Neural Reasoning for Sure Through Constructing Explainable Models

Tiansi Dong, Mateja Jamnik, Pietro Liò

Department of Computer Science and Technology, University of Cambridge
15 JJ Thomson Ave, Cambridge, UK
{td540, mj201, pl219}@cam.ac.uk

Abstract

Neural networks remain black-box systems, *unsure* about their outputs, and their performance may drop unpredictably in real applications. An open question is how to qualitatively extend neural networks, so that they are *sure* about their reasoning results, or *reasoning-for-sure*. Here, we introduce set-theoretic relations explicitly and seamlessly into neural networks by extending vector embedding into sphere embedding, so that part-whole relations can explicitly encode set-theoretic relations through sphere boundaries in the vector space. A *reasoning-for-sure* neural network successfully constructs, within a constant number M of epochs, a sphere configuration as its semantic model for any consistent set-theoretic relation. We implement Hyperbolic Sphere Neural Network (HSphNN), the first *reasoning-for-sure* neural network for all types of Aristotelian syllogistic reasoning. Its construction process is realised as a sequence of neighbourhood transitions from the current towards the target configuration. We prove $M = 1$ for HSphNN. In experiments, HSphNN achieves the symbolic level rigour of syllogistic reasoning and successfully checks both decisions and explanations of ChatGPT (gpt-3.5-turbo and gpt-4o) without errors. Through prompts, HSphNN improves the performance of gpt-3.5-turbo from 46.875% to 58.98%, and of gpt-4o from 82.42% to 84.76%. We show ways to extend HSphNN for various kinds of logical and Bayesian reasoning, and to integrate it with traditional neural networks seamlessly.

Code and Data — https://github.com/gnodisnait/hsphnn

Introduction

Besides being capable of human-like communication (Biever 2023), LLMs are successful in various reasoning tasks, such as mathematical findings (Davies et al. 2021), predicting protein structures (Abramson et al. 2024), and playing games (Silver et al. 2017; Schrittwieser et al. 2020). However, being able to communicate does not necessarily result in the ability to reason (Fedorenko, Piantadosi, and Gibson 2024). Though breaking complex tasks into multiple steps may improve reasoning performance (Creswell, Shanahan, and Higgins 2023; Wei et al. 2022; Lightman et al. 2023), it remains unclear whether LLMs really reason (Biever 2023; Mitchell 2023). They exhibit unpredictable behaviour (Park et al. 2024), errors in abstract reasoning (Eisape et al. 2024; Lampinen et al. 2024), and the irrationality of producing correct answers with incorrect explanations (Creswell, Shanahan, and Higgins 2023; Zelikman et al. 2022). Neural networks, including LLMs, are still not sure about their reasoning results. Here, we define the *reasoning-for-sure* property of neural networks as follows: *a neural network reasons for sure, if for consistent set-theoretic statements, it can successfully construct a model in the vector space, within M epochs, where M is a constant number.* We shall prove M's existence and the value of M without using data. This M enables neural networks to determine inconsistent statements: after M epochs, if the target model is not constructed, the neural network will conclude that there is no model and the conclusion is unsatisfiable.

Missing *reasoning-for-sure* property, LLMs may make wrong decisions in legal judgments (Deng et al. 2023), in clinical services (Hager et al. 2024), and bring unpredictable risks to our society (Bengio et al. 2024). The ability of syllogistic deduction is the pre-condition for inference and other types of rational reasoning. Legal judgments have a syllogistic backbone, where rules of laws serve as the major premise and a concrete case is the minor premise. Clinical decision-making is the reverse process that tries to infer which disease causes (deduces) the symptoms. Determining the consistency is a pre-condition to making correct deductions (Bucciarelli, Khemlani, and Johnson-Laird 2008; Johnson-Laird, Khemlani, and Goodwin 2015). For example, from *'All Greeks are human. All humans are mortal.'*, we can deduce *for sure* both that *'All Greeks are mortal'* and also that *'Some Greeks are mortal'*. However, this is a big challenge for supervised deep learning, partly because training data for one conclusion weakens the learning confidence of the other conclusion. The performance of LLMs on syllogistic reasoning has been systematically evaluated (Eisape et al. 2024; Lampinen et al. 2024): they inevitably mimic human errors in the training data. Cutting-edge neural research is to move from simulating fast associative thinking (System 1) to achieving slow *logical* thinking (System 2) (Kahneman 2011) for high-level cognitive tasks (Goyal and Bengio 2022). The first step is to develop *reasoning-for-sure* neural networks for Aristotelian syllogistic statements, the microcosm of human rationality (Khemlani and Johnson-Laird 2012; Khemlani 2021).

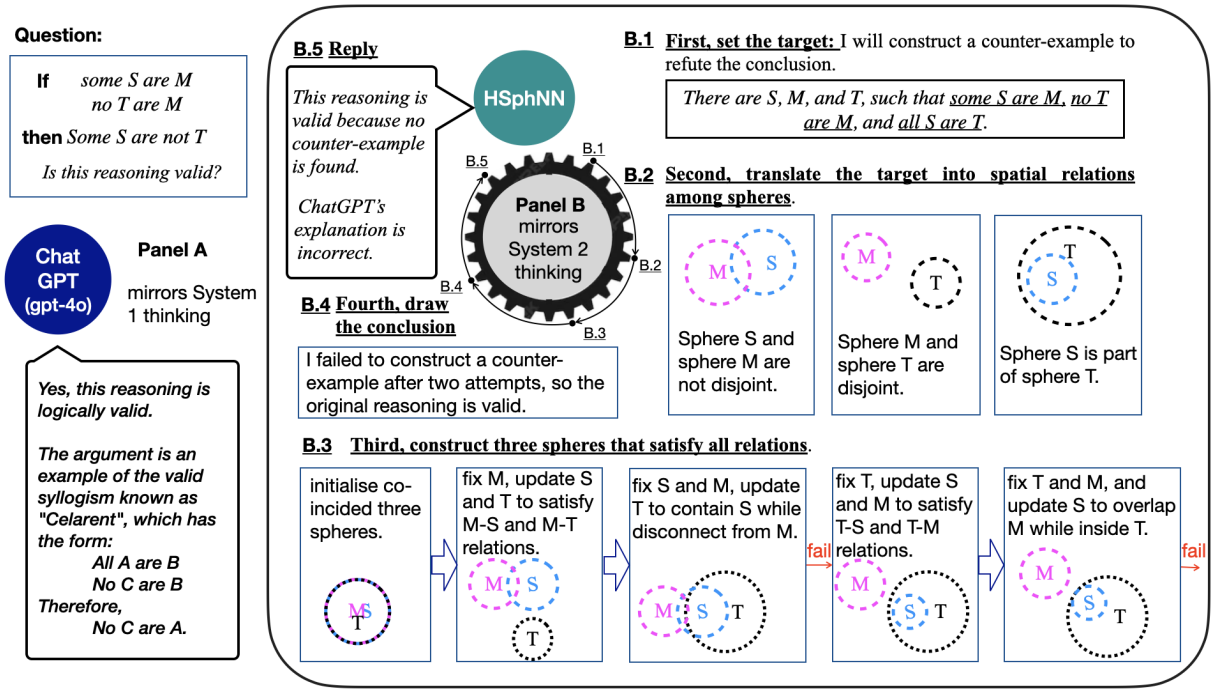

Figure 1: Panel A: Given a syllogistic reasoning task, ChatGPT replied with a correct decision with an incorrect explanation. Panel B illustrates how HSphNN mirrors System 2 rigorous logical reasoning by constructing models. To answer whether a reasoning is always true, HSphNN tries to refute it by finding a counter-example. (B.1) It sets the configuration of a counter-example; (B.2) Then, it translates the configuration into spatial relations among spheres; (B.3) HSphNN follows a well-designed procedure to construct the target configuration. This procedure guarantees that HSphNN can find the configuration if it exists. (B.4) If HSphNN fails to find the target configuration, it concludes with no counter-examples. Finally, it replies the original reasoning is valid. Using the same procedure, HSphNN can check whether ChatGPT's explanation is correct.

