Your red isn't my red! Connectionist Structuralism and the puzzle of abstract objects

Chris Percy

Long-form abstract

This paper presents a nine step argument for "Connectionist Structuralism" (CS), an account of the ontology of abstract objects that is neither purely nominalist nor purely platonist. CS is a common, often implicit assumption in parts of the artificial intelligence literature, but such discussions have not presented formal accounts of the position or engaged with metaphysical issues that potentially undermine it. By making the position legible and presenting an initial case for it, we hope to support a constructive dialogue between AI researchers and philosophers of metaphysics that helps both sides to refine the position.

CS proposes that each abstract object we can draw on in human analysis corresponds to a particular subset of an individual person's brain structure whose functionality is isomorphic to a subset of the nodes and connections in a suitable connectionist network. In other words, abstract objects are physically realised, but in individual brains, rather than only in the referent objects (pure nominalism) or in metaphysical universals (pure platonism).

This paper's minimum claim is that CS can account for all abstract object predicables regarding sensible properties, such as "is red" or "is a square". Using evidence from cognitive neuroscience, machine learning, and evolutionary biology, as well as a fully traceable toy example, we describe how CS can support our core cognitive uses of such sensible properties and can account for our core phenomenal experiences of them. In the former, CS provides sufficient albeit imperfect inferential safety, whose limitations are argued to strengthen rather than weaken the case for CS as describing human behaviour. In the latter, four target phenomenal features are accounted for – abstract objects as feeling intangible, non-located, transparent, and unchanging - along with accounts for further phenomena such as semantic refinement, the Stroop effect, synaesthesia, semantic clarity, sensory overload, and satiation.

Our minimum claim concerns the day-to-day usage and felt experience of abstract objects, but we suggest also an extended claim in which CS can form the basis of a pragmatic sufficient logic. As such, the initial outline of a response is provided for five common objections to positions that seek to ground abstract objects without reference to metaphysically stand-alone universals: referential opacity; identity of indiscernibles; infinite regression; non-physical concepts; and necessary truths. These outlines lay the foundation for (but do not seek to formally demonstrate) an extended CS account that addresses other abstract objects, including issues relating to their use in mathematics and formal logic.

The CS account leads to a four-layer hierarchy of similarity for whether your "red" is the same as mine. Considering both semantic and phenomenal similarity, we conclude that our "reds" are likely non-identical but can be made close enough for practical purposes. Finally, we describe how future work could elaborate CS as a metaphysical project and how confidence in it could be tested through empirical research.

Key words: Connectionism; Nominalism; Abstract Objects; Universals; Artificial Intelligence

Introduction

Many have enjoyed debating whether one person's experience of red is the same as another's, whether we could ever know for sure, and whether it would matter if it weren't. Often discussed while pointing at a particular red-coloured object, the topic can also be productively phrased in the language of universals and abstract objects. When one person reflects on the concept of "redness", without any particular object in mind, does that experience feel similar to another person's (phenomenal angle) or, at least, are they invoking the same semantic concepts (utility angle)?

This paper assembles "Connectionist Structuralism" (hereinafter "CS") as a proposal for the ontological metaphysics of abstract objects - such as our mental concepts of redness or squareness - drawing on findings from computational neuroscience, machine learning, and brain biology. In a sentence, the proposal is that each abstract object we can draw on in human analysis corresponds to a particular subset of an individual person's brain structure whose functionality is isomorphic to a subset of the nodes and connections in a suitable connectionist network

This proposal or something similar is often implied in the machine learning and artificial intelligence literature, but such accounts do not engage with standard metaphysical objections or human phenomenology, although critiques do run in the reverse direction (see, e.g., Froese, & Taguchi, 2019; Olah et al., 2020). By making the position legible and presenting an initial case for it, we hope to support a constructive dialogue between AI researchers and philosophers of metaphysics that helps both sides to refine the position.

This paper's minimum claim is that CS can account for all abstract object predicables that are accessible to external sensory experience ("sensible properties"), e.g. those instantiable in observed physical objects of the type exemplified by "is red" or "is a square". Moreover, that CS can account for such sensible properties in a way that respects the phenomenology commonly reported for abstract objects.

Many metaphysical arguments have been assembled against different accounts of abstract objects, with the debate remaining contested in recent literature (e.g. Himelright, 2022; Carmichael, forthcoming; Imaguire, 2022). The extended claim is that the same core CS principles can be combined with other structures to account for all other abstract objects - such as mathematical propositions, counterfactuals, physically-impossible concepts, intentions, and emotions – hence forming the foundation for a formal logic that can meet the pragmatic needs of day-to-day usage. This paper sketches an outline towards some of these other abstract objects, but full treatment is reserved for future work.

Under CS, abstract objects do exist (helping to dissolve some critiques against traditional nominalism), but *ante rem* or *in re* universals do not. Abstract objects have locations outside of their referents or instantiations in the physical world; specifically, they have bounded physical presences inside each information processing mechanism that has both capability and cause to derive them. The proposal does not, however, require reductionist or radical connectionism, the notion that all human cognitive faculties, mental experiences, and consciousness can be explained by connectionist-style networks. Instead, a much weaker and better evidenced claim is sufficient: that various connectionist networks can mechanically encode abstractions of various types and associate them with language, and that the necessary minimum capabilities exist in various brain systems.

The paper proceeds as follows. Section one introduces the target problem, differentiating the phenomenology of abstract objects from their semantic utility, and situates CS within the literature.

Section two sets out the structure of the argument in nine steps, defining terms and providing subsections on connectionist models that implement the definitions including a summary of a fully worked toy example (full details in the appendix), evidence that human brains have the necessary minimum connectionist capabilities, and details of how such networks can account for the utility and phenomenal explananda from section one.

Section three provides a response to three key objections raised against nominalists that potentially bind on the paper's minimum claim (referential opacity; identity of indiscernibles; infinite regression). It also sketches an outline response to a further two challenges beyond the paper's minimum claim (non-physical concepts; necessary truths).

Section four interprets the age-old question of interpersonal subjectivity in the context of this account of abstract objects. Specific, measurable circumstances are described under which "your concept of redness" can be described as either the same as mine or different from mine. A four-layer hierarchy of sameness tentatively suggests that such concepts are likely to be non-identical, but nonetheless sufficiently isomorphic that we can work with them in a spirit of pragmatism, humility, and iterated interactions.

The conclusion summarises the claims and suggests further work to elaborate CS, including how it might be falsified, or at least refined, through lab experiments.

1. Context: The target explananda and academic context

The puzzle of where abstract objects reside is long-standing, but a brief account can be provided here to define the terms and context for the rest of the paper.

1.1. Defining the physical and non-physical realms

It is helpful to start with defining two realms of existence where entities might reside, termed "physical" and "non-physical".

The "physical realm" consists of what we can perceive with our external senses (touch, eyes, ears, etc.) and interact with via our motor functions (fingers, voice, etc.). The physical realm includes physical human bodies and brains, including those external sense and motor organs. The objects in the physical realm appear to have locations in space-time, but the realm can be posited without assuming that the four dimensions we perceive are an accurate or complete reflection of all that exists and without assuming our perceptions are exact. In other words, the picture can be consistent with indirect realism.

Objects are defined broadly to include complex objects built up from fundamental entities (whatever those may be), as well as patterns and events consisting of those objects. Such objects are typically perceived as bounded in spacetime, i.e. there are specific spatiotemporal locations where they appear to be present and locations where they appear not to be, albeit the precise boundaries may be fuzzy for a range of reasons[1]. However, we do not need to rule out (or categorically confirm) phenomena that appear to be universally located and can be considered single entities, such as universal quantum fields or a fabric in which objects reside (spacetime in the theory of relativity). We do rule out anything which has "no location" or for which the concept of location, universal or bounded, is meaningless.

The existence of this physical realm and particularly its intersubjective consistency is an assumption, since we only know of it through our senses which can be flawed or tricked. This paper makes this assumption but acknowledges it cannot be conclusively proved (see Comesaña & Klein, 2019, for an overview). However, in any case, the physical realm may not be the only realm of existence.

The "non-physical" realm is largely defined as anything not in the physical realm. Entities exist in this realm, but whatever location or substance they might have there, they have no independent location or substance in the physical realm. From the perspective of our external sense and motor organs, these entities are intangible. If we are capable of interacting with entities in this realm, it must be via something else, perhaps certain mental faculties. Multiple such non-physical realms might exist. Such non-physical realms could be proposed as the "place" of existence of various human experiences, including sensations of spirits or deities, ideas or mental experiences, consciousness, mathematics, universals, and abstract objects. This paper focuses on abstract objects, particularly on sensible properties.

1.2. Summarising the problem

The target explananda derive from our day-to-day experience and usage of abstract objects. For instance, in the physical realm, we never sense "squareness" except in its instantiation as part of something else: various square objects or embedded communications of various definitions (e.g. written down on paper, encoded as words in sound waves etc). Nonetheless, within our minds, we can typically conceive of "squareness" divorced from any one instantiation and from all instantiations. However, in the physical realm as defined above, nothing can exist divorced from all instantiations.

The ability to operate with abstract objects conveys considerable linguistic and cognitive benefits ("its utility") and is accompanied by a particular felt experience ("its phenomenology"). An ideal position should maintain the former and account for the latter. The primary explanatory challenge for physicalists comes in doing so with reference only to the physical realm. This section will first describe the utility and phenomenal explananda and then situate CS with reference to several broad classes of solution proposed in the literature.

[1] E.g. because our measurement instruments, senses, or reasoning faculties are not precise enough to see their precise edges, because the edges change in infinitesimal units of time or at least more rapidly than we can complete a measurement cycle, because of uncertainty/unknowability principles in physics, or because they are ontologically fuzzy in some sense.

1.3. The utility explananda

Abstract objects play a major role in our conventions of language and communicated reasoning, as well as being central to the generalisations and categorisations of the natural sciences. Several ways exist of formalising the subsequent linguistic challenge. This paper will follow Himelright's (2022) *problem of inferential safety* as its central utility explanandum, noting other features of language, learning and cognition as part of the broader discussion.

To paraphrase, the problem is how we can be confident ("safe") in conclusions about concrete (i.e. physical realm) objects arrived at where any premises about concrete objects are true and any premises about abstract objects are such that platonists would regard them as true, i.e. if there truly were a non-physical realm with a perfectly true benchmark to measure them against. Those who believe in non-physical realms and our ability to access them have a straightforward answer to this, but nominalists have to explain how their account of abstract objects enables this.

Himelright's phrasing neatly encapsulates the core of what we want a theory of abstract objects to achieve. Semantic interactions between abstract and concrete objects are possible to motivate if the abstract objects are, in some sense in the proposed ontology, capable of instantiation (e.g. predicable) and of being separable from their instantiations. This enables us to derive abstract objects from their physical instantiations and to use abstract objects to talk and reason about shared features of concrete objects.

Himelright constructs a formal language in which various classes of sentences about the properties of concrete objects are safe. He notes that his language does not (yet) help in all circumstances, such as mereological nihilism, but his footnote 1 references other safety proofs that can apply in other circumstances, albeit with their own caveats. Many other solution approaches exist, including the reductive A+ approach from White (2022) in which interpreting universals as discussions about patterns in particulars is argued to be metaphysically sufficient. CS is not inconsistent with such formal language approaches to nominalism, but additionally proposes an ontological lens they can adopt which helps account for certain phenomenal explananda and metaphysical objections.

1.4. The phenomenal explananda

Philosophers have traditionally devoted more of their metaphysical writing to the issues of language and logic in §1.3, but several recent papers take an explicitly phenomenological approach: asserting both that there is something it feels like to think and commenting on what that something is.

