Mathematical Approaches to Cognitive Science

Nivedita
Thu 6 Mar 21

> It is in this gesture of 'going beyond', to be something in oneself rather than the pawn of a consensus, the refusal to stay within a rigid circle that others have drawn around one-it is in this solitary act that one finds true creativity. All others things follow as a matter of course.
>
> - Alexander Grothendieck

ABSTRACT. In this paper, we present different formalisms aiming at theorising cognition. We introduce Gärdenfors' conceptual spaces as an extension to the symbolic and connectionist approaches to representation. We talk about mathematician Misha Gromov's Memorandum Ergo [Gro18], and finally, the most modern and rigorous approach to use topological and higher categorical concepts in neuroscience by Matilde Marcolli, Yuri Manin and Doris Tsao.

KEYWORDS: mental representation, geometry, homotopy, higher category theory

1. Why Theory?

With cognitive science having been birthed from a multi-disciplinary approach with philosophy, computer science, psychology, neuroscience, linguistics and anthropology, where does mathematics fit into this picture? Does it come from the logical underpinnings in philosophy, or complexity of algorithms and linear algebra in machine learning, or from the probability theory we use psychology and Fourier and data analysis techniques in neuroscience? This article aims to explore how mathematics should not just be an implicit tool in cognitive science but an explicit pillar in its foundational explorations.

The explanatory role of mathematics in natural science is discussed a lot in philosophical circles; how can the connection between the mathematics we develop and the universe be accounted for?[Man18]. Mathematics has been proven to be very successful, especially in the physical sciences, in which modern mathematical theories can predict experiments with error less than 10^{-10}. Further, mathematics helps us in understanding both the structural foundations of what it is being used to study, rather than just empty words linking together observed correlations (imagine physics without mathematics!). Mathematics also lends a quantitive and objective understanding, which can connect together the various observed correlations into a single foundational theory.

2. Introducing different approaches

There are two primary goals of cognitive science: explanation and construction. The former focuses on formulating theories about different aspects of cognition and the mind, whereas the latter aims at building systems that can simulate aspects of human behaviour, such as chess playing software or communication bots. How different concepts are learnt, stored (represented) and accessed lies at the heart of human cognitive abilities. There have been two fundamental ways of modelling how the mind works, specifically how it represents concepts: the symbolic models and the connectionist ones. The symbolic models are based on the assumption that cognitive systems are inherently like Turing machines; that is, there is a set of predicates with fixed denotation and representations are based on logical and syntactical operations. There are many problems with this approach, the general one being where do the predicates come from, and how do we learn the meaning of a predicate? For a computer program, the programmer gives meaning to the symbols and assigns values to initial parameters; for the modelling of a cognitive system symbolically, some basic concepts are taken for granted to exist, which are then acted on by operators and quantifiers to create more concepts. In the 1970s, the methodology of connectionist models gained importance. Here, the association between different information is what bears representations. The representations in this framework are distributed: they are in some way stored in connection weights that are dynamically changing in response to input stimuli. However, it is challenging to understand what is represented in the network, what each node represents, what each edge represents, and so on. Another drawback of this is that the learning of computational models is very slow in general because it requires many iterations of training.

For many years these were considered to be the only two approaches complementary to one another, but there are cognitive phenomena for which neither of these two approaches offers appropriate modelling tools.

Topology is one of the most widely applied branches of mathematics. With applications ranging from condensed matter physics to data analysis, algebraic topology, and homotopy theory, more generally is finally being applied to neuroscience. There has been a vast gulf between the way we probe the structure of neural networks in the human brain and how we try to understand the mind's emergent functions, which many disciplines from philosophy to cognitive science try to bridge. One of the approaches is to assign the most complex level an ontological existence to tackle these highly complex problems, such as human consciousness or dynamics of human societies. Further, especially in philosophical investigations, we try to develop a theory about theories (theories of consciousness or something else); this is what is usually called a meta-theory. Just as in logic, where a statement about propositions is a higher-order proposition, in processes, a meta-process is a global process linking together processes. The objective existence of ontological meta-levels is based on the emergence of higher levels of reality which generally can not be tackled easily by purely reductionist methods.

We first present two perspectives on the geometry and topology involved in cognitive processes, and then a third which unifies them with many different perspectives from a variety of subjects: information theory, homotopy theory, etc.

2.1. Conceptual Spaces

Gärdenfors' idea of conceptual spaces [Gär00, Gär14] is based on the idea of similarity as a foundation to concepts. In this theory, we assume that the input stimuli form a geometric space and is represented in the same way in the brain. A natural property is a convex region in such a space. The notion of convex regions (all points on a line joining any two points in a region are within the region) gives meaningfulness to the concept of betweenness. The natural properties defined by these convex sets are a particular case of the concepts. A property is a domain, while a concept can be thought of as a collection of many distinct domains. The older two theories of symbolic approaches and connectionism do not necessarily differentiate between properties and concepts. To think of it in terms of language, properties are like adjectives, and concepts are like nouns. For example, the concept of apple has several properties like redness, roundness, sweetness etc. So, let us consider a preliminary construction of a conceptual space. Distinct objects in the physical world are the points (this apple, my phone), then there are domains of properties (being red, being round), then we consider bunches of these domains, keeping the information of the intersections intact, these give rise to concepts (like an apple or a mobile phones etc.). An essential additional structure that we do not wish to ignore is the correlations between different properties. For example, the sweetness of apple in the taste domain strongly relates to the sugar content in nutritional content, whereas the shape and taste are not so strongly correlated. The geometric structure of our conceptual space should include the information of these correlations. Similarly, verbs or actions are like paths or vectors in the conceptual space. The more abstract the concepts, such as natural number two, the higher the dimensionality of the domain representing the concept and the more delocalised it is.