Syllogistic reasoning appears deceptively simple but has dominated logic research for over two thousand years and psychological research on human rationality for over 100 years till today (Khemlani and Johnson-Laird 2012; Johnson-Laird, Byrne, and Khemlani 2024). A recent analysis doubted whether traditional methods hit a dead end (Brand, Riesterer, and Ragni 2023). Developing neural syllogistic reasoning was once believed as Utopian (Khemlani and Johnson-Laird 2012). Only recently were supervised neural networks developed to approximate a substantial part of syllogistic reasoning (Wang, Jamnik, and Liò 2018, 2020), in part because vector embedding cannot explicitly represent set-theoretic relations.

Venn diagrams were used in the set-diagram network architecture to explicitly represent set-theoretic relations (Rosenblatt 1962). In recent years, regions in Euclidean space have been used in neural reasoning (Lv et al. 2018; Dong et al. 2019; Ren, Hu, and Leskovec 2020; Zhang et al. 2021), because regions can easily represent part-whole relations that exist in all material and conceptual domains, and are used in all disciplines (Smith 1996).

Compared to Euclidean, hyperbolic geometry better computationally simulates our vision reasoning systems (Fang et al. 2023). Here, we design a *reasoning-for-sure* network, Hyperbolic Sphere Neural-Networks (HSphNN), that reasons with Aristotelian syllogistic statements *for sure* with $M = 1$. That is, to determine the validity of the reasoning *'some S are M, no T are M; ∴ some S are not T'*, HSphNN tries to construct a counter-example in the first epoch and fails. With $M = 1$, it stops iteration and concludes that the reasoning is valid. In contrast, ChatGPT may make a correct conclusion with a wrong explanation, as shown in Figure 1.

The main contributions are as follows:

1. We define the property of *reasoning-for-sure* for neural networks.
2. We introduce into neural networks the methodology of *reasoning as explicit model construction* (Johnson-Laird and Byrne 1991; Goodwin and Johnson-Laird 2005; Knauff 2009), which allows neural networks to reason without training data, thus, go beyond the statistical paradigm (Goyal and Bengio 2022; Gigerenzer 2022).
3. We develop HSphNN, the first neural network that achieves syllogistic reasoning *for sure* without training data. Our experiments show that HSphNN can determine the validity and the satisfiability of all Aristotelian syllogistic reasoning, it can successfully identify all hallucinations of ChatGPT gpt-3.5-turbo and gpt-4o, and it can improve their performance significantly (as long as ChatGPT does not ignore prompt feedbacks).
4. We show how to extend HSphNN to other types of rational reasoning, and how to integrate it with traditional neural networks seamlessly.

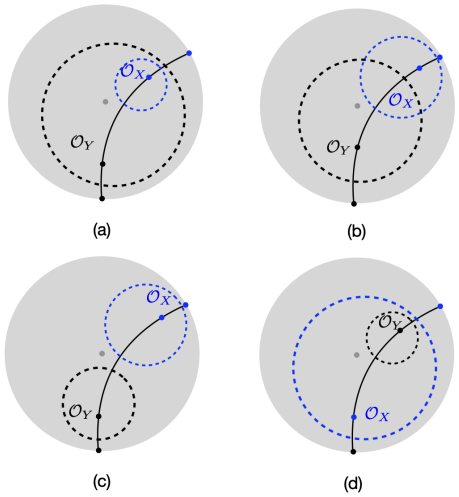

Figure 2: Four syllogistic relations are interpreted as four spatial relations between Poincaré spheres. The black curve is the Geodesic passing the Poincaré centres of $\mathcal{O}_X$ and $\mathcal{O}_Y$.

Spatialising Syllogistic Statements

Syllogistic statements only have four forms, each being a part-whole relation between sets (Hammer and Shin 1998), and can be represented by Euler diagrams (Wang, Jamnik, and Liò 2018). We propose a one-to-one mapping between syllogistic statements and spatial relations between Poincaré spheres (Nickel and Kiela 2017) (section 1 in the supplementary document), and define these spatial relations using the part-whole relation, as shown in Figure 2 and as follows:

1. *All X are Y*, or $all(X, Y)$ for short, iff $\mathcal{O}_X$ is part of $\mathcal{O}_Y$, $\mathbf{P}(\mathcal{O}_X, \mathcal{O}_Y)$, where $\mathcal{O}_X, \mathcal{O}_Y$ are Poincaré spheres (Figure 2(a)).
2. *Some X are Y*, or $some(X, Y)$ for short, iff there is $\mathcal{O}_Z$ that is part of $\mathcal{O}_X$ and $\mathcal{O}_Y$, $\exists \mathcal{O}_Z \mathbf{P}(\mathcal{O}_Z, \mathcal{O}_X) \wedge \mathbf{P}(\mathcal{O}_Z, \mathcal{O}_Y)$, or $\neg\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Y)$ (Figure 2(b)).
3. *No X are Y*, or $no(X, Y)$ for short, iff there is no $\mathcal{O}_Z$ that is part of $\mathcal{O}_X$ and $\mathcal{O}_Y$, $\nexists \mathcal{O}_Z \mathbf{P}(\mathcal{O}_Z, \mathcal{O}_X) \wedge \mathbf{P}(\mathcal{O}_Z, \mathcal{O}_Y)$, or $\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Y)$ for short (Figure 2(c)).
4. *Some X are not Y*, or $some\_not(X, Y)$ for short, iff $\mathcal{O}_X$ is not part of $\mathcal{O}_Y$, $\neg\mathbf{P}(\mathcal{O}_X, \mathcal{O}_Y)$ (Figure 2(d)).

We define the mapping function ψ: (i) $\psi(all) = \mathbf{P}$; (ii) $\psi(some) = \neg\mathbf{D}$; (iii) $\psi(no) = \mathbf{D}$; (iv) $\psi(some\_not) = \neg\mathbf{P}$. We introduce *$Y$ contains X*, $\psi(contain) = \overline{\mathbf{P}}$, which is equivalent to *all X are Y*. With this ψ function, HSphNN translates a symbolic syllogistic statement into a spatial statement between Poincaré spheres. Let $\vec{O}$, l, and r be the Euclidean centre, the Euclidean norm of $\vec{O}$, and the Euclidean radius of a Poincaré sphere $\mathcal{O}$ in an n-dimensional Poincaré disk. $\mathbb{D}(r)$ is the Poincaré radius of $\mathcal{O}$:

$$\mathbb{D}(r) = \tanh^{-1}(r + l) - \tanh^{-1}(l - r).$$

Let $\vec{O}_1^*$ and $\vec{O}_2^*$ be Poincaré centres of $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively. The Poincaré distance between them is $d_{\mathbb{D}}(\vec{O}_1^*, \vec{O}_2^*)$:

$$d_{\mathbb{D}}(\vec{O}_1^*, \vec{O}_2^*) = \cosh^{-1}(1 + 2\frac{\|\vec{O}_1^* - \vec{O}_2^*\|^2}{(1 - \|\vec{O}_1^*\|^2) \cdot (1 - \|\vec{O}_2^*\|^2)})$$

where $\vec{O}_i^* = \tanh(\frac{\tanh^{-1}(r_i + l_i) + \tanh^{-1}(l_i - r_i)}{2})\frac{\vec{O}_i}{\|\vec{O}_i\|}$, $i = 1, 2$. We define Poincaré spheres as open and define relations in Poincaré metrics as follows: $\mathcal{O}_X$ is part of $\mathcal{O}_Y$, $\mathbf{P}(\mathcal{O}_X, \mathcal{O}_Y)$, if and only if the distance between their centres plus $\mathcal{O}_X$'s radius is less than or equals to $\mathcal{O}_Y$'s radius:

$$d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*) + \mathbb{D}(r_X) \leq \mathbb{D}(r_Y).$$