Phenomenology is, by definition, drawn from reports of first person felt experiences, so the strength of the account depends either on a reader's ability to empathise with such experiences through introspection (even if felt only sometimes or to differing degrees) or on a willingness to accept the reports as a genuine report of experience (even if they may disagree with particular conclusions that might be inferred from those feelings). As set out in §2, the core CS argument is not reliant on phenomenology alone, which is instead construed as supporting evidence.

Pitt (2004) argued that there is something that it is like to have thoughts about propositions, i.e. there is an accompanying phenomenal experience even for intentional thoughts as well as explicitly phenomenal states, such as seeing red or feeling pain. Propositions can contain abstract objects and are themselves an abstract object when held in the mind. Hence Pitt's experience and claim support the case in this subsection.

Pitt's (2004) account is primarily analytic in nature, supported by appeals to readers' introspection on reading his psycholinguistic sentences. He asserts that it would be impossible introspectively to distinguish conscious thoughts with respect to their content if there weren't something it is like to think them (in the sense of entertaining them rather than necessarily believing them). He concludes that each type of conscious thought must have a proprietary, distinctive and individuative phenomenology, although this has been challenged (e.g. Levine, 2011). A weaker conclusion is sufficient here: that thoughts about abstract objects have a phenomenology, whether fully or partially shared across objects, or even with nothing shared at all beyond all being thoughts.

Some of Pitt's language begins to convey the shape of his experienced phenomenology as well as his logical argument, such as "when we introspect, we turn our attention inward, toward the contents of our minds – which are mental if anything is." (p22). This emphasis on mental hints at a feeling of non-physicality, which we will see below is one of the most common descriptions of abstract objects and one which rules out a set of well-known felt experiences, being those involving physical, tangible objects.

Smith (2011) also explores the phenomenology of consciously thinking, reflecting on his experience of thinking through three specific propositions, notwithstanding the difficulty of putting such experiences into words (p359). He describes conceiving abstract objects in those propositions as a "distinctly non-sensory and non-perceptual form of experience" (p351) and as "ethereal" (p368). In some modes of consciousness, he reports that some more abstract propositions can feel "transparent" (p362) or at least "translucent" (p362), noting that the concept of redness can be thought of without experiencing red colour or the concept actually being itself red-coloured. He motivates translucency by arguing that our cognition is embodied: we are world-directed and our awareness or thinking is shaped unavoidably by the external world.

In addition to these explicit commentaries on phenomenology, the features of abstract objects that have been referred to in metaphysical discussions of their nature provide further supporting evidence. Reference sources explain Plato's Forms as abstract, perfect, unchanging ideals that transcended time and space (e.g. Meinwald, nd.; Silverman, 2022). Rettler and Bailey (2017) discuss a range of candidate features of abstract objects: predicates, not in space time, not sense-perceptible, obeying the Identity of Indiscernibles, instantiated (or at least instantiable), and capable of being in multiple locations. Carmichael (forthcoming) takes the central feature of abstract properties as predicables (i.e. they can be true or false of a given object) that "(perhaps contingently) lack locations", where platonic realism would further claim they are mind-independent. Several of these features relate more to requirements for linguistic and cognitive utility (§1.3), but others have a more direct phenomenological interpretation.

Collectively, this subsection suggests two primary phenomenal explananda as a dominant theme across many accounts: intangible and non-located, which incorporates notions of being non-physical, ethereal, and non-sense-perceptible. Two further potential explananda include transparency/translucency and being unchanging.

The purpose of this subsection is not to draw ontological conclusions directly from phenomenology. Indeed, taken at face value, felt experiences of ethereal, intangible objects not located in space lead directly to the ontology of some non-physical realm and may be one strong initial intuition that leads to platonism and related theories. However, whatever ontology is proposed should also account for these felt experiences, since we know those exist in some sense. The account can explain why the experiences are illusory or misleading, but should have an explanation for such, rather than being dismissive with no further reasoning.

1.5. Classes of solution and CS

The two broad classes of solution are platonism (very similar to realism in this context), in which abstract objects have some mind-independent existence, and nominalism, in which they do not. Realism might invoke concepts similar to *ante rem* universals in a non-physical realm[2]. Such accounts have straightforward explanations of the utility and phenomenal explananda. For the former, there is a single entity corresponding to each abstract object which we all reference and which exists regardless of physical realm objects. For the latter, those entities actually are non-physical and non-located, so it unsurprising we experience them as such. The primary explanatory challenges are, instead, what this non-physical realm is, what it means to exist in it, and how we can interact both with it and what appears to be the physical realm (see for instance Montero's 2022 discussion of how platonic abstracta might come to have causal relevance).

Nominalism, by contrast, has no explanatory challenge about interactions across realms, since there is only the physical realm, but it has no automatic account for the explananda. Many subtypes of nominalism have been proposed to provide such accounts, particularly for the utility explananda, including the summaries and counter-arguments in Cowling et al. (2023).

CS positions abstract objects as derived from concrete objects. They are not *ante rem*, and no non-physical realm is proposed. The theory is nominalist in the sense that it is anti-universals and anti-platonism. Abstract objects are derived from concrete objects rather than immanent within them, so *in re* positions are also rejected. But it differs from nominalism in identifying a physical space for abstract objects that is separate from their referents' original physicality.

[2] In differentiation from some platonist accounts, realism might invoke an *in re* argument in which the abstract objects exist only in the physical realm objects to which they apply, either as multiply-located universals or as separate type/token instantiations.

Under CS, abstract objects have a physical existence *in cerebro,* i.e. in each physical brain corresponding to a person invoking them.

2. Argument for Connectionist Structuralism (CS)

The argument proceeds in nine steps:

P1. An abstract feature shared by a set of sensed external objects, such as a common property or category, can be considered encoded in an information processing mechanism when a subset of the system has a state or states that uniquely correspond to that feature. In other words, the relevant state occurs if and only if presented with any one or more of the objects instantiating the relevant feature. Objects are understood in the general sense of physical realm entities described in §1.1 and information processing systems are understood in the connectionist sense, with an example in the appendix.
[Definition]

P2. A wide variety of simulated connectionist networks, from simple to complex and using different connection and updating logics, are able to encode features in the sense of P1 and can arbitrarily associate specific features with other simultaneous or approximately co-occurring phenomena, such as symbols or words.
[Simulation capability claim, backed by a fully worked toy example and evidence from the machine learning and computational neuroscience fields]

P3. The specific subset of nodes and edges in a connectionist network that corresponds uniquely to an abstract feature in the P1 sense can be considered the particular physical location and pattern that corresponds to the abstract object referencing that feature.
[Definition]

P4. Across a variety of contexts, there is evolutionary value in the capabilities that are achieved by the information processing mechanisms in P1.
[Backed mainly by evidence from evolutionary biology]

P5. The human brain has evolved to contain structures whose mechanisms can, *inter alia*, implement the functionality described in P2.
[Backed mainly by evidence from biological neural networks]

C1. Given P5, it is plausible that structures sufficiently similar to P2 are used in the brain to account for at least some of the occasions when P4 capabilities are present.

- "Sufficiently similar" in the sense that it could be pragmatically modelled by some form of connectionist architecture, acknowledging that there is considerable uncertainty over exactly how the brain processes inputs and that it is highly likely to be more complex and nuanced than the various examples presented in support of P2.
- If deliberately simple connectionist architectures can achieve the result, then it is plausible a fortiori that more complex connectionist architectures also achieve it in a way that still corresponds to the general definitions in P1 and P3.

P6. The P1 and P3 definitions suffice for human usage of abstract objects in way that resolves the problem of inferential safety, subject to caveats which correspond with practical experience. Connectionist network features also resonate with particular features of our cognitive abilities, such as semantic refinement, the Stroop effect, and synaesthesia. *[Narrative based account with examples]*

P7. When combined with a mechanism generating or corresponding to conscious awareness (being neutral about what this mechanism might be), the definition in P3 and examples in P2 can account for four phenomenal features – most importantly, intangibility and non-location, followed by transparency/translucency and unchangingness – and resonate with several others, such as degrees of semantic clarity, sensory overload, and semantic satiation. *[Narrative based account with examples]*

C2. Given P5, the supporting evidence in P6 and P7 makes CS a strong candidate for an ontology that can account for the utility and phenomenology of a particular set of abstract objects ("sensible properties") as used by humans. This evidence remains in development, but the insights so far give some confidence that further aspects of utility and phenomenology could be explained by a CS-underpinned account, emphasising again the point in C1 that the brain may have other systems, both connectionist and non-connectionist, that combine with CS to provide a full account.

Subsections now provide more detail on the substantive claims in P2 (2.1), P4 and P5 (2.2), P6 (2.3), and P7 (2.4). The worked examples in 2.1 also illustrate the definitions in P1 and P3. The plausibility assertions in C1 and C2 stand on the strength of surrounding arguments.

2.1. The capabilities of simulated connectionist networks (P2)

Connectionism refers to a broad class of networks, defined abstractly as units (or nodes) joined together by various inward, outward, or bidirectional connections (or edges). Nodes can have multiple connections. There can be different types of node and connection which have different activation/deactivation rules, given both inward activity and elapsed time.

In a physical realm sense, such networks are only capable of receiving inputs, being activated, or interacting physically *if they are also embedded in a physical substrate* that can implement the rules of the network. Such networks are described here as substrate-neutral but not substrate-independent: they need a substrate to exist but any substrate that meets the requirements will do (e.g. discussion in Gómez-Emilsson & Percy, 2022). One such substrate is being implemented on a digital computer, termed here as simulated networks. As an aside, a description of a network can exist without that network being substrate-implemented (i.e. having its own existence) or even being implementable. Such descriptions of course also need to be embodied to exist, whether embodied in the paper you are reading or embodied in my brain while I write the paper.

The purpose of this subsection is to demonstrate that a mechanical information processing network can extract sensible properties and categories from sensed physical realm objects and cross-associate them with words sensed alongside those objects. Two forms of evidence are provided. First, a fully worked toy universe example is provided in the appendix and summarised here. Second, a brief account is provided of more sophisticated connectionist network simulations that have been able to process input information in diverse ways.

The appendix describes the full mechanics and stable output of a 68-node network that is trained through solely local, unsupervised example-based Hebbian principles. The network abstracts two colours and two shapes from a toy universe consisting only of 12 possible active inputs that combine blue squares, red squares, blue rectangles, and red rectangles (see A5 for details). The word-associated colour properties and shape categories of images are encoded reliably in the network in a way that implements two basic information processing capabilities: feature extraction and cross-association. The objective is not to design a system capable of more sophisticated information processing tasks, like generalisation or error-correction, but rather to provide a simple enough example that it can be implemented in full by hand or followed through in the text. The concrete approach allows us to trace exactly why the described patterns occurred as they did and how they interact with the rest of the network. Despite its simplicity, the network can encode colour and shape based the image inputs alone (A6, A7) and cross-associate sound and image such that sound-related nodes are activated based on image-only inputs and vice versa (A8).

These encodings meet the definition in P1 such that specific trained edges within the network can be pointed to that capture the different abstract objects. For instance in A8, redness is cleanly differentiated from blueness within the mechanism by any one of three edges (q_2q_4, r_2d_2, and s_2e_2). The abstract object of redness exists multiply as a physical object in each edge. All three are activated if redness is seen and heard. The first activates if any redness is seen (irrespective of auditory input). The second activates on either red images or sounds but would activate sound-related nodes even if only the image is seen. The third also activates on either but would activate image-related notes even if only the sound is heard.

