linear ones, and (2) Euclidean distance provides the highest correlation with both SF and delta. These provide further experimental evidence that the Euclidean distances between BERT's word vectors are the best statistical predictors of degrees of contextuality.

6.5 Interpretation of contextuality

The results we obtained demonstrate that quantum-like contextuality can be found in ambiguous schemas of natural language. However, whether that implies human languages are genuinely quantum in nature is most certainly a question that needs further explanation. An important point to note is that in the sheaf-theoretic notion of contextuality, quantum mechanism is just one example of a system that exhibits contextuality. Contextuality or quantum-like contextuality is not defined by quantum mechanics, and a contextual system does not necessarily have to be quantum-mechanical in nature.

To understand what does it mean for a PR-anaphora schema to be contextual in the sheaf-theoretic sense, we need to recall that an empirical model is contextual if there does not exist a global joint distribution which marginalises to all the local joint distributions in the empirical model, given that the empirical model is non-signalling. According to Fine's theorem, the existence of a global joint distribution is equivalent to the existence of a non-contextual realistic hidden-variable model. Hence, we should understand what a hidden-variable model looks like in the context of the PR-anaphora schema. A hidden variable dictates the pre-existing outcome of every observable in the system. In the PR-anaphora schema, the observables are the adjectives, and the outcomes are the nouns. For example, in the apple-strawberry schema, a hidden variable would assign either apple or strawberry to each of the adjectives red, sweet and round. To be more clear, the following table shows all possible hidden-variable assignments for the apple-strawberry schema:

	red	sweet	round
λ_1	apple	apple	apple
λ_2	apple	apple	strawberry
λ_3	apple	strawberry	apple
λ_4	apple	strawberry	strawberry
λ_5	strawberry	apple	apple
λ_6	strawberry	apple	strawberry
λ_7	strawberry	strawberry	apple
λ_8	strawberry	strawberry	strawberry

In the sheaf-theoretic framework, if the probabilistic mixture of these hidden-variable assignments $P(\lambda)$ could explain an observed non-signalling empirical model, then the empirical model would be non-contextual, otherwise it is contextual. For signalling empirical models, we still assume there is a probabilistic mixture over hidden-variable assignments. However, when signalling is present, such mixtures alone cannot fully explain the empirical model; the ontological model must also deviate from such mixture to include signalling effects to match the observed data.

While in the CbD framework, contextuality is defined in terms of the existence (or non-existence) of a global coupling of all random variables involved. Specifically, a system is contextual if it is impossible to construct a joint distribution that (1) reproduces the empirical distributions within each context as marginals, and (2) achieves maximal possible agreement (coupling) between random variables representing the same observable in different contexts. This approach does not require the assumption of pre-existing outcomes for observables, and is applicable even in the presence of signalling.

6.6 Discussion and Conclusion

In this paper, we set to find out whether quantum contextuality can occur in large language models. We built a linguistic schema and modelled it over a quantum contextual scenario. We then instantiated this schema using a snapshot of the Simple English Wikipedia. Probability distributions of the instances were collected using the masked word prediction capability of the large language model BERT. Since natural language data is signalling, one should work in more general frameworks, such as the Contextuality-by-Default (CbD) and the signalling corrected version of the sheaf-theoretic model of contextuality. Computing degrees of contextuality in either of these frameworks led to the discovery of many contextual instances in both the CbD and sheaf-theoretic framework. In order to investigate the reason behind this discovery, we worked with features of BERT's predicted vectors and degrees of contextuality, and derived an equation between the two. More specifically, we showed that the differences in BERT's logit scores can be used as a proxy for the values of ε. The former is directly proportional to Euclidean distance, and the latter is used to compute the entries of the empirical tables.

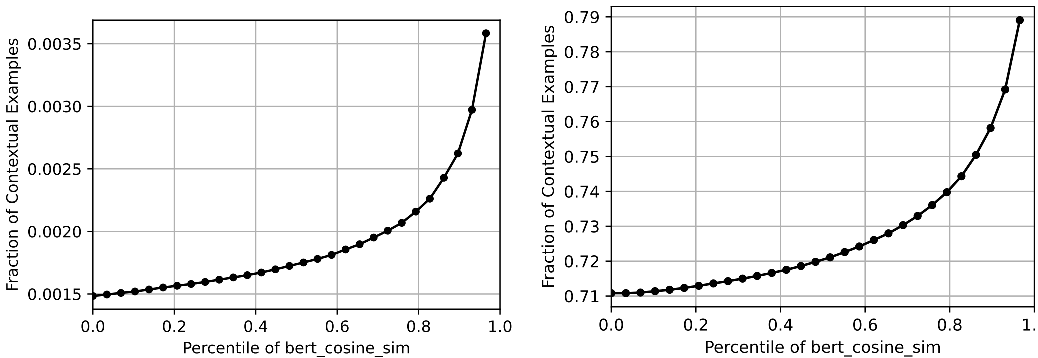


Figure 6.11: The fraction of sheaf-contextual (left) and CbD-contextual (right) instances at different subsets of the dataset created by considering the most similar noun pairs at different percentile thresholds.

The percentage of contextual instances and the degrees of correlation with contextuality were much higher in the subset of the dataset with semantically similar nouns in comparison to the full dataset. Figure 6.11 plots the R^2 for different thresholds for the full dataset. This plot shows that the number of contextual instances increases as we increase the similarity thresholds. These results match the ones we obtained previously, also using BERT predictions [92, 93], where we had a much smaller dataset, consisting of only 11 pairs of nouns and 11,052 empirical models (vs 866,108 pairs of nouns and 51,966,480 empirical tables of the current paper). The noun pairs of that dataset were chosen to be highly similar, e.g. (cat, dog), (girl, boy) and (man, woman). Working with highly similar noun pairs led to 350 sheaf-contextual (3.1%) and 9,321 CbD-contextual (84%) models. To measure semantic similarity, we used the cosine similarity, which is defined as the cosine of the angle between two word embedding vectors. This is one of the most commonly used measures of semantic similarity between two words given their word embeddings. Two more similar words will have smaller angles between their vectors, and thus a higher cosine similarity. We conjecture that the increase in the degree of contextuality in the similar-noun subset is because Euclidean and cosine distances are related to each other. Note that the cosine similarity is related to the Euclidean distance via the identity

$$\|u-v\|^2 = (u-v)\cdot(u-v) = \|u\|^2 + \|v\|^2 - 2\|u\|\|v\|\cos(u,v), \tag{6.9}$$

where u and v are the word embeddings of the two words. If the vector norms are approximately equal, i.e. $\|u\| \approx \|v\|$, and for illustration purposes, we assume unit norm $\|u\| = \|v\| = 1$, the identity simplifies to $\|u-v\|^2 = 2-2\cos(u,v)$. This indicates a linear relationship between the cosine similarity and the Euclidean distance squared. In fact, the mean of vector norms in the similar-noun subset is 1.19 with a standard deviation of 0.14.