To transform $\mathcal{O}_X$ inside $\mathcal{O}_Y$, we shall decrease $d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*)$ or $\mathbb{D}(r_X)$, or increase $\mathbb{D}(r_Y)$. We introduce the inspection function $\mathcal{I}^{\mathbf{P}}(\mathcal{O}_X, \mathcal{O}_Y)$ as follows:

$$\mathcal{I}^{\mathbf{P}}(\mathcal{O}_X, \mathcal{O}_Y) \triangleq \max\{0, d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*) + \mathbb{D}(r_X) - \mathbb{D}(r_Y)\}.$$

The target $\mathbf{P}(\mathcal{O}_X, \mathcal{O}_Y)$ is achieved, if and only if $\mathcal{I}^{\mathbf{P}}(\mathcal{O}_X, \mathcal{O}_Y) = 0$, otherwise, $\mathcal{I}^{\mathbf{P}}(\mathcal{O}_X, \mathcal{O}_Y) > 0$.

$\mathcal{O}_X$ disconnects from $\mathcal{O}_Y$, $\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Y)$, if and only if the distance between their centres is greater than or equals to the sum of their radii:

$$d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*) \geq \mathbb{D}(r_X) + \mathbb{D}(r_Y).$$

To reach $\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Y)$, we shall either increase $d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*)$ or decrease $\mathbb{D}(r_X)$ or $\mathbb{D}(r_Y)$. The inspection function $\mathcal{I}^{\mathbf{D}}(\mathcal{O}_X, \mathcal{O}_Y)$ is designed as follows:

$$\mathcal{I}^{\mathbf{D}}(\mathcal{O}_X, \mathcal{O}_Y) = \max\{0, \mathbb{D}(r_X) + \mathbb{D}(r_Y) - d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*)\}.$$

The target relation $\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Y)$ is reached, if and only if $\mathcal{I}^{\mathbf{D}}(\mathcal{O}_X, \mathcal{O}_Y) = 0$, otherwise, $\mathcal{I}^{\mathbf{D}}(\mathcal{O}_X, \mathcal{O}_Y) > 0$.

$\mathcal{O}_X$ partially overlaps with $\mathcal{O}_Y$, $\mathbf{PO}(\mathcal{O}_X, \mathcal{O}_Y)$, if and only if the distance between their centres is greater than the absolute difference between their radii and less than the sum of their radii:

$$|\mathbb{D}(r_X) - \mathbb{D}(r_Y)| < d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*) < \mathbb{D}(r_X) + \mathbb{D}(r_Y).$$

Its inspection function $\mathcal{I}^{\mathbf{PO}}(\mathcal{O}_X, \mathcal{O}_Y)$ is the sum of $\max\{0, |\mathbb{D}(r_X) - \mathbb{D}(r_Y)| - d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*)\}$ and $\max\{0, d_{\mathbb{D}}(\vec{O}_X^*, \vec{O}_Y^*) - \mathbb{D}(r_X) - \mathbb{D}(r_Y)\}$.

The *Reasoning-for-Sure* Property of Neural Syllogistic Reasoning

We define the *reasoning-for-sure* property in the setting of neural syllogistic reasoning through constructing configurations of Poincaré spheres as follows:

> *For any* satisfiable *syllogistic reasoning, a* reasoning-for-sure *neural network can construct a Poincaré sphere configuration as its Euler diagram at the global loss of zero within M epochs, where M is a constant.*

With this property, HSphNN achieves the symbolic level of syllogistic reasoning as follows:

- A syllogistic reasoning statement (e.g., '$r_1(X, Y)$, $r_2(Y, Z) \therefore r_3(X, Z)$') is *satisfiable*, if and only if HSphNN constructs a configuration of Poincaré spheres satisfying $\psi(r_1)(\mathcal{O}_X, \mathcal{O}_Y)$, $\psi(r_2)(\mathcal{O}_Y, \mathcal{O}_Z)$, and $\psi(r_3)(\mathcal{O}_X, \mathcal{O}_Z)$ within M epochs.

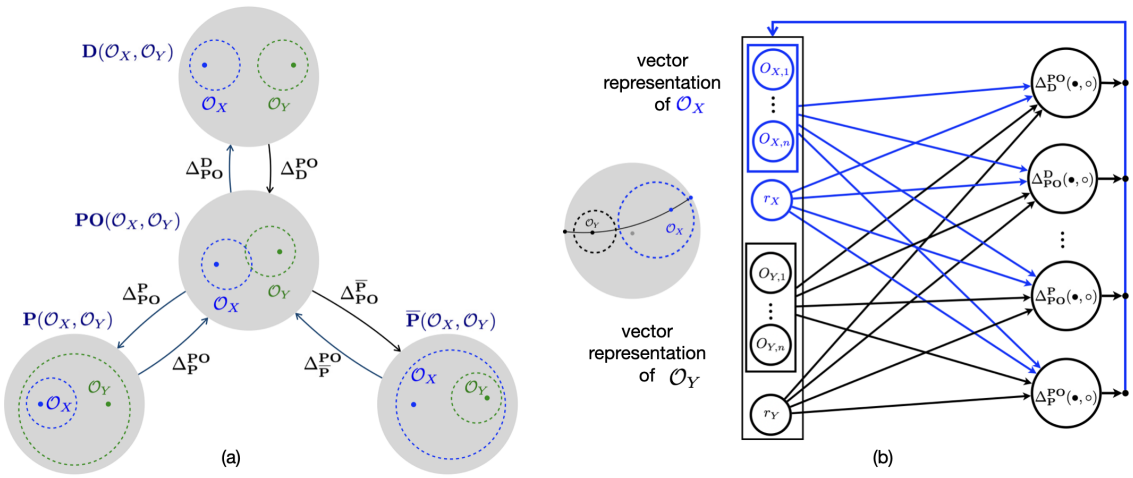

Figure 3: (a) The basic architecture of HSphNN is an RNN of neighbourhood spatial relations. (b) The input of a relation node is two $(n+1)$-dimensional vectors. Each represents an n-dimensional sphere, whose first n components represent the centre, and the $(n+1)^{th}$ component represents the radius. Edges are directed, and each is a Δ function responsible for a neighbourhood transition. When the Δ function returns 0, the two spheres have reached the neighbourhood relation. The blue colour in (b) means that $\mathcal{O}_X$ is dynamic and $\mathcal{O}_Y$ is fixed.