A wide variety of more sophisticated connectionist networks have been developed to implement information processing in a much broader sense than this toy example. Simple Hebbian networks have been shown capable of conducting principle components analysis i.e. generalised feature extraction from quantitative inputs (Oja, 1982), learning to extract letter cases and operate basic procedural memory (Wong, 2019), and implementing mirror neuron functionality (Keysers & Gazzola, 2014). Many extensions are being explored to reflect biological neural network features that are more complex than simple Hebbian learning (e.g. Burns, 2021), such as inhibitory neurons (O'Reilly, 2001), network plasticity (Chen et al., 2019), derivative-driven temporal relations (Zappacosta et al., 2018), as well as the potential for specific informational mechanisms or concepts to be hard-coded in the network upon initialisation (Ngiam et al., 2010).

In a trivial sense, any network whose outputs classify a particular abstract feature is also successfully encoding the accompanying abstract object in a P1 and P3 sense (output nodes are also part of the network). Nonetheless, recent work has been increasingly examining internal or middle layers of neural networks to identify the latent encoding of particular abstract objects (e.g. Olah et al., 2020; Lees et al., 2021), similar to the appendix example but for more sophisticated models.

The most powerful modern connectionist networks are trained using gradient descent and backpropagation, using supervised or semi-supervised learning. Such networks sometimes also incorporate Hebbian style unsupervised learning into their structures (e.g. Lagani et al., 2021; Stanley et al., 2016). These networks can display a broader range of information processing functions, from general capabilities like single neuron polysemanticity (Elhage et al., 2022) and transformer-like attention (Ellwood, 2023), to human-comparable (or better) domain-specific abilities in such areas as language translation (De Vries et al., 2018) and playing Go (Wang et al., 2016).

Considerable debate exists about the reasoning abilities of recent large language models (LLMs) such as Bard, GPT-4, and several others, e.g. Bubeck et al. (2023) in support and Berglund et al. (2023) against. Such debates notwithstanding, the evidence of progress in connectionist models is undeniable. Nonetheless, CS does not require the claim that today's connectionist models or their descendants are capable of general reasoning abilities. The cited evidence of a range of more humble abilities is adequate to support P2.

2.2. Connectionist structures in the human brain (P4, P5)

Behavioural ecologists often assume that natural selection will produce organisms that process information about their environment, action space, and likely outcomes to make optimal decisions, although emphasising that such rationality would be environment specific (Trimmer & Houston, 2014) and likely bounded by satisficing principles (Ben-Haim, 2012).

Where diversity generation mechanisms, heritability, and evolutionary selection pressures are present, there is reduced pressure to motivate why any one specific connectionist structure should exist that happens to work well in a particular setting. Indeed, considerable structural connectionist diversity is found in biological brains. This part of the paper assumes the human brain, as a biological system, is the product of evolutionary processes, but does not make or require any equivalent assumptions about phenomenal consciousness.

Simple Hebbian learning and its direct extensions are of high interest to cognitive neuroscientists as biological neural networks in the human brain have long been confirmed to have sophisticated implementations of the same core functionality (e.g. Sumner et al., 2020). Evidence also exists on inhibitory neurons in the brain (Swanson & Maffei, 2019), diverse time-dependent learning processes (Zappacosta et al., 2018), and sophisticated spike-timing dependent plasticity (Song et al., 2000; Caporale & Dan, 2008). While most attention has focused on neural networks, the CS argument is not weakened if other brain cells or structures have the necessary capabilities and instead implement the necessary networks.

Historically, backpropagation has been considered unlikely to apply in a biological context, since it relies on biologically-challenging global information transfer across the network (O'Reilly, 2001). This limits the relevance of the most powerful connectionist models in accounting for aspects of human cognition, but recent research is uncovering ever more biologically-plausible approximations of backpropagation (Whittington & Bogacz, 2017) and gradient descent (Berlemont & Nadal, 2022).

These remain active areas of research and it is plausible to conclude that greater biological understanding of biological neural networks will continue to uncover new and useful connectionist functionality in the human brain. In any case, the worked example in the appendix relies only on unsupervised Hebbian learning, of the type that has been widely demonstrated in biological brains.

2.3. Resemblance to human linguistic and cognitive features (P6)

The CS approach provides a workable path to resolving the problem of inferential safety, admittedly with a dose of pragmatism and humility.

As discussed in 1.3, the CS ontology of abstract objects allows them to exist separately from their instantiations, i.e. as sub-parts of a connectionist network that is capable of encoding them and to be associated by that information processing mechanism with the physical realm objects that instantiate them. This works for mereological components and abstract predicables, e.g. a table's "having legs" and "legs" as a separately observable object, as well as a table's "being red", where redness can never be observed independently of an instantiation. From within an information processing mechanism, the activation of the subsystem that encodes a specific abstract object can be independently and safely connected to various other information processing tasks, including internal cogitation (such as recall or planning) and motor functions (such as triggering a flight response or language).

The limitation of the CS approach, from the perspective of inferential safety, is it does not provide perfect safety in two distinct respects. First, within the mechanism, the inferences are only safe based on the abstractions derived within the system. These abstractions may not reflect all relevant external structure, perhaps due to faulty senses or system training, or because insufficient differentiated samples have been sensed to encode the right abstract objects. Second, in linguistic communication between two such mechanisms, it is challenging to be 100% certain that that the abstract object one has encoded as redness is the same as another's (see §4 for details). We cannot aspire to claim perfect access to a perfect realm of forms to guarantee our own accuracy or trust someone else's claimed access to the same realm to guarantee agreement.

These caveats are, however, features to be welcomed as support for the theory rather than bugs to be fixed via a different metaphysics. Seeking out and learning from new experiences and refining concepts is something we can continually and hopefully eternally, both as individuals and as communities, and is something we should be proud of. Moreover, the need to take care over defining terms and being mindful of linguistic miscommunications is likely of little surprise to most readers. Many scholars have convincing accounts of imperfect language alignment both with reality and among its users (e.g. Wittgenstein, 1953; Derrida, 1967). The familiar, pragmatic solution is to invest an amount of effort in refining one's own abstract objects and ensuring alignment to others that is proportionate to a particular task. We may not be able to guarantee certainty for eternity, but we can take steps to be confident enough for now.

There are further parallels to be found between the toy model and cognition. The uncertainty in the toy model regarding what “squareness” really means is a simple example of what a word really encodes in a system and how new samples might be needed to refine it to match its common English referent (A10). A child’s learning journey between terms like square, rectangle, and quadrilateral might take many iterations and corrections before settling on the technical and precise English usage.

As a second example, the toy model’s internal confusion in response to contradictory inputs (A11) perhaps foreshadows the results of experiments that show slower response time in association tests when faced with incongruent inputs (Stroop effect, e.g. MacLeod 1991). Doubtless, far more is going on in the human case, but a delay triggered by efforts to resolve an initial internal contradiction can be imagined in both cases (see 3.4). Likewise, the phenomenon of synaesthesia may be partly accounted for by the potential for arbitrary many-to-many associations in a connectionist network.

2.4. Resemblance to human phenomenology (P7)

Our experience of a first person perspective remains something hard to explain within a third person empirical framework, with some arguing that such subjectivities are fundamentally unfalsifiable or outside the domain of science (e.g. Cohen & Dennett, 2011). Nonetheless, no shortage of mechanisms have been proposed (e.g. the incomplete list of 22 theories in Table 1 from Seth & Bayne 2022), even if epistemologies other than falsificationism should prove necessary to form a view on which is likely to be true. Since our phenomenology is experienced from a first person perspective, some set of mechanisms needs to be assumed before a third person theory, like the account so far, can be translated into a felt experience.

A broad class of physicalist and embodied theories of (phenomenal) consciousness have a common feature in that they are capable of incorporating a wide range of experiences within the same general mechanism, corresponding to the wide phenomenal experiences humans typically report. In other words, the mechanisms are neutral or independent of any specific “content” of experience. Provided perhaps there is at least some content to activate the mechanism, any one of the possible contents would suffice to form an experience.

For instance, in global workspace theory, we might become consciously aware of any particular content once it has been broadcast to a workspace, being a specific but as yet not fully confirmed set of systems in the brain (Dehaene, 2014). In certain electromagnetic field theories, axiomatically-conscious fields integrate all information encoded in the electromagnetic activity corresponding to the field, such as the firing of neurons in a neural network (Jones, 2016). In integrated information theory (IIT), the specific pattern of embodied causal connections that contribute to the system’s phi value encode certain informational content, of which the system is then said to be conscious (Oizumi et al., 2014).

For theories that specify a general mechanism, it is plausible that when the mechanism integrates CS-defined abstract objects into their contents, it would translate into phenomenal consciousness of that object. In principle, dualist, panpsychist, or non-physical theories of consciousness could also invoke similar arguments, but may prefer to invoke other ontologies they assume as part of their core theory.

We can interrogate CS against this idea further using the discussion in 1.4. The primary phenomenal explananda were intangibility and non-location. As seen in the worked example (A8), abstract objects ("redness") can be invoked without firing the corresponding external senses ("seeing a red hat"), i.e. without being sense-perceived. In other words, the phenomenology of the abstract object firing on its own can be expected to feel intangible, particularly as we do not have touch receptors in the brain tissue. The brain does not directly sense the physical act of its neurons firing. We sense the result of neurons firing in some different manner, i.e. via some consciousness-generating mechanism as described above.

Non-location is equally natural to explain. The brain has an information encoding structure that encodes spatial locations in the physical realm as properties (CS abstract objects) connected to internal representations (CS abstract objects) of those physical realm objects. It appears not to do this for CS abstract objects, perhaps because there has not yet been sufficient evolutionary utility to do so. Abstract objects are typically felt as locationless simply because our brains do not typically encode locations against them. In principle, we could probably trick the brain into forming such associations and this might be one indirect phenomenological test of the theory.

There are also plausible accounts for the two speculative explananda. Abstract objects appear transparent/translucent for the same reason as they feel intangible – they can be activated absent from triggering colour perception at or near the input layer. They may feel "unchanging" because in any time instance when those nodes activate they are only and exactly what they are. In a stable network, each time they are activated again, they would be unchanged. While they do not exist "outside of time", it is understandable if felt "unchangingness" might convey such a phenomenology.

All of this phenomenology is real and can be accounted for as above, but it is misleading. I could, in principle, cut out the network substructure (brain cells) that corresponds to my redness and put it on the table. You could touch it and see its location. It would have colour (albeit typically not the same redness it encodes). It would be easily changed into something else, probably unavoidably so when you touch it.

Various structural features of the appendix model predict other phenomenal aspects that resonate with personal experience. Future work could derive such predictions in a structured manner and test them across several individuals' introspection as part of formal phenomenological research.

For example, the multiplicity of redness encoding described in 2.1 is a feature rather than a bug. Although the features possess a common core of informational associations within the network, they each pick up complementary facets of what it is to experience redness, in a way that reflects the different phenomenology, e.g. conceiving redness while seeing it is different from conceiving redness while hearing it and from invoking redness through an internal cogitation without a live connection to an externally-sensed redness. Variations in felt semantic clarity in these cases may correspond to levels of "contamination" in the network encoding (A10).

Other brief examples include increased phenomenal intensity in response to increased inputs as measured in network activity (A9), leading eventually perhaps to experiences of sensory overload, and how sensory data of an object can in principle be separated from its reifications and associations. For the latter, experienced meditators talk of seeing the visual inputs of, say, a glass without their brain categorising them as a "glass". A related experience can be evoked by reading the same word a hundred times until it feels wrong somehow (semantic satiation), no longer like the original word with its original meaning. Objects encoded in deeper network layers (in the sense of A2) may also feel "more" abstract or intangible in some sense. Indirect biological evidence for the latter might be seen in brain imaging data that show overlapping but distinct parts of the brain responding to concrete and abstract words (Binder et al., 2005).