The fact that there are overwhelmingly more CbD-contextual models than sheaf-contextual models in our results is intriguing and raises many questions. This discrepancy highlights fundamental differences in the criteria for contextuality between these two frameworks. The contextual bounds in the (Δ, SF) space of empirical models are orthogonal to each other, as depicted in Figure 6.6. This orthogonality suggests that the two frameworks are capturing different aspects of contextuality. In the sheaf-theoretic framework, a signalling empirical model is considered contextual if any hidden-variable explanation of the empirical data must exhibit more signalling than the observed data. (See Section 4.1.8 for more details.) In the CbD framework, a system is considered contextual if it is impossible to construct a global coupling (of all the random variables involved) that simultaneously (1) reproduces the empirical distributions within each context as marginals, and (2) achieve the maximal possible coupling for random variables representing the same content in different contexts. Failure to achieve this means that the empirical model is contextual in the CbD sense.

These differences underscore the complexity of contextuality in signalling data and exploring their significance in natural language data and tasks is left for future work.

To conclude, we demonstrated that a variant of quantum contextuality can be observed in the predictions of large language models. Contextuality leads to quantum advantage [9]. It remains to show whether quantum-like contextuality also leads to advantage, and if so what kind of advantage will it be and how can it be obtained. Finding answers to these questions is a future direction. In the meantime, one also needs to substantiate how this potential advantage can be used in improving methods that natural language tasks. Our linguistic schema is closely related to a well known coreference resolution task known as the Winograd Schema Challenge (WSC) [46]. WSC was proposed as a benchmark for measuring machine intelligence. The idea behind it is that solving the task requires common sense and access to external knowledge, which humans have, but machines do not. In previous work, we showed that the measurement scenario of the original WSC is too simple to host contextuality [111]. The schema presented in this paper offers a suitable generalisation of it. In this paper, we showed how machines, i.e. the large language model BERT can be used to solve it. It remains to collect human judgments and compare their performances.

An advantage of transformer-encoder models with bidirectional attentions such as BERT over the state-of-the-art decoder models such as the GPTs is that its encoder architecture allows masked language modelling, which is a crucial tool for obtaining probability distributions for the instances of our linguistic schema. It could be possible to use GPTs for our purpose with a carefully designed prompt, which is a future direction.

Coreference ambiguity has given rise to other historical challenges in Linguistics and Computational Linguistics. The most difficult cases arise when pronouns are used together with quantifiers and indefinites. The term "donkey anaphora" is used to denote a family of such challenges, defying compositionality and posing many challenges to existing formal models of syntax and semantics. Donkey anaphora have been treated using sheaves [112]. Collecting data for these examples using large language models and investigating whether they can host quantum contextuality is another future direction. in all possible ways;

Chapter 7

Generalised Winograd Schema

In this chapter, we present our approach for the generalisation of the Winograd Schema, enabling the potential observation of contextuality. We will first explain why the original Winograd Schema is not sufficiently complex to exhibit contextuality, and then propose a generalised Winograd Schema that is sophisticated enough to host contextuality.

7.1 Modelling Winograd Schemas as measurement scenarios

To study the contextuality in the original 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 [78, 79], where an ambiguous word is considered an observable in a measurement scenario, and a pair of ambiguous words form a measurement context.

In the original Winograd Schema, one ambiguous pronoun is used in the twin pair of sentences. If we follow strictly the approach of "ambiguous words as observables", then we will end up with a trivial measurement scenario, where there is only one observable, that is, 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 linguistical contexts. The possible outcomes of each of the two observables are the candidate referents of the pronoun.

Definition 20 (Winograd Schema scenario)**.** *Given a Winograd Schema with two noun phrases A and B; an ambiguous pronoun* ***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* $X = \{(\boldsymbol{p}, s),\, (\boldsymbol{p}, a)\}$*;*
- *contexts* $\mathcal{M} = \big\{\{(\boldsymbol{p}, s)\}, \{(\boldsymbol{p}, a)\}\big\}$*;*
- *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 $X = \{(\textbf{they}, \textit{feared}),\ (\textbf{they}, \textit{advocated})\}$;
- contexts $\mathcal{M} = \big\{\{(\textbf{they}, \textit{feared})\},\ \{(\textbf{they}, \textit{advocated})\}\big\}$;
- outcomes $O = \{\text{city councilmen},\ \text{demonstrators}\}$.

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.

7.2 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:

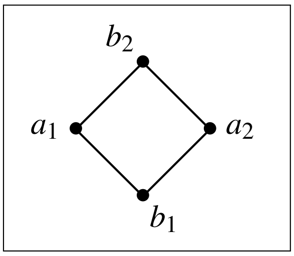
1. A generalised schema consists of four slightly differing sentences. The first sentence contains two special words s_1 and s_2. Similar to the original Winograd Schema, s_1 can be replaced by an alternate word a_1 and s_2 can be replaced by an alternate word a_2. The possibility of replacing special 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.

(a) 

(b)

	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
(a_1,b_1)	$1/2$	0	0	$1/2$
(a_1,b_2)	$3/8$	$1/8$	$1/8$	$3/8$
(a_2,b_1)	$3/8$	$1/8$	$1/8$	$3/8$
(a_2,b_2)	$1/8$	$3/8$	$3/8$	$1/8$

Figure 7.1: (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).

Definition 21 (Generalised Winograd Schema scenario)**.** *Given a Generalised Winograd Schema with two noun phrases A and B; two ambiguous pronouns $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ can each refers 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* $X = \{(\boldsymbol{p}_1,s_1),(\boldsymbol{p}_1,a_1),(\boldsymbol{p}_2,s_2),(\boldsymbol{p}_2,a_2)\}$
- *contexts* $\mathcal{M} = \{\{(\boldsymbol{p}_1,s_1),(\boldsymbol{p}_2,s_2)\},\{(\boldsymbol{p}_1,s_1),(\boldsymbol{p}_2,a_2)\},$ $\{(\boldsymbol{p}_1,a_1),(\boldsymbol{p}_2,s_2)\},\{(\boldsymbol{p}_1,a_1),(\boldsymbol{p}_2,a_2)\}\};$
- *outcomes* $O = \{A,B\}$.

Such a measurement scenario is called a Generalised Winograd Schema scenario, *or GenWino scenario in short.*

The generalised WS scenario is isomorphic, i.e. identical upon relabelling, to the Bell-CHSH scenario shown in Figure 7.1. It has long been known that the Bell-CHSH scenario can host contextual models [3, 5]. Thus, a carefully designed generalised Winograd Schema would be able to demonstrate contextuality.