- A syllogistic reasoning statement is *valid*, if and only if the conclusion is true, whenever the premises are true (Jeffrey 1981). This is equivalent to the negation of the conclusion being *false*, whenever the premises are true. That is, if HSphNN cannot construct a configuration of Poincaré spheres within M epochs, which satisfies $\psi(r_1)(\mathcal{O}_X, \mathcal{O}_Y)$, $\psi(r_2)(\mathcal{O}_Y, \mathcal{O}_Z)$, and $\neg\psi(r_3)(\mathcal{O}_X, \mathcal{O}_Z)$, it will conclude that the original syllogistic reasoning is valid.

Without the value of M, if HSphNN fails to reach the global loss of zero in the current epoch, then it cannot conclude *for sure* that a syllogistic reasoning statement is *unsatisfiable*.

Readers familiar with Euclidean geometry may refer to Sphere Neural Networks (SphNN) in Euclidean space, the twin of HSphNN. We prove $M = 1$ for both HSphNN and SphNN (Dong, Jamnik, and Liò 2024).

Neural Reasoning As Neighbourhood Transitions of Poincaré Sphere Configurations

Following Herbert A. Simon (Simon 2019), reasoning, as a typical problem-solving task, can be viewed as navigation in a graph – each node represents a situation, and each edge represents an action that transforms one node into its neighbourhood till the target node is reached. For example, if $\mathcal{O}_X$ currently disconnects from $\mathcal{O}_Y$, through repeated decreasing of the distance between their centres or increasing of the radius of $\mathcal{O}_X$, they will be partially overlapped. We define an action function $\Delta^{\mathbf{PO}}_{\mathbf{D}}(\mathcal{O}_X, \mathcal{O}_Y) \triangleq \max\{0, d_{\mathbb{D}}(\vec{O}^*_X, \vec{O}^*_Y) - \mathbb{D}(r_X) - \mathbb{D}(r_Y)\}$. HSphNN gradually reduces the value of $\Delta^{\mathbf{PO}}_{\mathbf{D}}(\mathcal{O}_X, \mathcal{O}_Y)$. When the value reaches 0, it arrives at the neighbourhood node. Each action function has the form $\Delta^{\mathbf{T}}_{\mathbf{T}_1:\mathbf{T}_2}(\mathcal{O}_X, \mathcal{O}_Y)$ that transforms the relation between $\mathcal{O}_X$ and $\mathcal{O}_Y$ from $\mathbf{T}_1$ to its neighbourhood relation $\mathbf{T}_2$ towards the target relation $\mathbf{T}$, where $\mathcal{O}_Y$ is fixed. We design each $\Delta^{\mathbf{T}}_{\mathbf{T}_1:\mathbf{T}_2}(\mathcal{O}_X, \mathcal{O}_Y)$ to satisfy three conditions:

1. non-negative, $\Delta^{\mathbf{T}}_{\mathbf{T}_1:\mathbf{T}_2}(\mathcal{O}_X, \mathcal{O}_Y) \geq 0$;
2. strict monotonous, when $\Delta^{\mathbf{T}}_{\mathbf{T}_1:\mathbf{T}_2}(\mathcal{O}_X, \mathcal{O}_Y) > 0$;
3. if the target relation is reached, $\Delta^{\mathbf{T}}_{\mathbf{T}_1:\mathbf{T}_2}(\mathcal{O}_X, \mathcal{O}_Y) = 0$.

HSphNN has an RNN architecture and repeatedly updates locations and sizes of Poincaré spheres using a neighbourhood transition map, as illustrated in Figure 3. Thus, training data is not required. Without constraints, HSphNN can realise every neighbourhood transition and, thus, has constant architectural complexity (the length of the longest path is a constant number) (Zhang et al. 2016). HSphNN can transform three spheres to satisfy two relations in one epoch (Theorem 1, in the supplementary document, page 22).

Neighbourhood Transformation With Constraints HSphNN may be stuck in non-zero local minimums. For example, let the target configuration be $\mathbf{P}(\mathcal{O}_Y, \mathcal{O}_X)$, $\mathbf{P}(\mathcal{O}_Z, \mathcal{O}_X)$, and $\mathbf{D}(\mathcal{O}_Z, \mathcal{O}_Y)$, and currently $\mathcal{O}_Y$ be part of $\mathcal{O}_X$, and $\mathcal{O}_Z$ be disconnected from $\mathcal{O}_X$, as shown in Figure 4(a). When $\mathcal{O}_Z$ is getting closer to $\mathcal{O}_X$ following the Geodesic (the shortest path to $\mathcal{O}_Y$ in a Poincaré disk), it will overlap with $\mathcal{O}_Y$. This will increase the global loss and be stuck in a non-zero local minimum, shown in Figure 4(b). Then, $\mathcal{O}_Z$ needs to circumvent $\mathcal{O}_Y$ by rotating around $\mathcal{O}_Y$, as illustrated in Figure 4(c, d). Rotating $\mathcal{O}_Z$ around $\mathcal{O}_Y$ keeps the relation with $\mathcal{O}_Y$ and allows the search for a new location to improve the relation with $\mathcal{O}_X$. We introduce constraint optimisation $COP^{\mathbf{T}_{ZX}}_{\mathbf{T}_{ZY}}(\mathcal{O}_Z|\mathcal{O}_X; \mathcal{O}_Y)$ that optimises $\mathcal{O}_Z$ to satisfy the relation with $\mathcal{O}_X$ while preserving its relation with $\mathcal{O}_Y$. It works as follows: given $\mathcal{O}_Z$, $\mathcal{O}_X$, and $\mathcal{O}_Y$, where $\mathcal{O}_X$ and $\mathcal{O}_Y$ are fixed, HSphNN gradually reduces the value of $\Delta^{\mathbf{T}_{ZX}}_{\mathbf{S}_{ZX}}(\mathcal{O}_Z, \mathcal{O}_X) + \Delta^{\mathbf{T}_{ZY}}_{\mathbf{S}_{ZY}}(\mathcal{O}_Z, \mathcal{O}_Y)$,

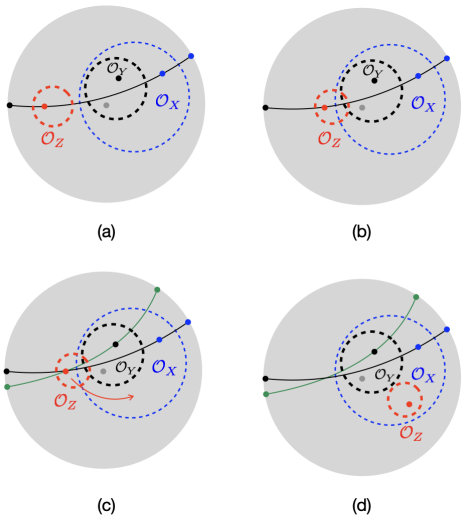

Figure 4: (a) $\mathcal{O}_Z$ is disconnected from $\mathcal{O}_X$; (b) when $\mathcal{O}_Z$ is approaching $\mathcal{O}_X$, a non-zero local minimum will be reached; (c) $\mathcal{O}_Z$ rotates around $\mathcal{O}_Y$; (d) $\mathcal{O}_Z$ successfully reached a target state.

where $\mathbf{S}_{ZY}$ and $\mathbf{S}_{ZX}$ are current relations of $\mathcal{O}_Z$ to $\mathcal{O}_Y$ and to $\mathcal{O}_X$, respectively. After each gradual descent step, HSphNN checks whether $\mathbf{T}_{ZY}(\mathcal{O}_Z, \mathcal{O}_Y)$ is broken. HSphNN will first repair the broken relation before continuing, as described in Algorithm 1. We prove that $COP^{\mathbf{T}_{ZX}}_{\mathbf{T}_{ZY}}(\mathcal{O}_Z|\mathcal{O}_X;\mathcal{O}_Y)$ is a gradual descent (Theorem 2, in the supplementary document, page 36-37).