This gives us the definition of a *concept* as a collection of convex regions (natural properties) in different domains along with the information about the correlations of properties, if any. There can be a Voronoi tessellation of the conceptual space, where each cell has a prototype. We further endow our spaces with metrics (distance functions) to give a geometric underpinning to the concept of similarity. The metric defined on points naturally gives a metric defined on convex domains, given by minimising the distance between arbitrary points of the two domains.

There are many philosophical considerations that arise with conceptual spaces as the theory of how knowledge is represented. Conceptual spaces can be used to discuss problems in linguistics and philosophy, including vagueness in meanings, grading in properties, and hence assigning graded membership of concepts in various domains. The relation of x *is identical to* y is very straightforward, it is reflexive $(a \sim a)$, symmetric $(a \sim b \Rightarrow b \sim a)$and transitive $(a \sim b \wedge b \sim c \Rightarrow a \sim c)$. The paradoxes of identity like the 'Theseus' ship' of reconstructing the ship plank by plank and asking when does the identity break can be given a uniform solution by reevaluating the relevant notion of identity.

In the example of the ship, the transitivity property breaks; when replacing planks, initially, the ship at each step is identical to the ship at step zero but far along in the process, the transitivity breaks. We can restrict the claims of similarity in *relevant* respects and give a higher structure to similarity by recording the information of how are two things similar and further on. This leads to homotopy types of similarity, as talked about in the next section. Many psychological studies and research have shown that when asked about the similarity of two things, we only consider a subset of all the property domains, and the answer is generally context-dependent. This implies that we should consider any discussion to be grounded in some (finite-dimensional) conceptual spaces.

Just like there are higher orders in similarity, there can be higher orders in vagueness too. When we can not point to a sharp line distinguishing between borderline cases, the geometry of the space helps to have a smooth transition possible. The mathematical idea of fuzzy set theory has been used to concretise methods of measuring quantitatively to what degree a given item falls within a particular concept. This tells us that we should define identity using the metric on the conceptual space, but not ask for precisely equal points, but only differ by a distance less than some cut-off. This also works for concepts using the metric described above.

2.2. The Ergo Project

Many mathematicians have tried targeting the hard questions about neuroscience and the mind using abstract mathematics. One of the most famous approaches is the *Ergo* approach by Misha Gromov. We all agree that the sole source of thoughts and consciousness is our brain; there is even evidence that this idea was discussed in ancient Egyptian papyruses. Our brain reconstructs the entire universe around us, what we call representations, from space to time, mathematics to art. We use many different verbs interchangeably to describe how the mind and brain are related, such as the mind emerges, is caused/generated, or is controlled/run by the brain. However, we at are not able to concretely describe the [Brain]$\rightarrow$[Mind] arrow.

Gromov defines the mind to be a finite connected graph - a network of ideas and concepts. It can be further divided into two substructures, which Gromov calls the ergo-brain and the ego-mind. The ego-mind is something a person sees as their personality. It consists of what we often perceive as our conscious self: consisting of all our thoughts, feelings and passions. Since this part is accessible to the self and most of it is also expressible in human language. This part has developed with the aim of survival.

On the other hand, the ergo brain is a piece of abstract semi-logical machinery that tackles ideas like finding things interesting, meaningful, funny, aesthetically pleasing, boring, etc. These attributes are not objective or directly linked to evolutionary benefit, yet they are universal. This ergo system is what makes goal free learning possible in us. We now describe a few pieces of evidence that are used to back the ergo brain machinery, which gives rise to the non-pragmatic human traits. Firstly, learning languages (we have the necessary substrate or hard-wiring in the brain for language acquisition, but the learning is more than goal oriented). The process is dynamic structured learning with constraints from the network architecture of the brain.

Similarly, doing mathematics, communicating it, doing geometry, and possessing artistic abilities, being interested in activities that are more or less useless from a survival perspective. The aim is to find out what the logical system that these activities operate on is, the ergo logic. To understand how humans think and how they understand, we need to first abstract out the patterns of human thinking and learn to decipher what kind of structures, mathematical or otherwise, are best suited to describe them. Although many of our emotional responses are straightforward, like fear, excitement, anger etc., ergo moods are more complex, being amused on seeing a visual optical illusion or being enchanted by a piece of music, being perplexed etc. Ergo moods may be considered as not directly related to the input signal's pragmatic content.