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

- The trophy doesn't fit into the suitcase because **it**$_1$ is too [s_1 = *small* / a_1 = *large*]. Nonetheless, **it**$_2$ is [s_2 = *light* / a_2 = *heavy*].

The corresponding generalised WS scenario is given by:

- observables $X = \{(\textbf{it}_1, \textit{small}), (\textbf{it}_1, \textit{large}), (\textbf{it}_2, \textit{light}), (\textbf{it}_2, \textit{heavy})\}$
- contexts $\mathcal{M} = \begin{array}{l}\{\{(\textbf{it}_1, \textit{small}), (\textbf{it}_2, \textit{light})\}, \{(\textbf{it}_1, \textit{small}), (\textbf{it}_2, \textit{heavy})\}, \\ \quad \{(\textbf{it}_1, \textit{large}), (\textbf{it}_2, \textit{light})\}, \{(\textbf{it}_1, \textit{large}), (\textbf{it}_2, \textit{heavy})\}\};\end{array}$
- outcomes $O = \{\text{trophy}, \text{suitcase}\}$.

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

- 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 [113, 114].

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.

7.2.1 An example of the generalised Winograd Schema

As our goal is to uncover contextual models in natural language, we need to gather judgement data from human participants to build empirical models for generalised Winograd Schema instances. Crucially, deterministic systems lack contextuality. Therefore, our generalised 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 Schema[1]:

- 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*]$_2$.

We decided to use the referring phrase **one of them** instead of the third-person pronoun **it** to improve the naturalness of the example. Note that we have chosen to name the two candidate referents by alphabetical symbols, A and B, rather than by actual names. This is done to minimise any bias towards the candidates. For instance, one might prefer to choose a name with a specific gender. The reason why debiasing is so important here is that a model would have a higher contextual fraction if it is closer to a PR-box. Now recall that PR-boxes are outcome-symmetric. By making the two candidate referents symmetric by construction, we ensure that empirical models of the example to be outcome-symmetric as well. While the two candidates

[1] It was pointed out by one of the reviewers during submission to QPL that the original version of the example contains two incorrect uses of English. Here we provide the corrected version of the example.

are constructed to be as symmetric as possible, perfect symmetry is not fully achieved: A is always mentioned before B, and subtle ordering effects or associations may still influence interpretation. For further details on outcome-symmetric models and their implications, refer to Section 5.3.

Enforcing the outcome-symmetry brings also two additional benefits. Firstly, all outcome-symmetry models are also non-signalling. This allows us to use the standard sheaf-theoretic framework without signalling corrections. On top of that, the CbD framework agree with the sheaf-theoretic one for non-signalling models. Secondly, it has already been shown that LLMs have reached human-level at coreference resolution. Thus, having a single correct solution is no longer challenging for the modern LLMs, and therefore coreference resolution can no longer be thought of as a proxy test for reasoning capabilities. Instead, the LLM will be not resolving a single reference, but rather understanding the relationship between the first and the second referring phrases. We speculate that such understanding takes

7.2.2 Human judgements on the example

Following the methodology of prior studies that investigated contextuality in human decision-making [115, 116], we gathered human judgements for this example using a questionnaire administered via Amazon Mechanical Turk. 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 them**$_1$ and **one of them**$_2$. A screenshot of the questionnaire is shown in Figure 7.2.

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 respect the symmetry between A and B.

A total of 410 responses were collected on Amazon Mechanical Turk separately

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 ${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 ${word2}.

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

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

Figure 7.2: A screenshot of the template of the questionnaire. The placement holders ${word1} and ${word2} are instantiated with the two special words or the alternate words of the generalised Winograd Schema. In this example, ${word1} can be either *cannibalistic* or *herbivorous* and ${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 verb tense inconsistencies, with a mixture of present and past tenses. Unfortunately, we did not notice these until a reviewer pointed them out, after data collection.

on two dates: 20 Oct 2022 and 23 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. (Note that this symmetry filtering is not done in [115, 116].) The respondents were each 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 7.1. The model violates the Bell-CHSH inequality by 0.192. Since the model is symmetric in the outcomes by construction, it is non-signalling and thus the measure of contextuality CNT in the CbD framework coincides with the degree of violation [88]. The symmetry in the outcomes also allows the violation to saturate the bound defined by CF in sheaf-theoretic framework [87], i.e. the following equality is attained

$$\mathsf{CF} = \max\left\{0, \frac{1}{2}\,\mathsf{violation\ of\ Bell\text{-}CHSH\ inequality}\right\}. \tag{7.1}$$

(a)	(A, A)	(A, B)	(B, A)	(B, B)
(*canni*, *hungry*)	0.402	0.097	0.097	0.402
(*canni*, *alive*)	0.044	0.455	0.455	0.044
(*herbi*, *hungry*)	0.345	0.154	0.154	0.345
(*herbi*, *alive*)	0.344	0.155	0.155	0.344

(b)	(A, A)	(A, B)	(B, A)	(B, B)
...	$1/2$	0	0	$1/2$
...	0	$1/2$	$1/2$	0
...	$1/2$	0	0	$1/2$
...	$1/2$	0	0	$1/2$

Table 7.1: (a) The empirical model constructed with the 410 human judgements collected from Amazon Mechanical Turk. The violation of Bell's inequality of the model is 0.192. 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.

That means that the symmetry renders the Bell-CHSH inequality the strongest possible test for contextuality, that is, violations of any other inequality would be no larger than that of the Bell-CHSH inequality. As CNT reduces to the Bell-CHSH inequality for symmetric models, our model is considered contextual in both the sheaf-theoretic framework and the CbD framework.

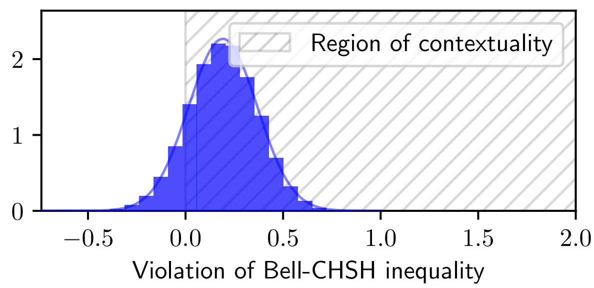


Figure 7.3: A normalised histogram of the Bell-CHSH inequality violation for 100,000 bootstrap samples from the model shown in Table 7.1. A positive violation, indicative of contextuality, is observed in 87% of the resampled models. The standard deviation of the distribution is 0.176.