Realising Negative Relations A negative target relation (e.g., $\neg\mathbf{D}$, $\neg\mathbf{P}$) may have an unintended realisation. For example, let the target be $\neg\mathbf{P}(\mathcal{O}_X, \mathcal{O}_Y)$, $\mathbf{P}(\mathcal{O}_Y, \mathcal{O}_Z)$, and $\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Z)$. If $\mathcal{O}_X$ and $\mathcal{O}_Y$ are realised as being partially overlapped, $\mathbf{PO}(\mathcal{O}_X, \mathcal{O}_Y)$, $\mathcal{O}_Z$ cannot be optimised to satisfy both relations with $\mathcal{O}_X$ and with $\mathcal{O}_Y$, as illustrated in Figure 5(a, b), because $\mathbf{PO}(\mathcal{O}_X, \mathcal{O}_Y)$ is inconsistent with the other two relations. Hence, the negation of this unintended realisation ($\neg\mathbf{PO}(\mathcal{O}_X, \mathcal{O}_Y)$) is a valid conclusion from the other two relations. Therefore, starting with realising the other two relations will avoid this unintended relation. This suggests that the process shall be restarted by fixing a different sphere. In the example, we restart with fixing $\mathcal{O}_Z$ and updating $\mathcal{O}_X$ to disconnect from $\mathcal{O}_Z$, and $\mathcal{O}_Y$ to be part of $\mathcal{O}_Z$, thus $\mathcal{O}_X$ will disconnect from $\mathcal{O}_Y$, as illustrated in Figure 5(c). We prove that restarting once is sufficient to construct a target configuration if the target exists (Theorem 4, on page 44 of the supplementary document), thus, $M = 1$.

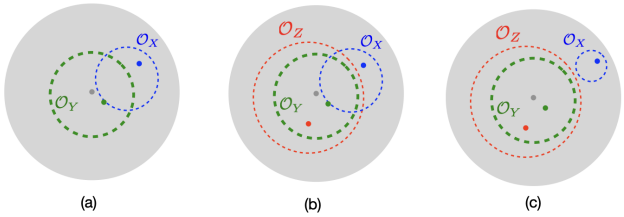

Figure 5: (a) $\neg\mathbf{P}(\mathcal{O}_X, \mathcal{O}_Y)$ is incorrectly realised as $\mathbf{PO}(\mathcal{O}_X, \mathcal{O}_Y)$; (b) no $\mathcal{O}_Z$ can satisfy both $\mathbf{P}(\mathcal{O}_Y, \mathcal{O}_Z)$ and $\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Z)$; (c) HSphNN fixes $\mathcal{O}_Z$ first, then updates $\mathcal{O}_X$ and $\mathcal{O}_Y$ to satisfy $\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Z)$ and $\mathbf{P}(\mathcal{O}_Y, \mathcal{O}_Z)$. Then, $\neg\mathbf{P}(\mathcal{O}_X, \mathcal{O}_Y)$ is correctly realised as $\mathbf{D}(\mathcal{O}_X, \mathcal{O}_Y)$.

***Reasoning-for-Sure* via Explicit Model Construction** A syllogistic reasoning statement '$p_1, p_2 \therefore q$' can be translated into three spatial statements $\psi_1(\mathcal{O}_1, \mathcal{O}_2), \psi_2(\mathcal{O}_2, \mathcal{O}_3) \therefore \psi_3(\mathcal{O}_3, \mathcal{O}_1)$, where $\psi_i \in \{\mathbf{D}, \mathbf{P}, \overline{\mathbf{P}}, \neg\mathbf{D}, \neg\mathbf{P}, \neg\overline{\mathbf{P}}\}$. For example, '*Some X_2 are not X_1. All X_2 are X_3. $\therefore$ Some X_1 are not X_3.*' is translated to $\neg\mathbf{P}(\mathcal{O}_1, \mathcal{O}_2)$, $\mathbf{P}(\mathcal{O}_2, \mathcal{O}_3)$, and $\neg\mathbf{P}(\mathcal{O}_1, \mathcal{O}_3)$. HSphNN initialises three coincided spheres $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{O}_3$; then, it fixes $\mathcal{O}_1$ and updates $\mathcal{O}_2$ and $\mathcal{O}_3$ to satisfy the relations with $\mathcal{O}_1$; thirdly, it fixes $\mathcal{O}_1$ and $\mathcal{O}_2$, and updates $\mathcal{O}_3$ to the relation with $\mathcal{O}_2$ while keeping its relation with $\mathcal{O}_1$. If the target is not achieved, it will fix $\mathcal{O}_3$ and update $\mathcal{O}_1$ and $\mathcal{O}_2$ to satisfy their relations with $\mathcal{O}_3$. Finally, it fixes $\mathcal{O}_3$ and $\mathcal{O}_2$, and updates $\mathcal{O}_1$ to satisfy the relation with $\mathcal{O}_2$ while keeping the relation with $\mathcal{O}_3$. If the target is still not achieved, HSphNN will conclude the original reasoning is unsatisfiable, because $M = 1$. The process is outlined in Algorithm 2.

Algorithm 1: Constraint optimisation.

Input: $\mathcal{O}_X, \mathcal{O}_Y, \mathcal{O}_Z, \mathbf{T}_{ZX}, \mathbf{T}_{ZY}$
Output: $\mathcal{O}_Z$

Optimise $\mathcal{O}_Z$ to satisfy $\mathbf{T}_{ZY}(\mathcal{O}_Z, \mathcal{O}_Y)$;
Get $\mathbf{S}_{ZX}(\mathcal{O}_Z, \mathcal{O}_X)$; $last\_gLoss \leftarrow +\infty$;
while $\Delta^{\mathbf{T}_{ZX}}_{\mathbf{S}_{ZX}}(\mathcal{O}_Z, \mathcal{O}_X) < last\_gLoss$ **do**
  $last\_gLoss \leftarrow \Delta^{\mathbf{T}_{ZX}}_{\mathbf{S}_{ZX}}(\mathcal{O}_Z, \mathcal{O}_X)$;
  do one step $\Delta^{\mathbf{T}_{ZX}}_{\mathbf{S}_{ZX}}(\mathcal{O}_Z, \mathcal{O}_X) + \Delta^{\mathbf{T}_{ZY}}_{\mathbf{S}_{ZY}}(\mathcal{O}_Z, \mathcal{O}_Y)$;
  Optimise $\mathcal{O}_Z$ to satisfy $\mathbf{T}_{ZX}(\mathcal{O}_Z, \mathcal{O}_X)$;
  Get $\mathbf{S}_{ZX}(\mathcal{O}_Z, \mathcal{O}_X)$;
return $\mathcal{O}_Z$