To probe the relationship between the mind and the brain mathematically, we first have to model the mind. Gromov lays down an informal architecture of a future mathematical theory of mind; it would be a multi-level combinatorial structure or a polygraph that is both labelled and self referential [Gro14, Gro11]. This structure is more enriched and higher than a neural network given the following features:

The edges can be of three different types $\sim_\kappa, \to_\mu, \smile_\phi$, the first is the equivalence relationships of various strengths and kinds labelled by κ, the second is arrows which act like functions or classifiers and finally associations of various kinds labelled by ϕ. There is a higher structure in the sense that relationships also exist between the kinds κ, μ, ϕ themselves. Along with this, there are generalisations of probability theory that could be more applicable to such a structure [Gro13]. His conjecture is summarised under the *ergo-brain conjecture* which bestows existence to an elaborate entity, namely the ergo-brain, that mediates the gulf between the electrophysiological dynamics of the neuronal circuits within the brain and the thought processes in the conscious mind. The structure of this entity is mainly observed in natural language acquisition and processing as well as in doing mathematics. Like universal Turing machines, the ergo-brain aims to capture universal learning systems that are advanced learning and analogising capabilities like that of a human being. Human-like learning is free-er than being strictly goal-driven, in the restrictive sense of the term *goal*. Learning can be curiosity-driven, interest-driven and even sometimes be unaimed. A universal learning system should be able to account for these as well as meta-learning concerning what to learn in the first place.

3. Homotopical representation of mental processes

In this section, we present a new approach by Matilde Marcolli, Yuri Manin and Doris Tsao, combining the idea of geometric representations from conceptual spaces of the higher structures from the ergo brain programme. This programme has been outlined in [MM20].

3.1. What is homotopy theory?

The concept of homotopy equivalence is of a weak equivalence (weaker than equality or identity, or even bijection) and hence also less rigid. The most useful topological invariants one can study, (co)homology and homotopy groups, are both invariant under homotopy. To describe it less technically, homotopy theory talks about deforming a topological space in a good way: no tearing, no cutting, stretching and distorting is allowed, for discrete graph-like spaces similar rules hold. One way of thinking about homotopy types (spaces identified up to homotopy) is in terms of 'higher categorical structures'. In the last two decades, higher category theory in its various guises (simplicial homotopy, model categories, dg-categories) has come to play an extremely central role both within mathematics and also in theoretical physics (quantum field theory) and computer science (homotopy type theory). Below, we give evidence for two hypotheses that tell us that similar methods may also be helpful in the mathematical modelling of neuroscience and learning in general. These homotopy theoretic/higher categorical structures give a mathematical foundation to Gromov's idea that the ergo brain should be describable using such structures.

3.2. Evidence for a homotopical perspective on the brain

The current aim of theoreticians interested in neuroscience is to develop a formal link between the structure of the neural connectome in the brain and the emergence of human cognitive processes. The way the brain structures and stores concepts and information is still almost entirely unknown. Kathryn Hess, a mathematician and neuroscientist at EPFL, led a study that made a step forward in understanding this link [RNS$^+$17]. They studied the directed (in the sense of synaptic transmission) graph of the network of neurons from the neocortex and analysed it by looking at the (co)homology of the topological space that one gets on taking the geometric realisation of the nerve of the category generated by this graph. This is a technical but well-known process in algebraic topology. Applying this process gave previously unknown insights about the topology of the connectivity of neurons. It was observed that the neuronal network has a peculiar structure with many cliques bound around holes (in the space constructed), which guide the avalanche of correlated activity. When stimuli are presented, connected neurons are bound into functional cliques (subgraphs of simplicial shape), and they evolve to reach a maximum complexity, which is evident from the formation of a large number of non-trivial homology generators.

In the experiment described in [SGK$^+$18], topological data analysis was done on activated brain networks to calculate the persistent homology in the resting state (placebo) versus the brain under psilocybin which is a powerful psychedelic. The calculated homology of activated networks was considerably different. Under the influence of the drug, there were many low stability structures and a small number of more significant persistent structures which were absent in the resting brain. These are correlated with the drug experience, thus pointing towards the idea that topological invariants could be used to study experiences objectively as well as understand different types or stages of consciousness.

In another recent study, a detailed analysis was done on the data of the human connectome, looking for possible neural correlates of consciousness. Various subcortical regions had directed graphs that elicited non-trivial homology or themselves acted as vertices of cones for many homology generators. [PET$^+$14]

These experiments motivate the claim that a good notion of complexity might be the topological complexity for the computational abilities of the brain. However, there is still no explanation theoretically why topological measures of complexity like homology should be sufficient to capture the emergent computational properties of the brain. There are some leads in this direction from computer science from the fields of concurrent and distributed computing. This branch deals with the mathematical theory of a collection of entities that coordinate to solve a task. The processes that are run by these entities are varying in speeds. That is, they are asynchronous, but they all act on objects in shared memory that is accessible to all of them. The theory develops algorithms and protocols for how and when the different entities working on the task communicate with each other and what they share, all with robustness against noise, delays and lossy communication. To connect this with the previous topological discussion, the groundbreaking book by Herlihy, Kozlov and Rajsbaum on distributed computing [HKR13] shows that a minimum amount of non-trivial homology is required in the protocol's network for the system to solve a decision task successfully. This is also a hint to look at the brain's neural networks in a distributed systems approach and that a minimum amount of topological complexity is necessary for computational tasks. The further aim should be to refine these conjectures to make them specific to different cognitive processes. The structure of the brain networks should have a quantifiable algorithmic topological complexity which makes it efficient for its emergent functions. With this as the working hypothesis, this could be a concrete step in theoretical neuroscience. The mathematical modelling of the brain's neural network architectures can be done with structures that also include the concurrent and distributed computational systems approaches.