Whether any of this amounts to "true" meaning, representation, or understanding is a topic for a future paper, noting diverse controversies and challenges (e.g. Millikan, 2017; Froese & Taguchi, 2019). Future work could explore in what ways phenomenal awareness of associations among *in cerebro* abstract objects and sensations of external physical objects might be sufficient to underpin an account for our felt experience of a particular word having a particular meaning. From there, it might be possible to build to the meanings of propositions and to at least one sense of the felt experience of "understanding".

3. Response to key philosophical objections

Section 2 explains how CS responds to the primary explanatory challenge in 1.2. The abstract objects that we use exist physically as subsets of the connectionist structures in my brain, typically derived based on interactions with multiple physical instantiations of that abstract object and counter-examples. This provides the metaphysical grounding for our claims. We thus generate knowledge of and operate with abstract objects without invoking any non-physical entities (and any felt non-physicality in the phenomenology is also accounted for). The sufficient consistency of a shared external environment and repeated interactions both with the environment and with each other generates sufficient, approximate consistency in our concepts to meet the needs of communication, being precise in some cases (requiring lots of consistency) and fuzzy in others.

This physicalist approach shares in common the rejection of universals that is typical of many flavours of nominalism proposed over the years. Examining the various counter-arguments and possible responses to those nominalist theories (see e.g. overview in Cowling et al., 2023) can help refine what is meant by CS and how an extended position on CS might form the basis of a formal grammar. As such, this section provides brief accounts of how CS would respond to five key applicable arguments, recognising that full accounts would require separate papers: (i) the referential opacity paradox, (ii) the identity of indiscernibles paradox given co-extensive properties, (iii) the problem of infinite regression, (iv) existence of non-physical concepts, and (v) the non-subjectivity of necessary truths. The last two in particular apply to other types of abstract objects far beyond the minimum claim of this paper, but a sketch of a response is provided nonetheless as an aid to its possible extensions.

3.1. Referential opacity paradox

One version of this paradox occurs when substituting one expression with another equivalent expression does not preserve the truth value of the statement (Bell, 1973). For example, "Lois Lane believes that Superman can fly" may be true, but "Lois Lane believes that Clark Kent can fly" may not be true, even though Clark Kent and Superman have the same referent (unknown to Lois).

This paradox is a problem only if words are assumed to have a singular universal meaning rather than a subjective meaning or if logics are intended in an infallible sense, unlike the pragmatic sense described in §2.3. In CS, words as internal objects only have subjective meanings, it just happens that a subset of us can work hard in particular environments to ensure we all adopt the same one. Lois has two abstract concepts in her head “Clark Kent” and “Superman”, which can be arbitrarily associated with different facts. She may in the future learn they in fact refer to the same person and the connections in her information processing system would be under pressure to update accordingly.

Another archetypal example: “I saw the Morning Star in the morning” is reasonable, but not “I saw the Morning Star in the evening”, even though the Morning Star refers to Venus which can indeed be seen in the evening. CS addresses this by recognising that the full state of an external object to be assigned to a word can be more than just its component physical parts, it can also be the time of day it is seen or any other part of its context. The core concept of Venus as a set of connected nodes would thus also link to at least two other concepts, Evening Star and Morning Star, where those other concepts have other nodes that must fire before they are activated, notably the time of day. Until someone knows all three of these are the same object, they may exist as three separate, only weakly connected objects in the network (e.g. connected by all being “small shining lights in the sky”).

3.2. Identity of indiscernibles

Some variants of nominalism, such as Class Nominalism, are argued to reach the unreasonable conclusion that two objects with exactly the same properties must be the same object or, symmetrically, that two identical sets of objects can never be said to have more than one property in common (Katz 1983; noting rejections such as in French 1989). A related issue comes up frequently in statistics or machine-learning in the context of under-determination and a similar treatment can resolve it without the metaphysical dilemma.

In the first case, this issue can only arise if we only know a subset of the objects’ full set of properties, where that subset happens to be co-extensive across both objects. In other words, it is an issue of imperfect knowledge to be solved by learning about the objects, e.g. through our sense organs and the measuring devices we construct to augment them. Indeed, if two objects have exactly the same properties for every imaginable set of properties, including their spatiotemporal boundaries and all their constituents and contexts, a fully knowledgeable account would conclude correctly that they are the same and should be collapsed into a single concept. That concept might now be known by two names, but without any harm done by such linguistic redundancy.

In the second case, it is true that if all instances of properties F and G appear in the exact same set of objects, the encoding mechanisms described in CS would not be able to tell them apart. They would typically encode a different property FG, being the combination of F and G. If F and G truly only ever come together, then encoding them in combination conveys no harm in the actions and experiences a cogitating agent might take based on FG. However, with the addition of more examples where F and G do not exactly overlap, it becomes possible to encode the differences and replace FG with the separate F and G properties. Importantly, the encoding mechanism never claimed that FG was not itself composite, as with any other encoded association, e.g. when words are linked to abstract objects those words are often composed of multiple sounds.

In a human sense, the initial FG state represents a midpoint of learning about the objects that exist, their properties, and their relationships. Since we can never conclude we have learned everything possible, a state of humility about midpoint learning and the possibility of future distinctions should not be a concern. Critiques against such intensionalism are more contingent concerns about constrained epistemology, rather than evidence against CS as an ontological position.[3]

3.3. Problem of infinite regression

The regression problem can be stated in several ways and is levied particularly against resemblance nominalism (Bradley 1893; better defended in modern variants like ostrich tropes, Giberman, 2022). CS has a different set of responses available to it than resemblance nominalism. To summarise one way of presenting the problem: if two objects are similar to each other in sharing properties F and G, we must also invoke the property of "sharing F & G", as well as "F", and "G". But then those objects now share three properties, and so we must invoke a fourth property: "sharing "F", "G", and 'sharing F & G' "[4]. Some nominalists might argue this infinite regress can be tolerated, observing that we tolerate infinities in other settings as well. However, regardless of that argument's merits or demerits, CS cannot apply it, since there is clearly not infinite space in the available connections in our brain.

There are three possible, non-mutually exclusive responses that could work for CS, where any would be individually sufficient. First, to reject the need for a dedicated "sharing" property; F and G simply are shared by virtue of both being associated with the relevant external object, with these connections grounded in the physical world primitives that define CS. An alternative would be to encode one abstraction for similarity/resemblance/sharedness

[3] If two properties appear only in the same set of objects, an intensionalist position would argue there is in truth only one property, whereas CS holds we can only know them as one property (i.e. an epistemological position, not an ontological one). There are more issues here than addressed in this brief account (see e.g. Yli-Vakkuri & Hawthorne, 2022), reserving full treatment to future work.

[4] Critiques of trope-only ontologies might focus on resemblance relationships as primitive and derivative at the same time. This critique is less relevant for CS, as the primitive ontological basis of abstract objects and their connections to external objects or in internal reasoning are defined in the physical world.

(or at least a manageable set of such abstractions), encoded in the usual way of being inferred from examples where similarities do exist. It could then be sufficient to keep referencing one or more of those encoded abstractions each time. The brain might also have various (perhaps imperfect) fail-safes for arbitrarily breaking a loop to the same abstract object that might otherwise repeat for infinite time. A third option is that some notions of similarity might be hard-coded into how a substrate implements a connectionist structure, such as a mechanical reaction to the minimum number of nodes between two other nodes, number of shared edges, shared depth in a network, number of consecutive activations, intensity of activations, or even fixed structures that extract notions of distance or number and work with them mechanically.

3.4. Existence of non-physical concepts

We are capable of conceiving, at least in some senses, of various non-physical concepts. This includes concepts that cannot be instantiated in any finite physical realm, such as infinity, or at the least could not be observed even in an infinite physical realm. It also includes concepts corresponding to objects that do not exist physically, such as unicorns, concepts that cannot exist physically in our 3D space, such as a 5D object, and perhaps even definitionally impossible concepts such as square circles in Euclidian space. If all abstract objects are extracted from real world sensory data, how can CS arrive at such non-physical objects?

A sufficiently sophisticated connectionist network could apply the principles in section 2 in a complex, differentiated, and well-sensed environment to produce many rich concepts based on repeated abstractions and associations from real world objects, but not necessarily the non-physical ones listed here. For instance, we could arrive at concepts and linguistic labels for people, groups of people, common behaviours, events, outcomes, and so on, creating space for notions like leadership, justice, peace, and so on. However, it might be possible to encode such non-physical concepts if the section 2 information processing capabilities are combined with others, such as generalisation, directed creativity, and consistency-checking.

An abstract object in a network is, like matrices in mathematics, both an entity (e.g. a particular typed structure of nodes & connections) and a rule-set for transforming inputs into outputs. In other words, they can encode transformations like “add one” as well as states like “being one”. In one sense they necessarily encode both simultaneously, but need not be operationalised in both senses simultaneously by the rest of the network.

Generalisation, in the sense of extrapolation and interpolation, is a series of transformation rule-sets. For instance, seeing and interacting with many groups of 1,2,3,4, and more objects in the physical realm could drive the encoding of numbers and simple arithmetic in a sufficiently sophisticated network, leading to notions of quantity, continuous scales, and extensive properties. Later, we might observe a new set of physical realm objects that only occur in groups of 2 and 4. The abstract object corresponding to the “adding one” transformation can operate on the abstract object corresponding to the observation of these new physical objects to generalise them, i.e. creating new internal abstract objects that correspond to those objects appearing in groups of 3 and 5, even if this is unrealisable in the physical realm. The concept of countable infinity can likewise be encoded in combining the same transformation with a “never-stopping” abstract object, such as might trivially be abstracted from physical realm observations like how a plan to exit at the first corner would never succeed when walking around a circular path. This is one of many ways the abstract object of infinity could be encoded in a finite network.

Networks most easily derive abstract properties from multiple objects, some of which instantiate the property and some of which do not. As a result, it is reasonable to ask how a network could come to encode a property that only exists in one object. Direct identity relationships are straightforward, e.g. the property of "being the Eiffel Tower" as achieved via one-to-one encoding (A7 has an example relating abstract objects to unique sound inputs). Singular properties might also be encoded by applying transformations (encoded in the usual direct manner from diverse physical examples) onto other objects similarly derived to arrive at something that could not be directly derived. For instance, the mathematical property "having no elements" applies only to the empty set - "zero" of anything is hard to see directly. However, if positive integer counting and arithmetic transformations are derivable as above, then counting down from a positive integer can lead to the "set with no elements" concept.

Connectionist networks are trivially capable of arbitrary free association. The toy model in the appendix can associate any input with any other input, just as different human languages can (in principle) arbitrarily assign the meaning of new words. As a result, felicitous mistakes in words and associations might lead to some concepts like unicorns which prove entertaining and hence persistent cultural products once they emerge. However, accidental mistakes alone feel an unsatisfying explanation for our full ability to generate new concepts. A fuller form of creativity might be possible with the brain drawing on internally manufactured or coincidental forms of “randomness” (in the sense of being sufficiently uncorrelated to the topic in question) to fuel a free associative process, directed by perceptions or generalisations of what might be entertaining or otherwise suitable for the goal at hand.

Consistency-checking could be delivered once certain trained structures can be functionally locked in at a certain point. The appendix (A11) illustrates contradictory inputs encoded a specific pattern in a network, differentiable from inputs that are consistent with the locked structure. Experiencing multiple examples of contradictions and consistencies across different settings could lead to the encoding of an abstract rule-set object for checking the consistency of one object against another rule-set, i.e. in activating a particular output node structure.