Algorithm 2: The control process of HSphNN

Input: Target relations: $\psi_1(\mathcal{O}_1, \mathcal{O}_2)$, $\psi_2(\mathcal{O}_2, \mathcal{O}_3)$, $\psi_3(\mathcal{O}_3, \mathcal{O}_1)$;
Output: `SAT` or `UNSAT`;

Initialise $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{O}_3$ as coinciding;
if *all three relations are satisfied* **then return** `SAT`
fix $\mathcal{O}_1$, update $\mathcal{O}_2$ to satisfy $\psi_1(\mathcal{O}_1, \mathcal{O}_2)$;
fix $\mathcal{O}_1$, update $\mathcal{O}_3$ to satisfy $\psi_3(\mathcal{O}_3, \mathcal{O}_1)$;
do $COP^{\overline{\psi}_2}_{\psi_3}(\mathcal{O}_3|\mathcal{O}_2;\mathcal{O}_1)$, $\overline{\psi}_2(\mathcal{O}_3, \mathcal{O}_2) = \psi_2(\mathcal{O}_2, \mathcal{O}_3)$;
if *not all three relations are satisfied* **then**
  fix $\mathcal{O}_3$, update $\mathcal{O}_1$ to satisfy $\psi_3(\mathcal{O}_3, \mathcal{O}_1)$;
  fix $\mathcal{O}_3$, update $\mathcal{O}_2$ to satisfy $\psi_2(\mathcal{O}_2, \mathcal{O}_3)$;
  do $COP^{\psi_1}_{\overline{\psi}_3}(\mathcal{O}_1|\mathcal{O}_2;\mathcal{O}_3)$, $\overline{\psi}_3(\mathcal{O}_1, \mathcal{O}_3) = \psi_3(\mathcal{O}_3, \mathcal{O}_1)$;
if *all relations are satisfied* **then return** `SAT`
else return `UNSAT`

Experiments

Experiment 1

We challenge HSphNN to determine the *validity* of all types of syllogistic reasoning in one epoch.

Method To determine the validity of a syllogistic reasoning statement '$r_1(S, M)\ r_2(M, P) \therefore\ r_3(S, P)$', HSphNN tries to construct a counter-example, namely, $\mathcal{O}_S$, $\mathcal{O}_M$, and $\mathcal{O}_P$, satisfying $\psi(r_1)(\mathcal{O}_S, \mathcal{O}_M)$, $\psi(r_2)(\mathcal{O}_M, \mathcal{O}_P)$, and $\neg\psi(r_3)(\mathcal{O}_S, \mathcal{O}_P)$. If successful, HSphNN refutes it by providing a counter-example; otherwise, HSphNN concludes that this syllogistic reasoning is valid, as $M = 1$. Among 256 types of syllogistic reasoning statements, only 24 reasoning types are valid.

Setup Three spheres are randomly initialised as coinciding in a Poincaré disk, with the centres $\vec{O}$ following the uniform distribution and with the Euclidean radius $r = 0.3$. A Poincaré sphere is computed by setting its Euclidean centre to $0.9 \cdot \sin(\vec{O})$ and the Euclidean radius to $\sin^2(r)$. This is to prevent HSphNN from pushing them close to the boundary of the Poincaré disk and having NAN values. We set the learning rate to 0.0001 and the maximum number of epochs $M = 1$. All experiments were conducted on MacBook Pro Apple M1 Max (10C CPU/24C GPU), 32 GB memory. We challenged HSphNN to construct Poincaré spheres with the following dimensions $2, 3, 15, 30, 100, 200, 1000, 2000, 3000$.

Results HSphNN successfully determined 24 valid and 232 invalid syllogistic reasoning types by constructing Poincaré sphere configurations with dimensions from 2 to 3000, totalling 2304 cases, satisfying the *reasoning-for-sure* property. HSphNN needs more time to determine valid reasoning than invalid reasoning. The mean time cost to determine a valid reasoning is 17.71 seconds. The mean time cost to construct a counter-example for invalid reasoning is 3.46 seconds; 1732 among 2088 (82.91%) invalid reasoning cases are determined in less than 5 seconds; 169 among 216 (78.82%) valid cases are determined in less than 30 seconds.

Experiment 2

Trained by vast amounts of corpora, LLMs have achieved human-like communication and do better in syllogistic reasoning than humans, but also mimic human errors that appeared in their training corpus (Eisape et al. 2024; Lampinen et al. 2024). This experiment aims to demonstrate a pure neural dual-process model (Evans 2003; Sun et al. 2005; Kahneman 2011) for syllogistic reasoning – ChatGPT simulates System 1's fast associative thinking without guarantee, while HSphNN simulates System 2's slow logical reasoning that examines ChatGPT outputs, and sends feedback through prompts.

Setup We tested two ChatGPT versions (gpt-3.5-turbo and gpt-4o) with and without the feedback from HSphNN, on three kinds of syllogistic rules: (1) rules with meaningful words (e.g., '*All Greeks are human*'), (2) rules with simple symbols (e.g., '*All S are M0*'), and (3), rules with random symbols (e.g., '*All HWsF1eq9 are hONvNxop*').

We use the Level 6 TELeR prompt structure (Karmaker Santu and Feng 2023) to send a task request to ChatGPT, which consists of six parts: (1) assigning ChatGPT the role of a professional logician; (2) a detailed instruction, including output format; (3) a question; (4) a context; (5) an output explanation; and (6) an example. Our TELeR prompts instruct ChatGPT to decide the satisfiability of Aristotelian syllogistic reasoning. If ChatGPT answers satisfiable, it will output the configuration of spheres. HSphNN will check the correctness, including whether it is supported by its explanation. If one of the two is incorrect, HSphNN will send feedback to ChatGPT and let it think again.

Evaluation HSphNN is a neural model that automatically examines the decision and the explanation of ChatGPT and classifies it into six classes: (1) a correct decision with a correct explanation; (2) a correct decision with an incomplete explanation; (3) a correct decision with an incorrect explanation; (4) a correct decision with an irrelevant explanation; (5) an incorrect decision with a correct explanation; and (6) an incorrect decision with an incorrect explanation. The irrationality ratio is the percentage of cases in (2-5) among the 256 types of syllogistic reasoning.

Results and Discussions (1) Without feedback, ChatGPT (gpt-3.5-turbo) reached the best performance (correct decision and explanation) of 46.875% using simple symbols, with the irrationality ratio of 46.875%; ChatGPT (gpt-4o) reached the best performance of 82.42% using random symbols, with the irrationality ratio of 15.23%. For both versions, rules with meaningful words somehow disturbed the performance. (2) With a maximum of two rounds of feedback, ChatGPT (gpt-3.5-turbo) improved the performance to 58.98% and significantly reduced the irrationality ratio to 21.48%; ChatGPT (gpt-4o) improved the performance to 84.76% and reduced the irrationality ratio to 12.89%. (3) Increasing the number of feedbacks did not have much impact on improving the performance. For example, with a maximum of ten rounds of feedback, ChatGPT (gpt-4o) only improved the performance to 85.16%. This suggests that prompt engineering may not be an effective way to form the feedback loop from System 2 neural networks to System 1 neural networks.

Conclusion and Outlook

A valid logical conclusion explicitly states what is implicit in the premises (Simon 2019). Thus, additional training data is not needed, as demonstrated by HSphNN. Our method shows that sphere embedding is more suitable than vectors to explicitly represent set-theoretic knowledge, which exists widely in the spatial domain and high-level cognitive domains (Allen 1983; Bellmund et al. 2018; Benthem 1983; Dowty 1979; Smith 1996; Tversky 2019).