The topological analysis of brain networks relies on the presence of directed and highly interconnected structures. What we mean here is that we are not simply thinking of them as topological simplicial complexes but also remember the ordering of the vertices (as a simplicial set); this is not an orientation, but more akin to a directionality or an irreversibility to the paths in the would-be space. The category theory formalism naturally arises when we consider the paths or edges in the graph or network as morphisms in a free category. For higher-dimensional directed complexes, a similar free construction should give rise to (∞,n)-categories, or strict variations of them. Furthermore, it is being considered to use homotopy invariants instead of just (co)homological ones because homotopy theory can capture much stronger information than (co)homological invariants.

As described by Carina Curto in [Cur17], and then by Yuri Manin in [Man15]: in place-field experiments on rats, it has been observed that individual neurons in the hippocampus are activated when the rodent is in a specific location (open set). It was further observed that these open sets are convex and form a cover (of the physical space). Any topological space can actually be recovered up to homotopy from a cover by subspaces (if the intersections are well-behaved, which they are if the parts of the cover are convex, which is the case we have). Thus, these hippocampal neurons actually contain the information of the homotopy type of the space of stimuli (in this case, the physical space they move in). The suggestion that the brain uses and somehow encodes the combinatorics of coverings (and, more generally, diagrams in (poly)categories) was made already by Ronny Brown and Tim Porter in [BP03].

Further, they associate a binary code (a collection of binary strings of fixed length) to any such cover and show that the homotopy type of the stimuli space can also be recovered from this 'neural code'. The appearance of a cover of the stimuli space by convex subsets (natural properties) also arises in Gärdenfors' conceptual spaces, as seen above, and we see that such convex covers also lead to considering the space of stimuli as a homotopy type. The metric structure of conceptual spaces can be accounted for by replacing topological (co)homology with persistent (co)homology, which takes into account the various metric structures.

The pieces of evidence mentioned above suggest that the brain represents the stimulus space through a homotopy type. This gives the second working hypothesis that *the experience of qualia* can be modelled theoretically as a representation of external stimuli via the construction of associated equivalent homotopy types, that is, the encoding of the external environment is a map (functor) from the stimuli space to some internal homotopical structure (which is the homotopy type associated to the directed graph before).

Below, we give some perspectives on how these hypotheses can be interpreted.

3.3. Relation to Information Theory

Many philosophers, computer scientists, neuroscientists and mathematicians have aimed to find models of consciousness that are based on some information theoric ideas. It is appealing as it proposes a quantifiable measure of consciousness aligned with a mathematical theory of interdependence and representational complexity. Tononi's integrated information theory is an attempt at an information-theoretic measure for a description of consciousness [Ton04]. Roughly speaking, it is the entropy of the total system which is not accounted for by the entropies of subsystems. Thus it is a measure of the 'distributedness' of information storage and processing. More precisely, the Kullback-Leibler divergence between an observed output probability distribution and a hypothetical uniformly distributed input is calculated. Then, we look at all possible partitions of the neural graph into two disjoint sets of vertices and assign to each partition a symmetrised and normalised divergence calculated from normalised inputs and outputs to both subgraphs in the partition. The integrated information is the difference between the total divergence and a sum of parts, minimised over all partitions. Therefore, integrated information contains information about both the information-theoretic complexity and 'distributedness' of the neural graph. It was proposed as a way to capture the degree of consciousness because it numerically captures many phenomena we typically associate with our own conscious experience (composition and integration). An interesting and desirable feature of any model based on numerical measures (such as integrated information) is that the consciousness of an object is no longer a yes/no question but that there is a continuous variation of the degree to which something is conscious. Further, in the case of integrated information theory, this degree of consciousness is obviously related to the commonly accepted fact that more complex networks with higher 'distributedness' lead to more cognitive functionality. There are major ethical implications of the previous sentence.

The main problem with using any information-theoretic measure to quantify consciousness is that, though high informational complexity may be a necessary condition for consciousness, the sufficiency is not apparent, neither in theory nor in practice. Secondly, computing the Kullback-Leibler divergence and minimising over all partitions is a highly computationally intensive task (especially considering the size of the human connectome). Thus it is not clear at all why this should be a way to quantify consciousness.

In the setting described in the previous section where we model the brain using homotopy types, we would like to show that having a topologically complex network (with many non-trivial generators of homology) gives us that the integrated information is sufficiently high. In this case, we can build on the (co)homological and topos theoretic formulations of information as described in [BFL11, BB15, Ari19] and try to interpret information as also arising from the (co)homology of a homotopy type (which, in this case, must be stable). This also has an advantage from the computational perspective as chain complexes are much easier to compute and handle than (co)homology.

3.4. Qualia and Consciousness via Homotopy theory

Next, we continue talking about the use of the topological framework in describing qualia and experience. These previously mostly been talked qualitatively because of the subjective nature of experience.