This paper does not need to claim that today's connectionist simulation models already have such capabilities. The extended claim is that CS principles might be sufficiently applied in the human brain to achieve such outcomes, as part of a more complex mechanism along with other brain features that might aid information processing, such as field effects (Jones, 2016) and wave harmonics (Atasoy et al., 2018), as well as implications for the connectionist rule-set due to mechanics encoded in subneuronal structures, neurotransmitters, or others. The minimum claim requires none of the capabilities discussed in this subsection.

3.5. Non-subjectivity of necessary truths

Some researchers have a notion that certain truths, particularly mathematical facts, remain true even if we have not yet discovered them or even if there were no conscious agents around to conceive of them. For instance, the infinitude of the primes or the number of times 2 goes into 10. The fact that we can confidently assert some mathematical statements as wrong suggests some objective benchmark to measure against. However, the abstract objects in CS are (in a sense) subjective: they exist internal to – and are drawn on internally by - each cogitating agent that had both capability and cause to derive them. This objection is particularly relevant for the case against mathematical psychologism (e.g. Frege, 1884; Balaguer, 2023), rather than the physical-realm grounded properties that are the focus of this paper. As such, a full account goes beyond the scope of this paper, but an outline can be presented.

First, we can observe that any statement in modern mathematics is never absolutely true in a standalone sense. A statement's truth is typically evaluated relative to a set of accepted axioms or at least relative to certain intuitions[5]. Second, while many axioms are sufficiently intuitive that they do not generally face challenge or alternatives, not all of them are. Instead, mathematicians sometimes design different sets of axioms with attractive features that they choose to deploy in different circumstances (e.g. Feferman et al., 2000).

A mathematical statement can be true in one set of axioms, false in another, and undecidable in a third (e.g. the Continuum Hypothesis under variants of ZFC set theory; see e.g. Cohen, 2008). Rather than there being a single objective mathematical benchmark, it is more accurate to say there are several in use already and likely many more useful ones that might exist. Gödel's two incompleteness theorems ensure the exploration of mathematics can never be a finished task, in that there will always be statements about natural numbers that are true but unprovable within the system and in that a system of axioms cannot demonstrate its own consistency (Smullyan, 1992). Even the fact that some mathematical statements may be true in all the current commonly-used sets of axioms does not guarantee they are true in all possible sets in the future. Mathematical objectivity is contingent and we do not invoke an ontologically fixed benchmark to assess it.

In fact, in a Yoneda lemma's sense (e.g. Riehl, 2017), mathematical objects are absolutely defined by their relationships with other mathematical objects. They are patterns, not substrates, just like the information encoded in CS-sense connectomes. When CS-sense abstract objects are described as subjective, it is with a specific meaning that can preserve much of the functionality objectivity that modern mathematics desires.

Whether or not an abstract object is accessible to a particular person depends on whether they have encoded it in their connectome. In this sense, some abstract objects are accessible to some people and not to others (until they have been taught it) – they are subjectively accessible. But the whole point of CS is that all of these abstract object encodings have a physical existence. With the right measuring devices, we can observe the physical structure of any other person's abstract encoding of redness, tracing its origin through the whole system,

[5] For instance, where non-axiomatic approaches are preferred, e.g. as reflecting certain mental activities within the practice of doing mathematics as potentially distinct from the discipline of formalising proofs. An extended form of CS would seek to account for how the human brain engages with abstract objects in doing mathematics, which may not always be exact parallel with axiom-first formalisation.

including from retinal cells whose physical structures confirm they would send a particular signal in response to light of a particular wavelength in a particular context. This is an objective exercise. Indeed, provided redness has been encoded in some information processing mechanism somewhere in the universe, there is an objective sense in which we can say it does exist.

All abstract mathematical propositions that have been encoded in someone's brain exist objectively. Some also happen to be "false" according to common axiom sets – which is no news to a maths teacher. Being false in this sense is no contradiction with existence. Maths can be made objective among a given set of humans in a very precise CS manner. First, have each target human's connectome encode a set of axioms with sufficient effort that we are confident the axioms are isomorphically identical between humans for our given purposes. Second, have each target human agree on these axioms as the objectivity-measuring device for mathematical propositions.

Under CS, different sets of mathematical axioms and propositions should be thought of as tools, like hammers or spanners. Just because a hammer is not accessible (within reach) of a particular person does not mean hammers do not exist. Did redness exist prior to it being encoded in the first information processing mechanism capable of doing so? No. But the potential for it did, encoded in independent, objective, external structures. Red physical realm objects existed, from which redness could in principle be consistently abstracted by senses attached to processing networks that share a general set of features. Likewise, there was a time on earth when hammers did not exist but the potential to assemble them did. "Hammer"-ness as a predicate is itself a CS object: a mereologically complex pattern (with embedded human-specific instrumentality) that can be implemented in common physical substrates.

What about the "unreasonable effectiveness" of maths in the natural sciences (Wigner, 1960)? Some positive propositions may be argued necessarily true of both the physical realm and a particular set of maths axioms. In the former, we can never be 100% certain of them (even if can be certain enough for practical purposes), as argued by various well-known sceptical arguments such as Popper's anti-verificationism (Popper, 1935) or doubt-fuelling thought experiments that are hard to rule out entirely (Sinnott-Armstrong, 2004). In the latter, we can be 100% certain of them in principle within those axioms, since we have full visibility of them. However, as a trivial result of the first claim, we can never be 100% certain that any apparent isomorphism between mathematical structures and the physical realm captures the full structure of the physical realm. We can, nonetheless, choose to keep refining maths variants that better match what we see in the physical realm and choose to have sufficient faith in the isomorphism to make predictions or take actions accordingly.

4. Is my red the same as your red?

In a CS framework, there are specific observable spatiotemporal objects that correspond to abstract objects like redness. In each human who knows this concept, the object is a series of derived connections in the brain, most likely in a neural network of some sort. Since there is one in your brain and one in mine, we can arrive at disciplined objective ways to consider whether yours and mine are the same.

A hierarchy of sameness can be organised in approximately four layers of alignment: (1) core concept; (2) associations; (3) network structure; and (4) all possible properties. Right up until the top of layer (4), it is technically possible for your red to be exactly the same as my red, but it becomes unlikely and challenging to confirm early in layer 1. None of this prevents us having, or working to reach, a level of semantic similarity along those layers that is sufficient for a particular purpose, such as agreeing on a colour for a new carpet.

For phenomenal sameness, we would additionally need to assume we have the same consciousness-generating mechanism that experiences given semantic contents in the same way, an assumption that physicalist consciousness theorists may consider plausible (see 2.4). Having made that assumption, it may be that there is a midway level of semantic similarity within the layers that is sufficient for indistinguishable phenomenal experiences between people - perhaps (1) and (2) alone suffice, perhaps a subset of (3) is also necessary – but exploring this lies beyond the current scope.

In layer one, our core concept of redness can be considered aligned if we both consistently communicate that redness is present in response to a sequence of candidate red-coloured physical realm objects. However, this quickly becomes unlikely as colours exist along very finely graded spectra and we may not exactly agree when orangey-red becomes orange. One resolution to this is to agree on only specific wavelengths to mean red. However, the perception of redness is not entirely reducible to wavelengths, as shown *inter alia* by colour contrast, inferred lighting, and texture optical illusions (e.g. Gomez-Villa et al., 2020; Shapiro & Todorović, 2017). Since it is at least impractical to test every possible combination of wavelength and context that might matter for perception, we are unlikely to prove our core concepts are exactly aligned. However, within this layer, it remains plausible that steps could be taken such that they happen to be exactly aligned for at least some narrow core concepts like redness, even if we could never know it for certain.

The second layer considers the arbitrary associations networks can make. Our concepts of redness can be considered more similar if they have the same associations. For instance, a monolingual French speaker associates a different sound to redness than a monolingual Chinese speaker. It would not take much to align such linguistic differences. Unfortunately, the brain can encode so many different associations from its historical experiences, that any abstract object would typically become heavily-laden with associations, even if we are capable of semantically separating many of these for discussions about the core concept. For you and me to have the same emotional response to redness, we also need the exact balance of evoked notions of forbidding, slowing, fire, love, and anger, for instance, which would likely require us to have had identical histories – implausible in most cases. Nonetheless, the encodings all exist *in cerebro* and are, in principle, examinable with enough technology.

Thirdly, we might want our CS objects to have the same pattern of nodes and edges within their network structure. While it may not take much for the objects individually to have the same structure, we also want them to have the same interconnections and relative positions with respect to the full connectome. As such, layer 2 could be fully met without layer 3 being met. Given genetic and environmental variation, and the sheer size of our connectomes, meeting the full extent of layer 3 may be possible in principle but deeply improbable.

Finally, in layer 4, we ask for all possible properties to be shared. In a trivial sense and definitionally, our redness objects are different in that they necessarily occupy different parts of space-time (their properties of location and specific constituents are not shared). Under CS, the absolute sameness of your concept of redness and mine is nomologically out of reach, even though we might discuss them happily and productively.

Conclusion

This paper has proposed "Connectionist Structuralism" (CS) as an account of the ontology of abstract objects. CS proposes that each abstract object a human being can draw on corresponds to a particular subset of their brain structure whose functionality is isomorphic to the relevant nodes and connections in a suitable connectionist network. The paper's minimum claim is that this proposal can account for all abstract object predicables of the type "is red" or "is a square", i.e. sensible properties. The extended claim is that the same core principles can also account for other abstract objects, where only the outline of a supporting account is sketched in the current paper. The paper has described how CS can support our core cognitive uses of abstract objects and account for our core phenomenal experiences of them. The outline of a response is provided to five common philosophical objections, to form the basis of future research establishing the position more comprehensively.

Future work could formalise the CS response in the context of challenges directed specifically to the use of abstract objects in a formal language or grammar, including issues around intensionalism (e.g. Yli-Vakkuri & Hawthorne, 2022). For instance, CS does not have to make the assumption of some other nominalist accounts that only first order logics can be allowed. By permitting certain second order logics CS can tolerate a broader range of potentially semantic statements, at the cost of some of the completeness desired by some logic theorists. Future work could also extend the CS account to a broader range of abstract objects to situate it within the mathematical fictionalism literature (Field, 2016), including issues around non-constructive proofs and Kripke's plus vs quus (e.g. Warren, 2020), within indispensability argument schemas (Panza & Sereni, 2016), and within relevant theories of consciousness and representation.

Future work could also test the account in a human setting in several ways. Most easily and ethically, phenomenological predictions could be derived in a structured manner from networks that approximate aspects of the human brain for testing across actual phenomenological accounts (examples in §2.4). It may also be possible to create learning environments in which abstract objects come to have an unusual phenomenology as predicted possible within CS, such as feeling tangible or located. Finally, the kind of brain mapping and manipulation already begun in many studies and discussed in Gómez-Emillson and Percy (2023) could unlock lab experiments where abstract objects can be found and altered, testing subsequent objectively-measured capabilities and phenomenal reports from patients.

References

Atasoy, S., Deco, G., Kringelbach, M. L., & Pearson, J. (2018). Harmonic brain modes: a unifying framework for linking space and time in brain dynamics. *The Neuroscientist, 24*(3), 277-293.

Balaguer, M. (2023). Fictionalism in the Philosophy of Mathematics. In Zalta, E., & Nodelman, U. (Eds). *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. https://plato.stanford.edu/archives/spr2023/entries/fictionalism-mathematics

Bell, J. M. (1973). What is referential opacity? *Journal of Philosophical Logic*, 155-180.