To test the significance of this result, we performed a one-sided hypothesis test with the null hypothesis that the Bell-CHSH violation is non-positive (H_0 : violation ≤ 0). We used bootstrap resampling with 100,000 bootstrap samples to estimate the sampling distribution of the violation statistic. The resulting distribution is depicted in Figure 7.3.

The bootstrap p-value, calculated as the proportion of bootstrap samples with a non-positive violation, is approximately 0.13. Since this p-value is greater than the conventional significance level of $\alpha = 0.05$, we cannot reject the null hypothesis. This provides suggestive but not statistically significant evidence for contextuality. Consistent with this, the one-sided 95% bootstrap confidence interval for the violation is $[-0.10, \infty)$, which includes zero.

The current results are suggestive of Bell violation, but not yet statistically significant. More data would be needed to verify the contextuality of the example with high confidence.

7.3 LLM evaluation of the generalised Winograd Schema

In the previous section, we have seen that the generalised Winograd Schema (GenWino) exhibits contextuality when judged by human subjects. With the advent of large language models (LLMs) which have been demonstrated to be able to tackle a wide range of reasoning tasks [117, 118], in particular reaching human performance on the original Winograd Schema Challenge [50, 119], it is natural to ask whether LLMs can also be used to reproduce the contextuality of GenWino.

In this section, we evaluate LLMs on the contextuality of GenWino. To adapt the questionnaire for LLMs, we refined the instructions to include a clear response format and required the LLMs to provide reasoning for their choices. This adjustment aimed to enhance the transparency and interpretability of the evaluation process. The modified template is shown below:

GenWino prompt for LLMs

```
Instruction:
Please read the following short story, then select the two
interpretations you think are the most likely. Your answers
should be comma-separated, followed your reasons for choosing
the two and your reasons for not choosing the other two. An
```

```
example response: ‘1, 2. It is because ...‘.

Story:
A and B belong to the same {word1} species of animals. On
a hot afternoon south of the Sahara,
one of them was very hungry. They noticed each other while
they were roaming in the field. After a while, it was no
longer {word2}.

Question:
The following are four different interpretations of the story.
Which two of the following interpretations are the most likely?

(1) A was the very hungry {word1} animal.
    A was no longer {word2} after a while.
(2) A was the very hungry {word1} animal.
    B was no longer {word2} after a while.
(3) B was the very hungry {word1} animal.
    A was no longer {word2} after a while.
(4) B was the very hungry {word1} animal.
    B was no longer {word2} after a while.
```

Again, the placeholder {word1} and {word2} were instantiated with the two special words (*cannibalistic* or *herbivorous*) and (*hungry* or *alive*) respectively.

We evaluated all the models that support the `chat/completions` endpoint within the OpenAI API [120]. At the time of the experiment on 26 June 2024, a total of 13 models were found to support the endpoint, which includes variants of GPT-3.5 [121], GPT-4 [122], and GPT-4o [123]. To obtain an empirical model, we treat the probabilities outputs by the LLMs as the empirical probabilities of the four contexts. Due to practicality and security concerns, the OpenAI API would only report the probabilities of the top 20 tokens. As there are only 4 valid answers per token, it is unlikely the top 20 tokens do not contain the four valid answers. If this happens, we will assign a zero probability to the missing answers.

Model	Bell-CHSH Violation
`gpt-3.5-turbo`	-0.00
`gpt-3.5-turbo-0125`	-0.00
`gpt-3.5-turbo-1106`	$+0.00$
`gpt-3.5-turbo-16k`	-0.89
gpt-4	$+\mathbf{2.00}$
`gpt-4-0125-preview`	$+0.00$
`gpt-4-0613`	$+\mathbf{2.00}$
`gpt-4-1106-preview`	$+0.00$
`gpt-4-turbo`	$+0.00$
`gpt-4-turbo-2024-04-09`	$+0.00$
`gpt-4-turbo-preview`	$+0.00$
`gpt-4o`	$+0.42$
`gpt-4o-2024-05-13`	$+0.96$

Table 7.2: The violation of Bell-CHSH inequality for the 13 LLMs evaluated. A positive violation indicates contextuality. The maximum violation attainable is 2, which is attained by `gpt-4` and `gpt-4-0613`.

Table 7.2 shows the violation of Bell-CHSH inequality for the 13 LLMs evaluated. The GPT-3 variants exhibited a range of violations, from negative (-0.89) to slight positive ($+0.00$). The GPT-4 variants did not produce any negative violations but most of them have a zero violation. Notably, the more recent GPT-4 variants (`gpt-4` and `gpt-4-0613`) showed strong contextuality, with a maximum violation of 2. The GPT-4o variants, which were designed to be more cost-efficient compared to the GPT-4 variants, exhibited positive violations but not as strong as the GPT-4 variants. Hence, one can conclude that the more advanced models exhibited stronger contextuality. The empirical models can be found in Appendix A.3.

7.4 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 generalised Winograd Schema reminiscent of the Bell-CHSH scenario. Through crowdsourcing, we collected human judgements on an example of the generalised Winograd Schema and observed a contextual empirical model with a significance level of 87%.

We also evaluated the contextuality of the generalised Winograd Schema using LLMs. We observed that LLMs can exhibit contextuality and that more advanced models exhibit stronger contextuality. This provides indications that the generalised Winograd Schema is a promising setting for testing the commonsense reasoning capabilities of LLMs via the contextuality criterion. Currently, one example of the generalised Winograd Schema is created manually. More examples are needed to verify if GenWino can be used as a systematic benchmark for LLM evaluation. It would be interesting to see if LLMs themselves can create more examples of the generalised Winograd Schema and use them to self-assess their commonsense reasoning capabilities.

GenWino 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, potentially paving the way for better language models.

This work presents an approach that consists of deliberately constructing unnatural sentences that exhibit contextuality, which may invite criticism for its contrived nature. Another interesting approach could involve the application of mathematical frameworks designed for contextuality to analyse pre-existing natural language data, moving away from the intentional construction of examples with distinct features [124]. 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.

Part III

Conclusion

Chapter 8

Conclusions and Future Directions

We have demonstrated that human language exhibits contextuality by constructing examples specifically designed to do so. The PR-anaphora schema, which is logically contextual by construction, exhibits probabilistic contextuality when evaluated using an LLM on a set of examples generated for this schema. This work establishes a novel connection between quantum contextuality and natural language, providing a fresh perspective on how ambiguity in language can be modelled and analysed.