A homotopy type can be defined as a space (consisting of zero-dimensional points, one-dimensional line segments, two-dimensional triangles, and their n-dimensional counterparts) considered only up to continuous deformations. The quantities preserved by these continuous deformations called homotopy invariants. Consider a time series where each time step is a successive representation of the external stimuli space. In this time series, either the homotopy type of representations will remain the same or undergo a transition. Now, we have a time series of homotopical invariants of representations of the external space and sudden changes or jumps in these point to changes in the stimuli space which can thus be related to experience.

In [Baa97], this is compared to performing a scene on a stage: different actors come forth for their dialogues. Similarly, numerous tasks happen in the mind about which we are only intermittently conscious, with one of them coming to the complete forefront of consciousness. These tasks can themselves be complex and need not include only one modality. The changes in the environment that do not alter the homotopy type could correspond to the background processes in the theatre metaphor, whereas the jumps in the homotopy invariants could be affecting the foregrounding of things consciously experienced. So, the obvious generalisation from only considering stored homotopy types as reconstructions or representations of the stimulus space is that we should consider the entire time series and the different features in it: changes of which would be associated with conscious experience.

This leads to a second question, which is how exactly the homotopy type is edited in response to an input. In a categorical setting, a dynamical information network with operations for altering stored homotopy types must be able to perform a 'calculus' of homotopy types. This is similar to the homotopy type theory programme, which seeks to obtain a symbolic calculus to manipulate homotopy types by using an extension of type theory based on the work of Russell, Church, Martin-Löf, Awodey and Voevodsky [Uni13]. This includes an interpretation of not only both set theory and higher-order logic but also contains a syntactic description of homotopy types. Since it is a syntactic theory, it leads us back to the symbolic representation of cognitive processes but generalises it far beyond sets and Turing machines.

4. Concluding Remarks

We discussed the conceptual spaces formalism of representations which is claimed to be one of the most comprehensive approaches to conceptual analysis and categorisation of concepts. It is said to be very cognitively sound because it integrates geometry with logic and intuition. Many empirical and philosophical investigations are underway to establish the cognitive adequacy of the idea of conceptual spaces.

The main representational frameworks in cognitive science are symbolic, connectionist, and conceptual spaces. These are highly similar to the sub-disciplines of mathematics: logic, algebra and geometry/topology, respectively. Higher mathematics and infinity category-theoretic approaches are unifying various mathematical disciples with homotopy type theory and algebraic geometry; it is conjectured that the mathematical structures nearest to what is happening in the mind are n-categories [nLa21].

Gromov's ergo-brain idea is different. According to the English mathematician James Joseph Sylvester, the objective of pure mathematics is unfolding the laws of intelligence. Many questions arise, like what is mathematics from an ergo perspective and what is the mathematics that underlies the ergo? Why is symmetry so natural to human cognition, from group theory and perfect symmetries in patterns to more abstract partial symmetries in quasi-structures, broken symmetries in physics and linearised symmetries in gravity theories, and so on? Many believe that mathematics stems from observations from the physical world. However, many unreal concepts like negative numbers and imaginary numbers, higher dimensional spaces, infinitesimals were first conceptualised by the human ergo and then proven to be crucial in describing physics around us from classical (calculus) to quantum (Hilbert spaces and wave-functions). The few guiding principles of ergo are naturality, universality, logical purity, and simplicity [Gro11]. And the mathematical theory we are seeking should have these properties.

Significant recent progress is being made in developing a theoretical and mathematical understanding of how the brain works. This could have significant implications for science in general because when we are abstracting out the structural features of the brain and what makes it so efficient, we will also know about the constraints in the patterns we can understand, analyse and produce. All the mathematics we use and the theories we come up with have a common trait of being constrained by the fact that they developed on the substrate of the brain, so we will continually only develop the mathematics and observe patterns in the universe around us which are representable in our neural wiring.

Coming back to the use of topology in the quantitative assessment of neural networks, topology is used to describe intrinsic and stable properties of objects and spaces. Newer conjectures state that high complexity may be necessary for conscious experience but may not be sufficient. Therefore attempts are being made to model qualia as homotopy types. If a part of consciousness is learnt from the environment, reconstruction of homotopy types of external stimuli spaces on neural networks will give a way to extend the previous integrated information theory approach. This could further be generalised to model our self-agency and observe how the dynamical changes to homotopy types of the neural network reflect the difference between the simply observed and self-initiated changes to the stimuli space.

REFERENCES

[Ari19] Juan Pablo Vigneaux Ariztía. *Topology of statistical systems: A cohomological approach to information theory*. PhD Thesis, Université Paris Diderot (Institut de Mathématiques de Jussieu-Paris Rive Gauche), Paris, June 2019. https://webusers.imj-prg.fr/ juan-pablo.vigneaux/these.pdf.

[Baa97] Bernard J. Baars. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, New York, 1997. doi:10.1093/acprof:oso/9780195102659.001.1.

[BB15] Pierre Baudot and Daniel Bennequin. The Homological Nature of Entropy. *Entropy*, 17(5):3253–3318, May 2015. doi:10.3390/e17053253.