Ben-Haim, Y. (2014). Robust satisficing and the probability of survival. *International Journal of Systems Science, 45*:1, 3-19. https://doi.org/10.1080/00207721.2012.684906

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A". *arXiv preprint* arXiv:2309.12288

Berlemont, K., & Nadal, JP. (2022). Confidence-Controlled Hebbian Learning Efficiently Extracts Category Membership From Stimuli Encoded in View of a Categorization Task. *Neural Comput 34* (1): 45–77. https://doi.org/10.1162/neco_a_01452

Binder, J., Westbury, C., McKiernan, K., Possing, E., & Medler, D. (2005). Distinct Brain Systems for Processing Concrete and Abstract Concepts. *J Cogn Neurosci 17* (6): 905–917. https://doi.org/10.1162/0898929054021102

Bradley, F. H. 1893. *Appearance and Reality*. Oxford: Oxford University Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint* arXiv:2303.12712.

Burns, T. (2021). Classic Hebbian learning endows feed-forward networks with sufficient adaptability in challenging reinforcement learning tasks. *Journal of Neurophysiology 125*:6, 2034-2037.

Caporale, N., & Dan, Y. (2008). Spike timing–dependent plasticity: a Hebbian learning rule. *Annu. Rev. Neurosci., 31*, 25-46.

Carmichael, C. (forthcoming). Platonic Realism. In Mauren, A-S., & Fisher, A. (Eds). *The Routledge Handbook of Properties*. London: Routledge.

Chen, C., Jin, X., Jiang, B., & Li, L. (2019). Optimizing Extreme Learning Machine via Generalized Hebbian Learning and Intrinsic Plasticity Learning. *Neural Process Lett 49*, 1593–1609. https://doi.org/10.1007/s11063-018-9869-6

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in cognitive sciences, 15*(8), 358-364.

Cohen, P. J. (2008). *Set theory and the continuum hypothesis.* Courier Corporation.

Comesaña, J., & Klein, P. (2019). Skepticism. In Zalta, E. (Ed). *The Stanford Encyclopedia of Philosophy (Winter 2019 Edition)*. https://plato.stanford.edu/archives/win2019/entries/skepticism

Cowling, S., Rodriguez-Pereyra, G., & Giberman, D. (2023). Nominalism in Metaphysics. In Zalta, E., & Nodelman, U. (Eds). *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. https://plato.stanford.edu/archives/sum2023/entries/nominalism-metaphysics/

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis, 26*(4), 417-430.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts.* Penguin.

Derrida, J. (1967). *De la Grammatologie.* Paris: Éditions de Minuit.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., ... & Olah, C. (2022). Toy models of superposition. *arXiv preprint* arXiv:2209.10652. https://arxiv.org/abs/2209.10652

Ellwood, I. (2023). Short-term Hebbian learning can implement transformer-like attention. *bioRxiv preprint* 2023.05.31.543109. https://doi.org/10.1101/2023.05.31.543109

Feferman, S., Friedman, H., Maddy, P., & Steel, J. (2000). Does Mathematics Need New Axioms? *Bulletin of Symbolic Logic, 6*(4), 401-446. https://doi.org/10.2307/420965

Field, H. 2016. *Science Without Numbers, Second Edition*. Oxford: Oxford University Press.

Frege, G., 1884. *Der Grundlagen die Arithmetik*. Translated by J.L. Austin as The Foundations of Arithmetic, Oxford: Basil Blackwell, 1953.

French, S. (1989). Why the Principle of the Identity of Indiscernibles is not contingently true either. *Synthese 78*, 141–166. https://doi.org/10.1007/BF00869370

Froese, T., & Taguchi, S. (2019). The problem of meaning in AI and robotics: Still with us after all these years. *Philosophies, 4*(2), 14.

Giberman, D. (2022). Ostrich tropes. *Synthese 200*, 18. https://doi.org/10.1007/s11229-022-03494-4

Gómez-Emilsson, A., & Percy, C. (2022). The “Slicing Problem” for Computational Theories of Consciousness. *Open Philosophy, 5*(1), 718-736.

Gómez-Emilsson, A., & Percy, C. (2023). Don’t forget the boundary problem! How EM field topology can address the overlooked cousin to the binding problem for consciousness. *Frontiers in Human Neuroscience, 17*.

Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Bertalmío, M., & Malo, J. (2020). Color illusions also deceive CSNs for low-level vision tasks: Analysis and implications. *Vision Research, 176*, 156-174.

Heyting, A. (1956). *Intuitionism.* Amsterdam: North-Holland.

Himelright, J. (2022). Safety first: making property talk safe for nominalists. *Synthese 200*, 262. https://doi.org/10.1007/s11229-022-03714-x

Imaguire, G. (2022). What Is the Problem of Universals About? *Philosophica: International Journal for the History of Philosophy, 30*(1/2). https://doi.org/10.5840/philosophica20229135

Jones, M. (2016). Neuroelectrical approaches to binding problems. *The Journal of Mind and Behavior, 37*(2).

Katz, B. D. (1983). The identity of indiscernibles revisited. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 44*(1), 37-44.

Keysers, C., & Gazzola, V. (2014). Hebbian learning and predictive mirror neurons for actions, sensations and emotions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 369*(1644), 20130175. https://doi.org/10.1098/rstb.2013.0175

Lagani, G., Falchi, F., Gennaro, C., Amato, G. (2022). Evaluating Hebbian Learning in a Semi-supervised Setting. In: Nicosia, G., et al. (Eds.) *Machine Learning, Optimization, and Data Science. LOD 2021. Lecture Notes in Computer Science, vol 13164*. Springer, Cham. https://doi.org/10.1007/978-3-030-95470-3_28

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., ... & Dadson, S. (2021). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences Discussions, 2021*, 1-37.

Levine, J. (2011). On the phenomenology of thought. In T. Bayne, & M. Montague (Eds.) *Cognitive Phenomenology*. Oxford University Press.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

Meinwald, C. (n.d.) The theory of forms. *Britannica*. https://www.britannica.com/biography/Plato/Forms-as-perfect-exemplars

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information.* Oxford University Press.

Montero, B.G. (2022). Mathematical platonism and the causal relevance of abstracta. *Synthese 200*, 494. https://doi.org/10.1007/s11229-022-03962-x

Ngiam, J., Chen, Z., Chia, D., Koh, P., Le, Q., & Ng, A. (2010). Tiled convolutional neural networks. *Advances in neural information processing systems, 23*.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS computational biology, 10*(5), e1003588. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003588

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J Math Biol 15,* 267–273.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill, 5*(3), e00024-001. 10.23915/distill.00024.001

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural computation, 13*(6), 1199-1241.

Panza, M., & Sereni, A. (2016). The varieties of indispensability arguments. *Synthese 193*, 469–516. https://doi.org/10.1007/s11229-015-0977-9

Pitt, David. (2004). The Phenomenology of Cognition Or What Is It Like to Think That P? *Philosophy and Phenomenal Research Vol. LXIX*, No. 1.

Popper, K. (1935). *Logik der Forschung.* Vienna: Julius Springer Verlag.

Rettler, B., & Bailey, A. (2017). Object. In Zalta, E. (Ed). The Stanford Encyclopedia of Philosophy *(Winter 2017 Edition)*. https://plato.stanford.edu/archives/win2017/entries/object/

Riehl, E. (2017). *Category Theory in Context.* Aurora: Dover Modern Math Originals

Rodriguez-Pereyra, G. (2002). *Resemblance Nominalism. A solution to the problem of universals*. Oxford: Clarendon Press.

Seth, A., & Bayne, T. (2022). Theories of consciousness. *Nat Rev Neurosci 23*, 439–452. https://doi.org/10.1038/s41583-022-00587-4

Shapiro, A., & Todorović, D. (2017). *The Oxford Compendium of Visual Illusions.* United Kingdom: Oxford University Press.

Silverman, A. (2022). Plato's Middle Period Metaphysics and Epistemology. In Zalta, E., & Nodelman, U. (Eds). *The Stanford Encyclopedia of Philosophy (Fall 2022 Edition)*. https://plato.stanford.edu/archives/fall2022/entries/plato-metaphysics

Sinnott-Armstrong, W. (Ed.). (2004). *Pyrrhonian skepticism.* Oxford University Press.

Smith, D. W. (2011). The phenomenology of consciously thinking. In T. Bayne, & M. Montague (Eds.) *Cognitive Phenomenology*. (pp345-72).

Smullyan, R. M. (1992). *Gödel's incompleteness theorems.* Oxford University Press, USA.

Song, S., Miller, K., & Abbott, L. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nat Neurosci 3*, 919–926. https://doi.org/10.1038/78829

Stanley, K., Pugh, J., & Bowren, J. (2016). Fully Autonomous Real-Time Autoencoder-Augmented Hebbian Learning through the Collection of Novel Experiences. In *Proceedings of the Artificial Life Conference 2016* (pp. 382-389). MIT Press.

Sumner, R. L., Spriggs, M. J., Muthukumaraswamy, S. D., & Kirk, I. J. (2020). The role of Hebbian learning in human perception: a methodological and theoretical review of the human Visual Long-Term Potentiation paradigm. *Neuroscience & Biobehavioral Reviews, 115*, 220-237. https://doi.org/10.1016/j.neubiorev.2020.03.013.

Swanson, O. K., & Maffei, A. (2019). From hiring to firing: activation of inhibitory neurons and their recruitment in behavior. *Frontiers in molecular neuroscience, 12*, 168.

Trimmer, P.C., & Houston, A.I. (2014), An Evolutionary Perspective on Information Processing. *Top Cogn Sci, 6*: 312-330. https://doi.org/10.1111/tops.12085

Wang, F. Y., Zhang, J. J., Zheng, X., Wang, X., Yuan, Y., Dai, X., ... & Yang, L. (2016). Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica, 3*(2), 113-120.

Warren, J. (2020). Killing Kripkenstein's Monster. *Noûs, 54*(2): 257–289. doi:10.1111/nous.12242

White, B. (2022). A reductive analysis of statements about universals. *Synthese, 200*(1), 22.

Whittington, J., & Bogacz, R. (2017). An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity. *Neural Comput 29* (5): 1229–1262. https://doi.org/10.1162/NECO_a_00949

Wigner, E.P. (1960), The unreasonable effectiveness of mathematics in the natural sciences. Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959. Comm. Pure Appl. Math., 13: 1-14. https://doi.org/10.1002/cpa.3160130102

Wittgenstein, L. (1953). *The philosophical investigations.* Oxford: Blackwell.

Wong, E. (2019). Example Based Hebbian Learning may be sufficient to support Human Intelligence. *bioRxiv preprint* 758375; https://doi.org/10.1101/758375

Yli-Vakkuri, J., & Hawthorne, J. (2022). 'Intensionalism and Propositional Attitudes'. In Uriah Kriegel (ed.), *Oxford Studies in Philosophy of Mind Volume 2.* Pages 114–174. Oxford: OUP. https://doi.org/10.1093/oso/9780192856685.003.0005

Zappacosta, S., Mannella, F., Mirolli, M., & Baldassarre, G. (2018). General differential Hebbian learning: Capturing temporal relations between events in neural networks and the brain. *PLoS Comput Biol 14*(8): e1006227. https://doi.org/10.1371/journal.pcbi.1006227

Appendix

This appendix sets up and describes the results of a simple, fully local, self-supervised Hebbian network system, called CS1 for this example. The system encodes, in a CS sense, the abstract objects corresponding to two colours and two types, associating words and images with unique subsets of connections such that words alone evoke their images and vice versa.

A1. System set-up and notation

There are input and internal nodes for this example, with each being activated (valued 1) or deactivated (valued 0) at a particular point in time. Each individual connection joins exactly two nodes, but each node can have multiple connections. Connections can also be deactivated (valued 0) or activated (valued 1) at a point in time. The network system, in terms of the edges and nodes that exists, is fixed throughout. There are 68 nodes and 118 edges in the full structure (mapped in Figure 1), introduced in stages below to aid following the process. The state of the system at any point in time is a list of which nodes and connections are activated.