Building on this foundation, we generalised the Winograd Schema Challenge to include more than two ambiguous pronouns from the original one-pronoun setting. This extension effectively recreates the well-known Bell-CHSH measurement scenario in a linguistic context. By hand-crafting examples that exhibit contextuality—violating the Bell-CHSH inequality—we demonstrated that both crowd-sourced human evaluators and a range of LLMs exhibit contextual behaviour when resolving these ambiguities. This result highlights the potential of using linguistic contextuality as a tool for studying reasoning and decision-making processes in both humans and artificial systems.

We analysed specific classes of empirical models to better understand their properties. For the PR-anaphora schema, we studied PR-like models which are models that share the same support as the PR-box and derived formulas for various quantities of interest, including the contextual fraction and signalling fraction from the sheaf-theoretic framework, and direct influence from the Contextuality-by-Default framework. This helped clarify the relationship between contextuality and signalling in these models. For the generalised Winograd schema, we looked at outcome-symmetric models and showed that they are non-signalling if the symmetry is described by a transitive permutation group. These results provide a clearer picture of how contextuality operates in linguistic and cognitive systems.

One of the most intriguing observations from this work is the correlation between the degree of contextuality exhibited by the generalised Winograd schema and the size of large language models (LLMs). Since the original Winograd Schema Challenge has already been effectively solved by LLMs, we propose using the generalised Winograd schema as a new benchmark for evaluating LLMs. In this framework, the degree of contextuality serves as a novel performance metric, offering a unique lens through which to assess the reasoning capabilities of these models. Furthermore, we hypothesise that the degree of contextuality is linked to the reasoning capabilities of LLMs, a hypothesis that opens up exciting possibilities for future research in artificial intelligence.

Looking ahead, there are several promising directions for future work. First, expanding the dataset of contextual examples is a critical next step. While this thesis focused on carefully constructed examples, developing automated methods to generate diverse contextual examples could provide a broader foundation for analysis. Exploring other types of linguistic ambiguities, such as lexical, structural, or semantic ambiguities, could further enrich our understanding of contextuality in language. A comprehensive benchmark dataset of contextual linguistic phenomena would also enable standardized evaluation of language models.

Second, refining contextuality as an evaluation metric offers significant potential. Investigating whether the degree of contextuality correlates with performance on other reasoning tasks could validate its utility as a metric. Developing more nuanced contextuality measures that capture different aspects of linguistic understanding and reasoning could provide deeper insights into model capabilities. Comparing contextuality measures with existing evaluation metrics could also reveal complementary strengths and weaknesses, further enhancing our ability to assess language models.

Third, there are several theoretical questions that merit further exploration. Developing more sophisticated mathematical models of linguistic contextuality could better capture the richness and complexity of natural language. Exploring the relationship between contextuality in language and other quantum-like phenomena in cognition and decision-making could uncover deeper connections between these fields. Extending the contextuality framework to model pragmatic aspects of language use, such as conversational implicature or speech acts, could also provide valuable insights.

Finally, these findings may have implications for the development of future language models. Incorporating contextuality into model design could potentially improve their ability to handle ambiguity. Training objectives that take contextuality into account might also help enhance reasoning capabilities.

Beyond language and language models, the ideas in this thesis could be useful in other areas. Studying links between linguistic contextuality and contextuality in other cognitive areas, like decision-making or perception, could help us understand context-dependent behaviour better. Using the mathematical tools for modelling contextuality in other complex systems might also give helpful insights. Additionally, looking into connections between linguistic contextuality and quantum information processing could lead to new ways of creating quantum algorithms for natural language processing tasks.

In conclusion, this thesis has established a novel connection between quantum contextuality and natural language, demonstrating that carefully designed linguistic examples can exhibit contextuality when evaluated by both humans and language models. This finding not only deepens our understanding of the fundamental nature of language processing but also offers new perspectives on evaluating and improving language models. The framework developed here opens up exciting possibilities for future research at the intersection of quantum foundations, linguistics, and artificial intelligence.

Bibliography

[1] Neil Macdonald. Language translation by machine—a report of the first successful trial. *Computers and Automation*, 3, 1954.

[2] Kim Vallée, Pierre-Emmanuel Emeriau, Boris Bourdoncle, Adel Sohbi, Shane Mansfield, and Damian Markham. Corrected bell and noncontextuality inequalities for realistic experiments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2022.

[3] John S. Bell. On the einstein podolsky rosen paradox. *Physics Physique Fizika*, 1:195–200, 1964.

[4] Simon B. Kochen and Ernst P. Specker. The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics*, 17(1):59–87, 1967.

[5] John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23:880–884, 1969.

[6] Alain Aspect, Philippe Grangier, and Gérard Roger. Experimental tests of realistic local theories via bell's theorem. *Physical Review Letters*, 47:460–463, 1981.

[7] Janet Anders and Dan E. Browne. Computational power of correlations. *Physical Review Letters*, 102:050502, 2009.

[8] Robert Raussendorf. Contextuality in measurement-based quantum computation. *Physical Review A*, 88:022322, 2013.

[9] Mark Howard, Joel Wallman, Victor Veitch, and Joseph Emerson. Contextuality supplies the magic for quantum computation. *Nature*, 510:351–355, 2015.

[10] Sergey Bravyi and Alexei Kitaev. Universal quantum computation with ideal clifford gates and noisy ancillas. *Physical Review A*, 71:022316, 2005.

[11] Alexei Yu. Kitaev. Fault-tolerant quantum computation by anyons. *Annals of Physics*, 303:2–30, 2003.

[12] Max Planck. Zur theorie des gesetzes der energieverteilung im normalspektrum. *Verhandlungen der Deutschen Physikalischen Gesellschaft*, 2:237–245, 1900.

[13] Albert Einstein. über einen die erzeugung und verwandlung des lichtes betreffenden heuristischen gesichtspunkt. *Annalen der Physik*, 17:132–148, 1905.

[14] Louis de Broglie. *Recherches sur la Théorie des Quanta*. PhD thesis, Université de Paris, 1924.

[15] Erwin Schrödinger. Quantisierung als eigenwertproblem (erste mitteilung). *Annalen der Physik*, 79:361–376, 1926.

[16] Werner Heisenberg. Über quantentheoretische umdeutung kinematischer und mechanischer beziehungen. *Zeitschrift für Physik*, 33:879–893, 1925.

[17] Werner Heisenberg. Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, 43:172–198, 1927.

[18] Paul A. M. Dirac. *The Principles of Quantum Mechanics*. Oxford University Press, 1st edition, 1930.

[19] John von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Springer, 1932.

[20] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

[22] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[23] W. John Hutchins. The georgetown-ibm experiment demonstrated in january 1954. In Ralph E. Frederking and Kathryn B. Taylor, editors, *Machine Translation: From Real Users to Research. AMTA 2004*, volume 3265 of *Lecture Notes in Computer Science*, pages 198–202, Berlin, Heidelberg, 2004. Springer.