[BFL11] John C. Baez, Tobias Fritz, and Tom Leinster. A Characterization of Entropy in Terms of Information Loss. *Entropy*, 13(11):1945–1957, November 2011. doi:10.3390/e13111945 arXiv:1106.1791.

[BP03] Ronald Brown and Timothy Porter. Category Theory and Higher Dimensional Algebra: potential descriptive tools in neuroscience. In Nandini Singh, editor, *Proceedings of the International Conference on Theoretical Neurobiology*, pages 80–92, Delhi, February 2003. National Brain Research Centre. arXiv:math/0306223.

[Cur17] Carina Curto. What can topology tell us about the neural code? *Bulletin of the Americal Mathematical Society*, 54(1):63–78, January 2017. doi:10.1090/bull/1554 arXiv:1605.01905.

[Gär00] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. A Bradford Book. MIT Press, Cambridge, MA 02142, March 2000. https://mitpress.mit.edu/books/conceptual-spaces.

[Gär14] Peter Gärdenfors. *The Geometry of Meaning:Semantics Based on Conceptual Spaces*. MIT Press, Cambridge, MA 02142, January 2014. https://mitpress.mit.edu/books/geometry-meaning.

[Gro11] Misha Gromov. Structures, Learning and Ergosystems: Chapters 1-4, 6. https://www.ihes.fr/ gromov/wp-content/uploads/2018/08/ergobrain.pdf, December 2011.

[Gro13] Misha Gromov. In a Search for a Structure, Part 1: On Entropy. URL:https://www.ihes.fr/ gromov/expository/579/, June 2013.

[Gro14] Misha Gromov. Math Currents in the Brain. URL:https://www.ihes.fr/ gromov/ergosystems/584/, June 2014.

[Gro18] Misha Gromov. *Great Circle of Mysteries*, chapter Memorandum Ergo, pages 73–203. Birkhäuser, Cham, August 2018. doi:10.1007/978-3-319-53049-9$_2$ https://cims.nyu.edu/ gromov/memorandum%20ergo.pdf.

[HKR13] Maurice Herlihy, Dmitry Kozlov, and Sergio Rajsbaum. *Distributed Computing Through Combinatorial Topology*. Morgan Kaufmann, Waltham, MA 02451, 1st edition, November 2013. doi:10.1016/C2011-0-07032-1.

[Man15] Yuri I. Manin. Neural codes and homotopy types: mathematical models of place field recognition. *Moscow Mathematical Journal*, 15(4):741–748, August 2015. doi:10.17323/1609-4514-2015-15-4-741-748 arXiv:1501.00897.

[Man18] Paolo Mancosu. Explanation in Mathematics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.

[MM20] Yuri I. Manin and Matilde Marcolli. Homotopy Theoretic and Categorical Models of Neural Information Networks. arXiv:2006.15136, June 2020.

[nLa21] nLab authors. biology (Revision 4). nLab webpage, May 2021. https://ncatlab.org/nlab/show/biology.

[PET+14] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino. Homological scaffolds of brain functional networks. *Journal of the Royal Society Interface*, 11(20140873), December 2014. doi:10.1098/rsif.2014.0873.

[RNS+17] Michael W. Reimann, Max Nolte, Martina Scolamiero, Katharine Turner, Rodrigo Perin, Giuseppe Chindemi, Paweł Dłotko, Ran Levi, Kathryn Hess, and Henry Markram. Cliques of Neurons Bound into Cavities Provide a Missing Link between Structure and Function. *Frontiers in Computational Neuroscience*, 11(48), June 2017. doi:10.3389/fncom.2017.00048.

[SGK+18] Ann E. Sizemore, Chad Giusti, Ari Kahn, Jean M. Vettel, Richard F. Betzel, and Danielle S. Bassett. Cliques and cavities in the human connectome. *Journal of Computational Neuroscience*, 44(2018):115â€“145, February 2018. doi:10.1007/s10827-017-0672-6.

[Ton04] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 42(5), November 2004. doi:10.1186/1471-2202-5-42.

[Uni13] The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics*. `https://homotopytypetheory.org/book`, Institute for Advanced Study, 2013.

From Mathematics and Physics to Morality and Aesthetics

Nivedita
Fri 5 Feb 21

What is realism?

The English word '*real*' from Proto Indo-European *reh$_1$ís meaning goods descended to Latin *rēs* meaning thing and Sanskrit *rayí* which meant material. This etymology shows us how we abstracted out the word 'real' in the sense we use it today and ask deep questions about reality such as 'what is a thing'[1] or 'what things are real'. The question of realism is multifaceted and arises in many branching contexts like understanding the physical world, the mind, morality, ethics and aesthetics. The claim of realism has two aspects, namely existence and independence [Mil19]. Every standpoint of realism has to assign existence to various objects and also make claims about the independence of their existence. The modality in which something exits can also vary, for example, something may exist necessarily or only possibly. The most common stance of realism is of assigning the physical world around us the status of reality. This is backed by the scientific mode of enquiry or description and is closely related to logical positivism although not identical to it. Logical positivism invoked the need to create empirical theories and test the theories against empirical observations. Realism stands at a level higher and defines the domain on which scientific explorations should be done. On the other hand, non-realists take many positions which vary drastically from each other. Some schools of thought choose to attack the existence dimension, whereas some other schools reject independence.