Unique nodes (circles in the fig. 1) are denoted with a single capital letter followed by a subscript number and, for the direct sensory input columns, additionally with a preceding superscript number (denoting depth of exact copies of input data – see below), e.g. ${}^{1}Z_{1}$ or X_{1}. If no subscript is used, all nodes of that letter (and superscript if included) are indicated. Connections (single- or double-ended arrow lines in fig. 1), are denoted by the two nodes they join, using lower case letters, e.g. $z_{1}x_{1}$.

A2. Node and edge activation

Three types of connection exist within the network:

- "Fixed edges" are always one way and always convey a strength of 1 to their end node when their start node activates. These activations and conveyances are functionally instantaneous from the perspective of the model.
- "Oneway trainable edges" are all initialised as "untrained", but can become "trained" instantly if their start and end nodes fire independently. Once trained, they stay trained. In their untrained state, they cannot activate. In their trained state, they can activate only if their start node is active, in which case they convey a strength of 1 to the end node (in the direction of the arrow in fig. 1). Edge and any subsequent node activation requires one time-step.
- "Twoway trainable edges" are the same but can convey their strength (either 1 or) 2 in either direction, i.e. either node can then convey the stimulus strength to the other node once the edge is trained.

Input nodes become active based on some external stimulus, such as an inward connection from a sense organ. Input nodes are the first layer of a network (layer 0). To aid notation, all input nodes and their exact successor nodes have a preceding superscript denoting their depth, defined as the minimum number of connections required to reach the input node. Similarly to aid notation, input nodes are grouped by the sensory mechanism activating them.

- This example has two sets of sensory mechanisms: visual (entering pixel inputs to the network via 4 ${}^{0}V$ nodes) and auditory (entering sound inputs via 4 ${}^{0}A$ nodes).
- In this set-up, the raw input data are processed in a range of different ways, each starting from the raw data. As a result, the sensory mechanisms transmit identical signals down to four layers in the network along fixed edges in two columns: ${}^{1}V$-${}^{4}V$ and ${}^{1}A$-${}^{4}A$, e.g. ${}^{0}V_1$ has a fixed edge out to ${}^{1}V_1$, which in turn has a fixed edge to ${}^{2}V_1$.

Internal nodes become activated if the intensity of their inward connections, aggregated according to a defined function, exceeds a pre-specified and fixed activation threshold. In this case, all nodes have an activation threshold of 1 except P and B which have a threshold of 2.

As an aside, output nodes operate the same as internal nodes but their outward connections activate some external mechanism, such as a motor action or decision output. This model contains no output nodes, as none are needed for the phenomena to be demonstrated, but they could be added trivially to operate off any subset of the nodes and edges in the model in an extended setup. For instance, a single output node with inbound connections from all the desired set of xy edges and/or X Y nodes is set up to fire only when all of xy, X and/or Y are activated. All possible single output node variants could be present in the initial structure, with frequency and utility based rules driving selection and pruning, where utility levels could be defined by some set of internal network states. Other options are possible and more natural, but a full treatment lies outside the scope of this paper.

A3. Deactivation rules

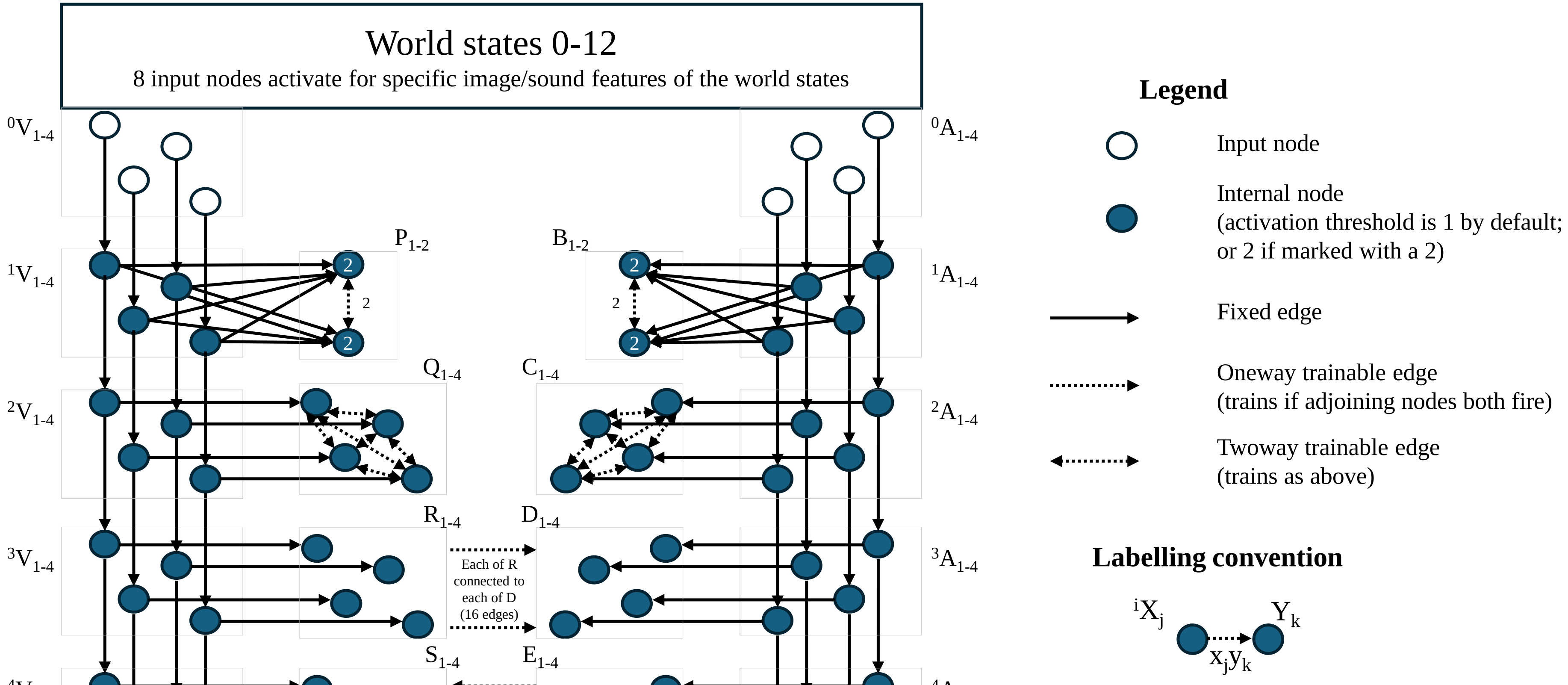
Nodes remain in an activated state until explicitly deactivated. Any initial input node activation triggers an internal timing mechanism, implemented in the substrate of the overall mechanism, that allows network activity for one timestep and then deactivates all nodes and edges in the second timestep. In other words, nodes may both fire instantly via fixed edges and after one timestep via trained edges, before the network is reset.

A4. Stability

The network system, in terms of edges and nodes, is fixed throughout, but as inputs are received and transmit through the network, the set of edges that are trained can change, which can alter the response of the network to future inputs. The system state at any given time is a list of which nodes and edges are activated. The “trained structure” at a point in time can be fully captured by the list of trained edges. The system is stable if there is no possible combination of input node activations that would trigger a cascade of states that further changes the trained structure.

This particular example has been designed to stabilised rapidly without path dependence, i.e. all possible sequences of input stimuli lead to the same stable trained structure. Since the network iterates only for 1 timestep per input state and trainable edges always take 1 timestep to fire, an edge can only ever be trained when both nodes are activated via fixed edges, which only occurs for direct input node information transmission. Because all combinations of input states occur eventually (A5), we also know that all possible such edges end up trained in the stable state, the only difference is the order they are trained in.

Fig. 1 CS1 model structure


World states 0-12
8 input nodes activate for specific image/sound features of the world states
⁰V₁₋₄
¹V₁₋₄
²V₁₋₄
³V₁₋₄
⁴V₁₋₄
⁰A₁₋₄
¹A₁₋₄
²A₁₋₄
³A₁₋₄
⁴A₁₋₄
P₁₋₂
B₁₋₂
Q₁₋₄
C₁₋₄
R₁₋₄
D₁₋₄
S₁₋₄
E₁₋₄
2
Each of R connected to each of D (16 edges)
Each of E connected to each of S (16 edges)
Direct visual column
Direct auditory column
ⁱXⱼ
Yₖ
xⱼyₖ

Legend

○ Input node

● Internal node (activation threshold is 1 by default; or 2 if marked with a 2)

Fixed edge

Oneway trainable edge (trains if adjoining nodes both fire)

Twoway trainable edge (trains as above)

Labelling convention

- X, Y = node group name (marked as square grouping the nodes)
- j,k = specific node in group
- i = depth from input nodes (only used to differentiate node groups in direct sensory columns)
- x_jy_k = edge for two upper-case equivalent nodes (fixed/trained edges convey strength of 1 by default; or 2 if marked with a 2; untrained edges convey 0)

A5. Toy universe set-up

The external environment to the CS1 system contains four active world state variants with both image and sound (1. blue square, 2. red square, 3. blue rectangle, 4. red rectangle), as well as eight variants for sound without image and image without sound. There is a default blank world state where none of the CS1 input nodes would fire, labelled 0. Only one variant may be displayed at any one time and for only one timestep. There is then a pause of one or more timesteps before the next state can occur, ensuring CS1 is fully deactivated at the start of each variant-sensing session. The 12 active variants and the blank state may nonetheless appear in any order and any number of times, including repetitions, provided each of 1-4 is displayed at least once, ensuring CS1 becomes stable from that point onwards.

CS1's eight input nodes always respond consistently to the relevant world state, according to the rules in Table 1. Relative to the speed of operations in the model, the image display, sound sequence, and their perceptions via the input systems are all functionally instantaneous. The squares are one pixel large and always appear in the location corresponding to the first pixel in the visual system. The rectangles are all two pixels large and always appear in the location corresponding to the first and second pixels combined.

Table 1. Environment states and input node responses

	Environment state		**Visual system** – 2 pixel input; 2 input nodes per pixel, which fire if they sense red or blue respectively in that pixel				**Auditory system** – 4 input nodes which fire if they hear blue, red, square, or rectangle respectively			
#	Image	Sound	$^0V_1$	$^0V_2$	$^0V_3$	$^0V_4$	$^0A_1$	$^0A_2$	$^0A_3$	$^0A_4$
0			0	0	0	0	0	0	0	0
1	🟦	Blue square	1	0	0	0	1	0	1	0
2	🟥	Red square	0	1	0	0	0	1	1	0
3	🟦🟦	Blue rectangle	1	0	1	0	1	0	0	1
4	🟥🟥	Red rectangle	0	1	0	1	0	1	0	1

** States 5-8 are images without sounds (V activates as above; A is all 0s); 9-12 are sounds without images (A activates as above; V is all 0s).*

A6. Layer 1: Encoding shape from visual input

Four fixed edges convey activations of strength 1 each outward from ^{1}V into a single node P_1, with a further identical four fixed edges connected to an identical single node P_2. P nodes have an activation threshold of 2 and a twoway trainable edge between them that conveys a strength of 2 once trained.

As can be seen in Table 1, P_1 and P_2 activate and train their connecting edge only when images of rectangleness are present, i.e. only in states 3, 4, 7, & 8, as in all other states only one input node is activated and P do not exceed their activation threshold. The trained edge p_1p_2 encodes rectangleness in line with the P1 definition, with squareness delivered by the inverse.

A symmetrical structure exists for the auditory column, which encodes "any active sound state", i.e. it activates when any of 1-4 or 9-12 occurs. This particular example has no further need of this pattern, but such nodes can be useful for other types of information processing.