[24] W. John Hutchins. The first public demonstration of machine translation: the georgetown-ibm system, 7th january 1954, March 2006. Accessed: 2025-03-25.

[25] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9:36–45, 1966.

[26] Terry Winograd. Procedures as a representation for data in a computer program for understanding natural language. In *Proceedings of the Fall Joint Computer Conference*, pages 235–253. ACM, 1971.

[27] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[28] Julian Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

[29] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[30] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 48–54, 2003.

[31] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jennifer C. Lai. Class-based n-gram models of natural language. In *Proceedings of the Workshop on Speech and Natural Language*, pages 283–289, 1992.

[32] Jeffrey L. Elman. Finding structure in time. In *Cognitive Science*, volume 14, pages 179–211, 1990.

[33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[34] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Technical Report*, 2018.

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Sylvain Gelly, Luke Zettlemoyer, and Mathilde Caron. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[39] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[40] Warren Weaver. Translation. In W. N. Locke and A. D. Booth, editors, *Machine Translation of Languages: Fourteen Essays*, pages 15–23. Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass., 1955. Originally written in 1949.

[41] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[42] Catherine F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 86–90. Association for Computational Linguistics, 1998.

[43] Hector L. Liu and Praveen Singh. Conceptnet: A practical commonsense reasoning engine. *BT Technology Journal*, 22(4):211–226, 2004.

[44] Jerry R. Hobbs. Resolving pronoun references. *Lingua*, 44:311–338, 1978.

[45] Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[46] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference*

on Principles of Knowledge Representation and Reasoning, pages 552–561. AAAI Press, 2012.

[47] Terry Winograd. Understanding natural language. In *Cognitive Psychology*, pages 1–191. Elsevier, 1972.

[48] Arpit Sharma, Nguyen H. Vo, Somak Aditya, and Chitta Baral. Towards addressing the winograd schema challenge: Building and using a semantic parser and a knowledge hunting module. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1319–1325. IJCAI/AAAI Press, 2015.

[49] Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. A knowledge hunting framework for common sense reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1958, Brussels, Belgium, 2018. Association for Computational Linguistics.

[50] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, 2021.

[51] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[52] Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. The defeat of the winograd schema challenge. *Artificial Intelligence*, 301, 2022.

[53] Hongming Zhang, Xinran Zhao, and Yangqiu Song. Winowhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault,

editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online, 2020. Association for Computational Linguistics.

[54] Weinan He, Canming Huang, Yongmei Liu, and Xiaodan Zhu. 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*, pages 3779–3789, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[55] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47:777–780, May 1935.

[56] David Bohm. *Quantum Theory*. Prentice-Hall, Englewood Cliffs, NJ, 1951.

[57] Max Born. Zur quantenmechanik der stoßvorgänge. *Zeitschrift für Physik*, 37(12):863–867, 1926.

[58] Louis de Broglie. La nouvelle dynamique des quanta. In *Electrons et Photons: Rapports et Discussions du Cinquième Conseil de Physique Solvay*, pages 105–132, Paris, 1928. Gauthier-Villars.

[59] David Bohm. A suggested interpretation of the quantum theory in terms of "hidden" variables i and ii. *Physical Review*, 85(2):166–193, 1952.

[60] Alain Aspect, Philippe Grangier, and Gérard Roger. Experimental realization of einstein-podolsky-rosen-bohm gedankenexperiment: A new violation of bell's inequalities. *Physical Review Letters*, 49:91–94, 1982.

[61] Gregor Weihs, Thomas Jennewein, Christoph Simon, Harald Weinfurter, and Anton Zeilinger. Violation of bell's inequality under strict einstein locality conditions. *Physical Review Letters*, 81:5039–5043, 1998.

[62] M. A. Rowe, D. Kielpinski, V. Meyer, C. A. Sackett, W. M. Itano, C. Monroe, and D. J. Wineland. Experimental violation of a bell's inequality with efficient detection. *Nature*, 409:791–794, 2001.

[63] B. Hensen, H. Bernien, A. E. Dréau, 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. Taminiau, and R. Hanson. Loophole-free bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526:682–686, 2015.

[64] M. Giustina, M. A. Versteegh, S. Wengerowsky, J. Handsteiner, A. Hochrainer, K. Phelan, F. Steinlechner, J. Kofler, J.-Å. Larsson, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, J. Beyer, T. Gerrits, A. E. Lita, L. K. Shalm, S. W. Nam, T. Scheidl, R. Ursin, B. Wittmann, and A. Zeilinger. Significant-loophole-free test of bell's theorem with entangled photons. *Physical Review Letters*, 115:250401, 2015.

[65] A. J. Leggett and A. Garg. Quantum mechanics versus macroscopic realism: Is the flux there when nobody looks? *Physical Review Letters*, 54:857–860, 1985.

[66] A. Palacios-Laloy, F. Mallet, F. Nguyen, P. Bertet, D. Vion, D. Esteve, and A. N. Korotkov. Experimental violation of a bell's inequality in time with weak measurement. *Nature Physics*, 6:442–447, 2010.

[67] Hemant Katiyar, Abhishek Shukla, K. Rama Koteswara Rao, and T. S. Mahesh. Violation of entropic leggett-garg inequality in nuclear spins. *Phys. Rev. A*, 87:052102, May 2013.

[68] Alexander A. Klyachko, M. Ali Can, Sinem Binicioğlu, and Alexander S. Shumovsky. Simple test for hidden variables in spin-1 systems. *Physical Review Letters*, 101(2):1–4, 2008.

[69] R. Lapkiewicz, P. Li, C. Schaeff, S. Ramelow, M. Wieśniak, and A. Zeilinger. Experimental non-classicality of an indivisible quantum system. *Nature*, 474(7352):490–493, 2011.

[70] Mark Um, Xinhua Zhang, Junhua Zhang, Dianmei Deng, Chih-Yuan Chien, Sang-Kyung Choi, and Kihwan Kim. Experimental certification of random numbers via quantum contextuality. *Scientific Reports*, 3:1627, 2013.

[71] Jerome R. Busemeyer and Peter D. Bruza. *Quantum Models of Cognition and Decision*. Cambridge University Press, 2012.

[72] Diederik Aerts, Marek Czachor, and Bart D'Hooghe. Towards a quantum evolutionary scheme: Violating bell's inequalities in language. *Foundations of Science*, 11:345–379, 2006.

[73] Emmanuel M. Pothos and Jerome R. Busemeyer. Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, 36(3):255–274, 2013.

[74] Diederik Aerts. Quantum structure in cognition. *Journal of Mathematical Psychology*, 53(5):314–348, 2009.