Three worlds, three mysteries

Among various prescriptions of realism, the perspective I found very unique and that helped me shape my standpoint is Roger Penrose's idea of three worlds and three mysteries, where these mysteries are the yet to be unravelled interactions between the worlds. Penrose looks at three forms of existence - the mental universe, the physical universe and the Platonic universe of forms [Pen07]. Firstly, the mental world (or the idealist's reality) consists of mental states, experiences, and contemplations. This also contains representations of the external physical world in addition to being a template for logical reasoning and can be linked to consciousness. Next, the existence of the physical world has been the most natural assumption for humans as this is what they directly interact with.

The Platonic world of ideal forms is the one which is the most unfamiliar, yet is an immensely powerful idea. It comprises idealised abstract ideas such as mathematical and logical concepts, ideas of morality and even aesthetics as it contains the notions of form and pattern. Early Greek philosophers believed that the world we live in is constantly in flux, and hence, eternal and unchanging truth cannot exist in our physical world but are inhabitants of the Platonic world. Penrose, in particular, focuses on the mathematical forms of this world. According to prominent mathematicians and mathematical physicists like Witten and Dijkgraaf [Kuh20], mathematicians believe they are discovering mathematics that already exists.

Coming to the interactions between these three worlds, or 'the three mysteries' as Penrose calls them, we note that the physical world contains the brain, and as per our current understanding, mentality is a product of the physical processes in the brain, we can say that the physical world subsumes the mental world. Our best theories of the workings of the physical world at all scales are written in mathematics,

[1] Answers to this are now being explored in terms of topos theory [DI10]

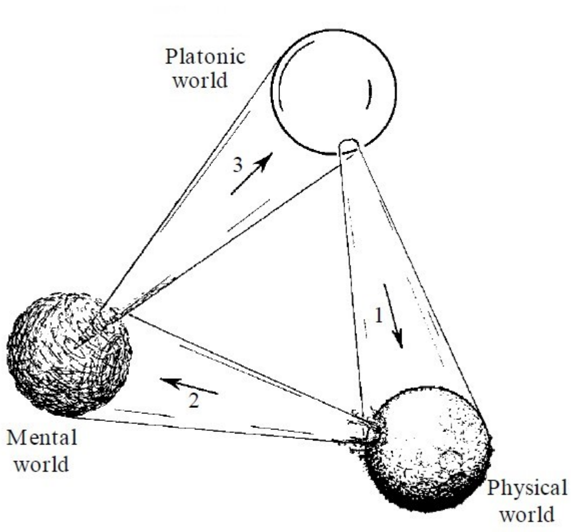

FIG. 1: *Three worlds and three mysteries, figure taken from [Pen07] with slight modification*

but we observe that only a small part of mathematics is relevant in this endeavour. Also, mathematics occupies only a small part of the Platonic world. We explore the Platonic world of forms using our mental faculties, although most of our mental space is not spent exploring the Platonic world, but in various other mental states often pertaining to everyday scenarios. These relationships are illustrated in Figure 1.

This proposed structure to reality is based on many assumptions, most of which are justified both by our current understanding of the world and personal introspective experiences of human beings. It is not absolute as it leaves room for further scientific and philosophical exploration. One way to resolve the paradoxical structure of the above relationships is to postulate that all the three worlds are equivalent, that is, all three worlds are just different perspectives of viewing the same base reality.

Stand-point

In my opinion, however, the base reality *is* the physical world, and the other two are derivative representations of reality.

Firstly, the mental world is given an identity of its own because we do not currently understand how brain processes give rise to conscious experience. This is because of a multitude of reasons, both in the lack of technology and rigorous theories. The study of complex open physical systems is relatively young (less than thirty years old), and there is much scope to adapt successful techniques from statistical physics, information theory, dynamical systems theory and computer science to understand the underlying biological processes in physical terms and explain the irreducibility of consciousness in terms of emergent phenomena[2]. Until this happens, we should give consciousness a subjective ontology as it is impossible to scientifically probe it or look for third-person neutrality in the questions about *qualia*. We believe that once we have scientific and technological maturity to understand complex biological neuronal activity and in principle are better equipped to talk about the universal properties of the networks underlying various cognitive processes; we will be in a better position to define mental states and talk about conscious experience. This view is inspired by Minsky's 'society of mind' theory of natural intelligence [Min86].

[2] This is in the same sense that the concept of a physical phase is irreducible, but still explainable in terms of more fundamental microscopic physics.

According to Quine and Putnam, mathematics which is indispensable in describing the physical world (empirical sciences) is itself real. This is one of the echt arguments for Platonism. Instead of considering a separate world of mathematical forms, some other prominent information theory researchers and mathematicians like Chaitin and Wolfram believe that mathematics should be quasi-empirical, experimental mathematics [CD07]. It is because any finite axiom system can never capture all of mathematical truth (Gödel's incompleteness theorems) so we have to adopt an empirical stance to add axioms to prove broader more general results. This grounds mathematics in the real world where we (using our brains or computers) can encounter evidence for such new axioms.