A7. Layer 2: Encoding colour based on visual input alone

Four fixed edges connect ^{2}V into four nodes $Q_{1\text{-}4}$. Q is fully interconnected with six twoway trainable edges that convey a strength of 1 once trained. Following the logic of A4, reading off table 1 for the equivalent V nodes shows only two of the Q edges will become trained: q_1q_3 from state 3 or 7; q_2q_4 from state 4 or 8. Table 2 shows how the stable end-state of Q responds to each of the four unique active input states in the visual column. As can be seen, the full subsystem displayed is cleanly separated between blue and red input states, using pale colour shading as a visual aid. Blueness is encoded in the trained edge q_1q_3; redness in q_2q_4.

Table 2. Environment states and Q subsystem responses in stable state [iv]

	Environment state		**Q nodes** *(fully interconnected)*				**Trained twoway edges**		**Untrained twoway edges** *(will never fire, shown as visual aid)*			
#	Image	Sound	Q_1	Q_2	Q_3	Q_4	q_1q_3	q_2q_4	q_1q_2	q_1q_4	q_2q_3	q_3q_4
1	🟦	Blue square	1		1		1					
2	🟥	Red square		1		1		1				
3	🟦🟦	Blue rectangle	1		1		1					
4	🟥🟥	Red rectangle		1		1		1				

** 0s left blank for ease of reading.*

The auditory column has a symmetrical structure, C. C has a stable set of four trained edges, of which a unique combination of three is activated in response to each one of the state inputs 1-4. C is not needed further in this example but may be useful in other settings.

A8. Layers 3 & 4: Associating colour and sound

Layer 3 is made up of R and D, as four-node duplications via four fixed edges of ^{3}V and ^{3}A respectively, such that $^{3}V_i$ drives R_i etc. R and D are fully connected via 16 oneway trainable edges, pointing from R towards D. There are no R-to-R or D-to-D edges. The stable state of trained edges can be read off from the equivalent Table 1 pairs: r_1d_1, r_1d_3, r_2d_2, r_2d_3, r_3d_1, r_3d_4, r_1d_4, r_2d_4, r_4d_2, and r_4d_4. Layer 4 of S and E is the same as layer 3, except the 16 oneway trainable edges now point from E (off the auditory column) towards S (off the visual column). The trained edges are the same as layer 3, except replacing r with s and d with e.

The edge activation map for all possible world states and several hypothetical states is shown in Table 3 and node activations in Table 4. As marked with the pale blue/red shading and black-outline rectangles, we can see unique sets of edges which uniquely identify the necessary properties and categories, when both sound and images are input (states 1-4). Blueness is r_1d_1 s_1e_1 (r_3d_1 s_3e_1 nearly qualifies but for the absent activation in state 5). Redness is r_2d_2 s_2e_2 (r_4d_2 s_4e_2 would qualify but for state 6). Rectangleness is r_3d_4 s_3e_4 r_4d_4 s_4e_4, and squareness by its inverse. These edges also activate when just the sound is heard or just the image is seen in real world states. The feature activations carry across with only mixed success to the hypothetical states. For instance, r_1d_1 & s_1e_1 fail to activate in state 17. Visual colour is encoded via the first pixel only, which works in all actual world states but not all imaginable states. As originally, the network is not designed to deliver perfect generalisation.

The edges connect to and activate nodes with an intimate relationship to the relevant original sound and image inputs, such that the sound "evokes" the image and vice versa, as illustrated in the example of D_2. D_2 can only activate (see Table 4) if at least one of the following three things happen: if the word red is heard, if a red square is seen, or if a red rectangle is seen. Only the first of these is linked solely via fixed edges from a sense input (${}^0A_2$), the others occur due to the set of trained edges above. However, D_2 does not activate ${}^0A_2$, which responds only to actual sounds in the world state. Hence, the right sound processing node is evoked by images alone but not as vividly in a total network information sense as if it is actually heard. Moreover, when it is actually heard, the sound activates the relevant nodes down the full A column, whereas it only appears in one layer of A when evoked (layer 3).

Table 3. Trained edge activation by input states in layers 3 & 4

#	Image	Sound	r_1d_1 s_1e_1	r_1d_3 s_1e_3	r_2d_2 s_2e_2	r_2d_3 s_2e_3	r_3d_1 s_3e_1	r_3d_4 s_3e_4	r_1d_4 s_1e_4	r_2d_4 s_2e_4	r_4d_2 s_4e_2	r_4d_4 s_4e_4	# live
1	🟦	Blue square	1	1		1	1		1				5
2	🟥	Red square		1	1	1				1	1		5
3	🟦🟦	Blue rectangle	1	1			1	1 □□	1	1		1 □□	7
4	🟥🟥	Red rectangle			1	1		1 □□	1	1	1	1 □□	7
5	🟦		1	1					1				3
6	🟥				1	1				1			3
7	🟦🟦		1	1			1	1 □□	1				5
8	🟥🟥				1	1				1	1	1 □□	5
9		Blue square	1	1 □		1 □	1						4
10		Red square		1 □	1	1 □					1		4
11		Blue rectangle	1				1	1 □□	1 □□	1 □□		1 □□	6
12		Red rectangle			1			1 □□	1 □□	1 □□	1	1 □□	6
Hypothetical states (do not exist in the world states)													
13		Blue	1				1						2
14		Red			1						1		2
15		Square		1		1							2
16		Rectangle						1 □□	1	1		1 □□	4
17	🟦						1	1 □□					2
18	🟥										1	1 □□	2
19	🟥	Blue square	1	1	1	1	1			1			6
20	🟥	Blue rectangle	1		1	1	1	1 □□	1	1		1 □□	8

** 0s left blank for ease of reading; stable state untrained edges will never fire so can be excluded.*

The discussion so far is sufficient to motivate a worked example that delivers on the capabilities claimed in P2. However, there are further interesting structures in the trained network, which build intuition and provide further examples for P6 and P7.

A9. Additional structure: Number of activations

From the last column of tables 3 and 4 and a definition of intensity from the perspective of the network as increased nodes or edges activated, we see that any increase in input node activation, holding other inputs fixed, leads to increased intensity, not just trivially in the A and V columns, but also in the internal trained edges.

Rectangles evoke more activations than other inputs. For states involving images, this is expected, since it corresponds to two activated pixels as differentiated from one activation for squares. However, this increased intensity carries over even when just the sound is heard, even though sounds have the same input intensity (2 per layer, as in Table 1). This occurs because the network structure encodes cross-associations, so some of the sound-related edges encode the image-related size of rectangles.

In extreme cases or for overly-connected, poorly parameterised networks, the mutual activation of many nodes (runaway excitation) or too many weights having extreme values (network saturation) can mean a network loses the ability to produce differential states on differential inputs, either in its activations or in its plasticity, e.g. when the whole network fires at once. In the above intensity sense, this might resonate with human notions of certain sensory overload or epileptic seizure experiences. In simulation networks, techniques like normalisation, inhibitory edges, and anti-Hebbian learning rules are used to prevent this dysfunctional outcome.

Table 4. Node activation by input states in layers 3 & 4

					Visual column: X = fired due to auditory column alone, i.e. cross-fired via a oneway trained edge				Auditory column: X = fired due to visual column alone, i.e. cross-fired via a oneway trained edge				# live in S&D
#	Image	Sound	V_{any}	A_{any}	S_1 =p1 bl.	S_2 =p1 rd.	S_3 =p2 bl.	S_4 =p2 rd.	D_1 =blue	D_2 =red	D_3 =sq.	D_4 =rect.	
1	🟦	Blue square	1	1	1	X	X		1		1	X	6
2	🟥	Red square	1	1	X	1		X		1	1	X	6
3	🟦🟦	Blue rectangle	1	1	1	X	1	X	1		X	1	7
4	🟥🟥	Red rectangle	1	1	X	1	X	1		1	X	1	7
5	🟦		1		1				X		X	X	4
6	🟥		1			1				X	X	X	4
7	🟦🟦		1		1		1		X		X	X	5
8	🟥🟥		1			1		1		X	X	X	5
9		Blue square		1	X	X	X		1		1		5
10		Red square		1	X	X		X		1	1		5
11		Blue rectangle		1	X	X	X	X	1			1	6
12		Red rectangle		1	X	X	X	X		1		1	6
Hypothetical states													
13		Blue		1	X		X		1				3
14		Red		1		X		X		1			3
15		Square		1	X	X					1		3
16		Rectangle		1	X	X	X	X				1	5
17	🟦		1				1		X			X	3
18	🟥		1					1		X		X	3
19	🟥	Blue square	1	1	X	1	X		1	X	1	X	7
20	🟥	Blue rectangle	1	1	X	1	X	X	1	X	X	1	8

** Excl. fixed edge activated only nodes, since responses unchanged between initialisation and stable system states. 0s left blank for ease of reading*

A10. Additional structure: Negative squareness

As seen earlier, squareness is only uniquely identified by the absence of activated edges, unlike the other three abstract objects encoded in this layer. This negative encoding reflects the reality of this toy universe: there are no active images in the world states without the square component. This is not a concern for utility. If a larger network wanted to operationalise squareness explicitly, inhibitory neurons could be included in its architecture to construct positive activations from absent inputs. However, there is more to explore.

In English, squareness and rectangleness denote a more specific set of properties than they do in the toy universe. Rectangleness most directly "adds a second pixel", which could have several interpretations in a more sophisticated universe. It can turn certain squares into rectangles, i.e. its original motivation here, but it also "adds one to one to make two" (but encodes no other insights about arithmetic). In this case, it also implies "the shape of (all) rectangles fully contains the shape of (all) squares" - this is a fact about the toy universe that is true only in some cases in our universe. The words, however, do not encode this hierarchical relationship. The sound inputs cleanly separate rectangles from squares. As a result, when sounds alone are heard (states 9-12), the trained edge response more cleanly separates rectangleness from squareness and vice versa (see grey-outline shapes added to table 3), with such additional separations only valid when comparing sounds-alone against sounds-alone, i.e. in isolation from other possible state-responses.

When images alone are heard, the association between image and rectangle results in some edges activating for rectangle inputs that do not meet this clean separation criteria, notably r_3d_1 & s_3e_1 in state 7 (blue rectangle image alone) and r_4d_2 & s_4e_2 in state 8 (red rectangle image alone). However, it is reasonable to ask which is nearer to the truth of the toy universe. The separation may be "cleaner" in the auditory inputs, but at the cost of discarding some information encoded in the messier structure of the external world states.

When both images and sounds are perceived at the same time, this effect multiplies. We can define a contaminated edge as an activated edge in a particular world state that, if it were deactivated, would result in a clean separation of either shape or colour. Each of the world states 1-4 has exactly one contaminated RD & SE edge-pair for shape and one for colour, which can be seen by examining table 3. For instance, for blue square, r_1d_4 & s_1e_4 contaminate otherwise clean shape separation. However, the network is robust to this contamination: other trained edges were already sufficient to encode the abstract objects of interest (A8). In a more sophisticated set-up, more diverse world states and a larger network with more sensitive sense organs could differentiate these structural elements.

A11. Additional structure: Contradictory inputs

Hypothetical states 19 and 20, not available for system training since they do not exist in the world states, show the stable system response if a red image accompanies a blue word: cross-activation of contradictory colours. Such cross-activation is not in itself a problem or a failure; the network is simply mechanically triggering nodes in response to inputs. As with elsewhere, output nodes could be run off such system states to trigger action to do something, such as investigate the contradiction. However, if the model were still open to training, the cross-activation would learn new nodes and break the separation of property encoding shown so far. Such issues are often mitigated in more sophisticated simulations by using gradual, frequency-based training and allowing for edge weakening, assuming that correct associations will be more frequently encountered.