[75] Andrei Khrennikov. *Ubiquitous Quantum Structure: From Psychology to Finance*. Springer, 2010.

[76] Ehtibar N. Dzhafarov and Janne V. Kujala. Contextuality is about identity of random variables. *Physica Scripta*, T163:014009, 2014.

[77] Ehtibar N. Dzhafarov and Janne V. Kujala. Context–content systems of random variables: The contextuality-by-default theory. *Journal of Mathematical Psychology*, 74:11–33, 2016.

[78] Daphne Wang, Mehrnoosh Sadrzadeh, Samson Abramsky, and Victor Cervantes. 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), pages 42–52, Groningen, The Netherlands, 2021. Association for Computational Linguistics.

[79] Daphne Wang, Mehrnoosh Sadrzadeh, Samson Abramsky, and Víctor H. Cervantes. Analysing Ambiguous Nouns and Verbs with Quantum Contextuality Tools. *Journal of Cognitive Science*, 22(3):391–420, 2021.

[80] Samson Abramsky and Adam Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics*, 13, Nov 2011.

[81] Arthur Fine. Hidden variables, joint probability, and the bell inequalities. *Physical Review Letters*, 48:291–295, 1982.

[82] Mateus Araújo, Marco Túlio Quintino, Costantino Budroni, Marcelo Terra Cunha, and Adán Cabello. All noncontextuality inequalities for the *n*-cycle scenario. *Physical Review A*, 88:022118, 2013.

[83] Ehtibar N. Dzhafarov, Janne V. Kujala, and Victor H. Cervantes. Contextuality and noncontextuality measures and generalized bell inequalities for cyclic systems. *Physical Review A*, 101(4):042119, 2020.

[84] Itamar Pitowsky. *Quantum Probability, Quantum Logic*, volume 321 of *Lecture Notes in Physics*. Springer, 1989.

[85] Itamar Pitowsky. Correlation polytopes: their geometry and complexity. *Mathematical Programming*, 50(1-3):395–414, 1991.

[86] Itamar Pitowsky. George boole's "conditions of possible experience" and the quantum puzzle. *The British Journal for the Philosophy of Science*, 45(1):95–125, 1994.

[87] Samson Abramsky, Rui Soares Barbosa, and Shane Mansfield. Contextual fraction as a measure of contextuality. *Physical Review Letter*, 119:050504, 2017.

[88] Janne V. Kujala and Ehtibar N. Dzhafarov. Measures of contextuality and noncontextuality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377(2157):20190149, 2019.

[89] Ehtibar N. Dzhafarov, Janne V. Kujala, Víctor H. Cervantes, Ru Zhang, and Matt Jones. On contextuality in behavioural data. *Philosophical Transactions of the Royal Society A*, 374(2058):20150234, 2016.

[90] Janne V Kujala and Ehtibar N Dzhafarov. Proof of a conjecture on contextuality in cyclic systems with binary variables. *Foundations of Physics*, 46:282–299, 3 2016.

[91] Sandu Popescu and Daniel Rohrlich. Quantum nonlocality as an axiom. *Foundations of Physics*, 24:379–385, 1994.

[92] Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield. A model of anaphoric ambiguities using sheaf theoretic quantum-like contextuality and bert. In *Proceedings of the Workshop on End-to-End Compositional Models of Vector-Based Semantics (E2ECOMPVEC)*, pages 23–34. Open Publishing Association, 2022.

[93] Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield. Developments in sheaf-theoretic models of natural language ambiguities. In *Proceedings of the 13th International Workshop on Developments in Computational Models (DCM)*, volume 408, pages 62–72. Open Publishing Association, 2023.

[94] John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

[95] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

[96] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[97] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.

[98] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, 2013.

[99] Philip Resnik. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453, Montreal, 1995.

[100] E. Bruni, G. Boleda, M. Baroni, and N. Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 136–145. Association for Computational Linguistics, 2012.

[101] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.

[102] F. Hill, R. Reichart, and A. Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

[103] Daniela Gerz, Mohammad Taher Pilehvar, Douwe Kiela, and Felix Hill. Simverb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[104] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

[105] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 1999.

[106] Julie Weeds, David Weir, and Diana McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1015–1021. Association for Computational Linguistics, 2004.

[107] Dimitrios Kartsaklis, Sanjaye Ramgoolam, and Mehrnoosh Sadrzadeh. Linguistic matrix theory. *Ann. Inst. Henri Poincare Comb. Phys. Interact*, 6(3):385–426, 2019.

[108] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[109] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[110] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2019.

[111] Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield. Generalised winograd schema and its contextuality. In *Proceedings of the 20th International*

Conference on Quantum Physics and Logic, pages 187–202. Open Publishing Association, 2023.

[112] Samson Abramsky and Mehrnoosh Sadrzadeh. Semantic unification: A sheaf theoretic approach to natural language. In *Categories and Types in Logic, Language, and Physics*, pages 1–13. Springer, 2014.

[113] Ehtibar N. Dzhafarov. The contextuality-by-default view of the sheaf-theoretic approach to contextuality. 2019.

[114] Ehtibar N. Dzhafarov. The contextuality-by-default view of the sheaf-theoretic approach to contextuality. In Alessandra Palmigiano and Mehrnoosh Sadrzadeh, editors, *Samson Abramsky on Logic and Structure in Computer Science and Beyond*, pages 573–593. Springer International Publishing, 2023.

[115] V. H. Cervantes and E. N. Dzhafarov. Snow queen is evil and beautiful: Experimental evidence for probabilistic contextuality in human choices. *Decision*, 5(3):193–204, 2018.

[116] I. Basieva, V. H. Cervantes, E. N. Dzhafarov, and A. Khrennikov. True contextuality beats direct influences in human decision making. *Journal of Experimental Psychology: General*, 148(11):1925–1937, 2019.

[117] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1060–1074, 2023.

[118] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Bäck. Reasoning with large language models: A survey. *arXiv preprint arXiv:2407.11511*, 2024.

[119] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[120] OpenAI. OpenAI API, 2020. Accessed: 2024-11-19.

[121] OpenAI. Introducing ChatGPT, 2022. Accessed: 2024-11-19.

[122] OpenAI. GPT-4, 2023. Accessed: 2024-11-19.

[123] OpenAI. Introducing GPT-4o and more tools to ChatGPT free users, 2024. Accessed: 2024-11-19.

[124] Daphne Wang and Mehrnoosh Sadrzadeh. The causal structure of semantic ambiguities. *arXiv preprint arXiv:2206.06807*, 2022.