HSphNN can seamlessly integrate with traditional neural networks by hosting latent feature vectors at the sphere centre, achieving neuro-symbolic unification: HSphNN specifies part-whole relations between spheres, while traditional networks optimise orientations of sphere centres, as illustrated in Figure 6(A.1, A.2, B). Neuro-symbolic unification

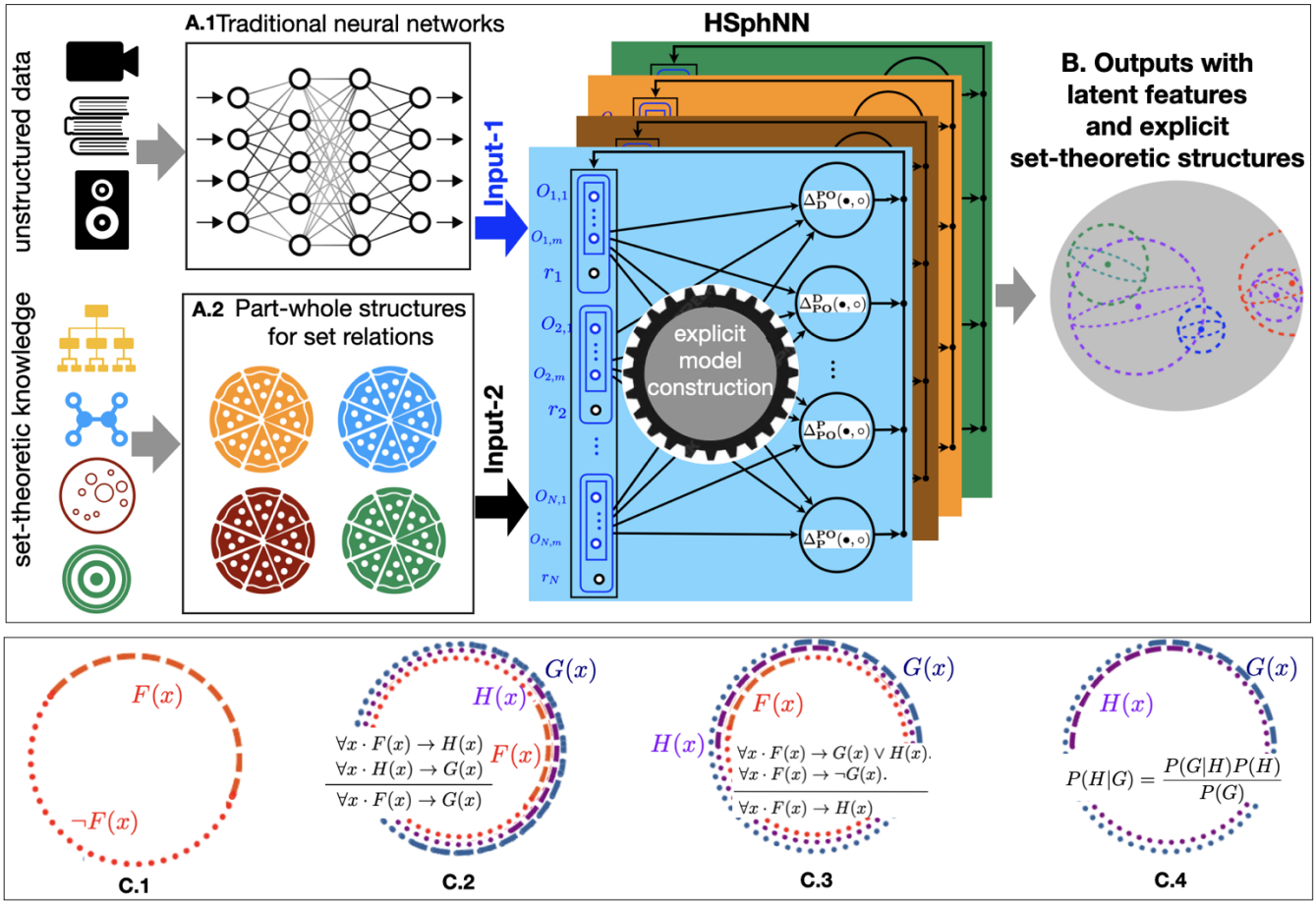

Figure 6: Traditional neural networks (A.1) learn vectors from the corpus, predicting centre orientations of spheres (Input-1). Set-theoretic knowledge is spatialised into part-whole relations (A.2), gearing lengths of centre vectors and radii of spheres (Input-2). Various set-theoretic domain knowledge introduces multiple channels. (B) HSphNNs construct a sphere configuration or refute the consistency of input set-theoretic knowledge. (C.1) 2-d arcs are closed under complement; arc configuration can be used for various logical reasoning, e.g., (C.2) hypothetical syllogism, (C.3) disjunctive syllogism, and (C.4) Bayesian reasoning.

brings mutual benefits. Using pre-training vectors as orientations of sphere centres may speed up the construction process; using boundary relations may prevent traditional neural networks from being biased by data.

Geometrically, a sphere can be understood as a set of points in *a universe*, whose distances to *a fixed point* (the centre of the sphere) are within *a constant* (the radius). If the universe is a circle, spheres will become arcs and can represent complement sets, as illustrated in Figure 6(C.1), and develop logical and Bayesian reasoning as follows.

Let $F(x)$, $H(x)$, and $G(x)$ be three sets that are represented by three arcs of the circle. If F arc is part of H arc, and H arc is part of G arc, then F arc is part of G arc (hypothetical syllogism). If F arc is part of the union of G arc and H arc, and F arc is part of G's complement arc, then we can prove that it is valid that F arc is part of H arc (disjunctive syllogism), as illustrated in Figure 6(C.2, C.3).

Knowledge of events can be understood and represented in the same way as knowledge of objects (Quine 1985). Let arcs represent events. The probability that the event H happens given the condition that the event G happens $P(H|G)$ is the ratio of the length of the intersection of H arc and G arc and the length of G arc, $P(H|G) = \frac{P(H \cap G)}{P(G)}$. Similarly, we have $P(G|H) = \frac{P(H \cap G)}{P(H)}$. Putting the two formulas together, we have the Bayesian rule $P(H|G) = \frac{P(G|H)P(H)}{P(G)}$. Like syllogistic reasoning having the Euler diagram as its spatial semantics, statistical reasoning (Pearl and Mackenzie 2018) can have arc configuration as its spatial semantics, as shown in Figure 6(C.4).

Therefore, arcs can serve as a unified semantic representation for both logical reasoning and statistical reasoning, the former considering whether arcs have part-whole relations, the latter considering what is the degree of part-whole relations. Cartesian product of arcs will be a spatial semantics for heterogeneous knowledge. In this way, neural networks using sphere embeddings can perform a variety of reasoning through model construction.

Acknowledgments

Part of this work was initiated by the Dagstuhl Seminar 21362 "Structure and Learning" and by the symposium of "AI-Assisted Mathematical Discovery" at the London Institute for Mathematical Sciences.

References

Abramson, J.; Adler, J.; Dunger, J.; and et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630: 493 – 500.

Allen, J. 1983. Maintaining Knowledge about Temporal Space. *Communications of the ACM*, 26(11): 832–843.

Bellmund, J. L. S.; Gärdenfors, P.; Moser, E. I.; and Doeller, C. F. 2018. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415).

Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Darrell, T.; Harari, Y. N.; Zhang, Y.-Q.; Xue, L.; Shalev-Shwartz, S.; Hadfield, G.; Clune, J.; Maharaj, T.; Hutter, F.; Baydin, A. G.; McIlraith, S.; Gao, Q.; Acharya, A.; Krueger, D.; Dragan, A.; Torr, P.; Russell, S.; Kahneman, D.; Brauner, J.; and Mindermann, S. 2024. Managing extreme AI risks amid rapid progress. *Science*, 384(6698): 842–845.

Benthem, J. v. 1983. *The Logic of Time*. Dordrecht, Holland: D.Reidel Publishing Company.

Biever, C. 2023. The easy intelligence tests that AI chatbots fails. *Nature*, 619: 686–689.

Brand, D.; Riesterer, N.; and Ragni, M. 2023. Uncovering Iconic Patterns of Syllogistic Reasoning: A Clustering Analysis. In *Proceedings of the 21st International Conference on Cognitive Modeling*.

Bucciarelli, M.; Khemlani, S.; and Johnson-Laird, P. N. 2008. The psychology of moral reasoning. *Judgment and Decision Making*, 3(2): 121–139.

Creswell, A.; Shanahan, M.; and Higgins, I. 2023. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. In *ICLR*.

Davies, A.; Velickovic, P.; Buesing, L.; Blackwell, S.; Zheng, D.; Tomasev, N.; Tanburn, R.; Battaglia, P. W.; Blundell, C.; Juhász, A.; Lackenby, M.; Williamson, G.; Hassabis, D.; and Kohli, P. 2021. Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887): 70–74.

Deng, W.; Pei, J.; Kong, K.; Chen, Z.; Wei, F.; Li, Y.; Ren, Z.; Chen, Z.; and Ren, P. 2023. Syllogistic Reasoning for Legal Judgment Analysis. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics.

Dong, T.; Jamnik, M.; and Liò, P. 2024. Sphere Neural-Networks for Rational Reasoning. *arXiv:2403.15297*.

Dong, T.; Wang, Z.; Li, J.; Bauckhage, C.; and Cremers, A. B. 2019. Triple Classification Using Regions and Fine-Grained Entity Typing. In *AAAI*.

Dowty, D. 1979. *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Eisape, T.; Tessler, M.; Dasgupta, I.; Sha, F.; van Steenkiste, S.; and Linzen, T. 2024. A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models. In *NAACL*.

Evans, J. S. 2003. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10): 454–459.

Fang, P.; Harandi, M.; Le, T.; and Phung, D. 2023. Hyperbolic Geometry in Computer Vision: A Survey. arXiv:2304.10764.

Fedorenko, E.; Piantadosi, S. T.; and Gibson, E. A. F. 2024. Language is primarily a tool for communication rather than thought. *Nature*, 630: 575–586.

Gigerenzer, G. 2022. *How to Stay Smart in a Smart World: Why Human Intelligence Still Beats Algorithms*. The MIT Press.

Goodwin, G.; and Johnson-Laird, P. 2005. Reasoning About Relations. *Psychological review*, 112: 468–93.

Goyal, A.; and Bengio, Y. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478.

Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; and Rueckert, D. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*.

Hammer, E.; and Shin, S. J. 1998. Eulers visual logic. *History and Philosophy of Logic*, 19(1).

Jeffrey, R. 1981. *Formal logic: Its scope and limits (2nd ed.)*. New York, NY:McGraw-Hill.

Johnson-Laird, P.; Byrne, R.; and Khemlani, S. 2024. Models of Possibilities Instead of Logic as the Basis of Human Reasoning. *Minds and Machines*, 34.

Johnson-Laird, P. N.; and Byrne, R. M. J. 1991. *Deduction*. Lawrence Erlbaum Associates, Inc.

Johnson-Laird, P. N.; Khemlani, S.; and Goodwin, G. P. 2015. Logic, probability, and human reasoning. *Trends in Cognitive Science*, 19(4): 201–214.

Kahneman, D. 2011. *Thinking, fast and slow*. Allen Lane, Penguin Books.

Karmaker Santu, S. K.; and Feng, D. 2023. TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14197–14203. Singapore: Association for Computational Linguistics.

Khemlani, S. 2021. Psychological theories of syllogistic reasoning. In Knauff, M.; and Spohn, W., eds., *Handbook of Rationality*. Cambridge, MA, USA: MIT Press.

Khemlani, S.; and Johnson-Laird, P. N. 2012. Theories of the Syllogism: A Meta-Analysis. *Psychological Bulletin*, 138(3): 427–457.

Knauff, M. 2009. A Neuro-Cognitive Theory of Deductive Relational Reasoning with Mental Models and Visual Images. *Spatial Cognition & Computation*, 9(2): 109–137.

Lampinen, A. K.; Dasgupta, I.; Chan, S. C. Y.; Sheahan, H. R.; Creswell, A.; Kumaran, D.; McClelland, J. L.; and Hill, F. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7).

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's Verify Step by Step. arXiv:2305.20050.

Lv, X.; Hou, L.; Li, J.; and Liu, Z. 2018. Differentiating Concepts and Instances for Knowledge Graph Embedding. In *EMNLP*, 1971–1979.

Mitchell, M. 2023. How do we know how smart AI systems are? *Science*, 381(6654): adj5957.

Nickel, M.; and Kiela, D. 2017. Poincaré Embeddings for Learning Hierarchical Representations. In *Nips*, 6338–6347.

Park, P. S.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2024. AI Deception: A Survey of Examples, Risks, and Potential Solutions. *Patterns*, 5(5).

Pearl, J.; and Machenzie, D. 2018. *The Book of Why*. Penguin Books.

Quine, W. V. 1985. Events and Reification. In Lepore, E.; and McLaughlin, B., eds., *Actions and Events: Perspectives on the Philosophy of Davidson*, 162–71. Blackwell.

Ren, H.; Hu, W.; and Leskovec, J. 2020. Query2box: Reasoning over Knowledge Graphs in Vector Space Using Box Embeddings. In *ICLR*.

Rosenblatt, F. 1962. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, USA: Spartan Books.

Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; Lillicrap, T.; and Silver, D. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588: 604—-609.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature*, 550: 354–359.

Simon, H. A. 2019. *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Smith, B. 1996. Mereotopology: A Theory of Parts and Boundaries. *Data and Knowledge Engineering*, 20: 287–303.

Sun, R.; Slusarz, P.; ; and Terry, C. 2005. The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112(1): 159–192.

Tversky, B. 2019. *Mind in Motion*. New York, USA: Basic Books.

Wang, D.; Jamnik, M.; and Liò, P. 2018. Investigating Diagrammatic Reasoning with Deep Neural Networks. In *Diagrams 2018*, 390–398.

Wang, D.; Jamnik, M.; and Liò, P. 2020. Abstract Diagrammatic Reasoning with Multiplex Graph Networks. In *ICLR*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. 2022. STaR: Bootstrapping Reasoning With Reasoning. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Zhang, S.; Wu, Y.; Che, T.; Lin, Z.; Memisevic, R.; Salakhutdinov, R. R.; and Bengio, Y. 2016. Architectural Complexity Measures of Recurrent Neural Networks. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Zhang, Z.; Wang, J.; Chen, J.; Ji, S.; and Wu, F. 2021. ConE: Cone Embeddings for Multi-Hop Reasoning over Knowledge Graphs. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 19172–19183. Curran Associates, Inc.