Appendix A

Appendix

A.1 Symmetric Group

A permutation of a set X is a bijective function from X to itself. It is common to denote a permutation as $\sigma : X \to X$. We have the following properties:

1. The identity permutation is the permutation that maps every element to itself.
2. The composition of two permutations is also a permutation, since the composition of two bijective functions is also bijective.
3. The inverse of a permutation is also a permutation, since the inverse of a bijective function is also bijective.

These properties satisfy the definition of a group. Hence, the set of all permutations of a set X forms a group under composition, called the *symmetric group* of X, denoted as S_X. The group action of $g \in S_X$ on $x \in X$ is defined as $g \cdot x = g(x)$.

A.1.1 Permutation Group

The symmetric group S_X contains all the possible permutations of a set X. In some cases, we may only focus on a subset of the permutations, depending on the context or specific properties of interest. A permutation group is a subgroup of the symmetric group, $G \subseteq S_X$.

The orbits of $x \in X$ under the action of G is defined as the set of all elements in X that can be reached from x by group action of some element in G. Formally, the orbit of x is

$$\mathrm{Orb}(x) = \{y \in X \mid \exists g \in G, y = g \cdot x\}.$$

A group action is called *transitive* if any two elements $x_1, x_2 \in X$ are reachable from each other, that is, there exists some $g \in G$ such that $g \cdot x_1 = x_2$. Hence, a transitive group action partitions the set X into a single orbit.

A.2 Inequality for signalling fraction and direct influence of PR-like models

Consider a PR-like model on the n-cyclic measurement scenario parameterised by ε_i for $i = 1,2,\ldots,n$. The signalling fraction of such a model is

$$\mathsf{SF} = \max_{i=1}^{n} |\varepsilon_i|, \tag{A.1}$$

while the direct influence is given by

$$\Delta = |\varepsilon_1 - \varepsilon_2| + |\varepsilon_2 - \varepsilon_3| + \cdots + |\varepsilon_{n-1} - \varepsilon_n| + |\varepsilon_n + \varepsilon_1|. \tag{A.2}$$

We would like to prove the following inequality:

$$2\mathsf{SF} \leq \Delta \leq \begin{cases} 2n\mathsf{SF} & n \text{ odd} \\ 2(n-1)\mathsf{SF} & n \text{ even} \end{cases}.$$

Proof. Without loss of generality, we can assume that $\varepsilon_1 = \max_{i=1}^{n} |\varepsilon_i|$. For the ease of notation, we will denote $\varepsilon_{n+1} := -\varepsilon_1$. Then direct influence can be expressed as

$$\Delta = \sum_{i=1}^{n} |\varepsilon_i - \varepsilon_{i+1}|. \tag{A.3}$$

Now since the sum of absolute values is greater than or equal to the absolute value of the sum, we have the following for the lower bound:

$$\Delta = \sum_{i=1}^{n} |\varepsilon_i - \varepsilon_{i+1}| \geq \left| \sum_{i=1}^{n} (\varepsilon_i - \varepsilon_{i+1}) \right| = |\varepsilon_1 - \varepsilon_{n+1}| = |\varepsilon_1 - (-\varepsilon_1)| = 2|\varepsilon_1| = 2\mathsf{SF}. \tag{A.4}$$

Here we used the telescopic property of the sum.

Now we will show the upper bound. If n is even, the alternating sign choice $\varepsilon_i = \mathsf{SF}$ for i odd and $\varepsilon_i = -\mathsf{SF}$ for i even gives each term in the direct influence as $2\mathsf{SF}$, except for the last term, which is 0. Hence, $\Delta = (n-1)2\mathsf{SF} = 2(n-1)\mathsf{SF}$. If n

is odd, the alternating sign choice $\varepsilon_i = \mathsf{SF}$ for i even and $\varepsilon_i = -\mathsf{SF}$ for i odd gives each term, including the last term, as $2\mathsf{SF}$. Hence, $\Delta = n2\mathsf{SF} = 2n\mathsf{SF}$. $\square$

A.3 LLM Empirical Models for GenWino

`gpt-3.5-turbo`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.00000	0.50000	0.50000	0.00000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.00000	0.50000	0.50000	0.00000
herbivorous	alive	0.00000	0.50000	0.50000	0.00000

`gpt-3.5-turbo-0125`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.00000	0.50000	0.50000	0.00000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.00000	0.50000	0.50000	0.00000
herbivorous	alive	0.00000	0.50000	0.50000	0.00000

`gpt-3.5-turbo-1106`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.00069	0.49931	0.49931	0.00069
cannibalistic	alive	0.00001	0.49999	0.49999	0.00001
herbivorous	hungry	0.00013	0.49987	0.49987	0.00013
herbivorous	alive	0.00000	0.50000	0.50000	0.00000

`gpt-3.5-turbo-16k`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.38770	0.11230	0.11230	0.38770
cannibalistic	alive	0.27934	0.22066	0.22066	0.27934
herbivorous	hungry	0.34671	0.15329	0.15329	0.34671
herbivorous	alive	0.32351	0.17649	0.17649	0.32351

`gpt-4`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.50000	0.00000	0.00000	0.50000
herbivorous	alive	0.49999	0.00001	0.00001	0.49999

`gpt-4-0125-preview`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.50000	0.00000	0.00000	0.50000
herbivorous	alive	0.00000	0.50000	0.50000	0.00000

`gpt-4-0613`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.50000	0.00000	0.00000	0.50000
herbivorous	alive	0.50000	0.00000	0.00000	0.50000

`gpt-4-1106-preview`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.49999	0.00001	0.00001	0.49999
herbivorous	alive	0.00008	0.49992	0.49992	0.00008

`gpt-4-turbo`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.50000	0.00000	0.00000	0.50000
herbivorous	alive	0.00000	0.50000	0.50000	0.00000

`gpt-4-turbo-2024-04-09`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.50000	0.00000	0.00000	0.50000
herbivorous	alive	0.00000	0.50000	0.50000	0.00000

`gpt-4-turbo-preview`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00000	0.50000	0.50000	0.00000
herbivorous	hungry	0.50000	0.00000	0.00000	0.50000
herbivorous	alive	0.00000	0.50000	0.50000	0.00000

`gpt-4o`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00002	0.49998	0.49998	0.00002
herbivorous	hungry	0.50000	0.00000	0.00000	0.50000
herbivorous	alive	0.10434	0.39566	0.39566	0.10434

`gpt-4o-2024-05-13`

word1	word2	(A, A)	(A, B)	(B, A)	(B, B)
cannibalistic	hungry	0.50000	0.00000	0.00000	0.50000
cannibalistic	alive	0.00001	0.49999	0.49999	0.00001
herbivorous	hungry	0.50000	0.00000	0.00000	0.50000
herbivorous	alive	0.23956	0.26044	0.26044	0.23956