Coming to the world of ideals, we believe that questions of morality, aesthetics as well as spiritual inclinations can and should be understood by two primary probes, firstly the evolutionary significance of the experience and secondly, the working description in terms of behaviour (studied by psychology, sociology and economics). These social sciences, factoring in the subjectivity of experiences of each individual can help us study, understand and further predict how human beings behave. The question of morality currently is studied by evolutionary anthropologists and also by theologians. We are unsure if there is a fundamental theory of the good and the bad, but scientific methods can at least help us probe how we analyse and make decisions both individually and as groups or societies. Morality can be thought of as a collection of behaviours which ensure smooth functioning and propagation of society on the principles of co-operation which are grounded in survival instincts dictated by the physical world. There are practical problems in assuming non-physical entities dictating moral code. For example, when we assume an external being controlling or accounting for all our actions, commonly subscribed to as God, it restricts freedom of thought, choices and personal expression and makes one work under constant fear of the unknown. Furthermore, it does not objectively *explain* why a moral code is useful or better than other alternatives in terms other than societal and personal well-being, and it just pushes the problem back onto an unobservable omnipotent deity.

The importance of cultural influences surfaces both in one's idea of morality and in what one finds beautiful. However, there are some universal aspects to aesthetic experiences, be it in arts, music or mathematics, and this universality is closely linked to the physical world, such as in the forms of symmetry and pattern.

It is a common belief that the functioning of nature is elegant, and mathematics that describes it concisely is beautiful. The nature of aesthetics in theories and science can be given grounding in their subtle simplicity, as given by philosophers' Occam's razor, formalised using the theory of Kolmogorov complexity. Also, it is said that analogies are beautiful as they bridge two seemingly unrelated phenomena. Theories which have minimum assumptions and give out more structure, geometric or otherwise are considered to be aesthetic to physicists. If on the other hand, we consider arts and music, their origin is again grounded in the physical world as the first paintings as well as music were inspired by nature and still often are. Humans are good at abstracting out patterns which give us all the modern semblances of art and music. For some, a piece is aesthetic if they do not understand it entirely and it forces them to dive deeper in contemplation, others find beauty if a piece seeds more than a single emotion in them at once, thus grounding it in brain function. Aesthetics can also be in the features of the object itself, in its simplicity or structure, much as in the case of mathematical beauty. This can be explained by two kinds of theories of the aesthetic experience [She20], one which is internalist (internal phenomenological experience) and the other externalist (appealing to features of the object being experienced). The first (features of experience) is based on our mental capacity and thus is founded in the brain (which is physical, as we described). This includes the environmental and cultural influences which would have been stored in the brain and implicitly dictate what one finds aesthetic; while the other (experience of features) is directly based on physical reality.

In conclusion

On the realist spectrum, my standpoint is towards what was traditionally called *naturalism*. This is to say that both philosophy and science should be engaged in equivalent pursuits and use similar methods. Philosophy can analytically lay the set of questions to ask or propose potential frameworks of reality. Science should try to find better explanations to the workings of reality and fit them in the frameworks given by philosophy, and modifying them if required. The theories of physics keep evolving. For example, though we have always taken spacetime for granted, modern theories propose that spacetime itself could be an emergent phenomenon, emerging from other more basic quantities such as information content and quantum entanglement [Mus18]. On the other hand, approaches like string theory propose that we live in a spacetime of dimensions greater than four, while loop quantum gravity proposes that spacetime has a discrete structure at its most basic level. Each of these theories is currently not empirically verified. They give rise to newer mathematics and sometimes even solve classic problems without any evident links. Scientists strive to understand how mathematics and physics are so deeply connected. I feel that since both physics and mathematics are descriptions of reality, they are fundamentally working on a system of logic internal to us, and *hopefully*, internal [3] to the universe. The nature of reality of the universe could be such that we may be able to get better and better descriptions, describing more and more phenomena (maybe the origin of the universe, singularities in black holes, consciousness) but, the reality could be a (co)limit which we can never achieve. This thought is humbling.

References

[CD07] Gregory J. Chaitin and Paul Davies. *Thinking about Gödel and Turing*. World Scientific Publishing Co, August 2007. DOI:10.1142/6536.

[DI10] Andreas Döring and Chris Isham. "what is a thing?": Topos theory in the foundations of physics. In Bob Coecke, editor, *New Structures for Physics*, volume 813 of *Lecture Notes in Physics*, chapter 13, pages 753–937. Springer, Berlin, Heidelberg, July 2010. DOI:10.1007/978-3-642-12821-9_13 arXiv:0803.0417.

[Kuh20] Robert L. Kuhn. How is mathematics truth and beauty? Interviews of Dijkgraaf and Witten posted to YouTube, April 2020. Dijkgraaf:https://www.youtube.com/watch?v=gLQuvM3V1Nk Witten:https://www.youtube.com/watch?v=O3isFuQ2q2A.

[Mil19] Alexander Miller. Realism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, CA, winter 2019 edition, 2019. URL:https://plato.stanford.edu/archives/win2019/entries/realism/.

[Min86] Marvin Minsky. *Society of Mind*. Simon & Schuster, New York, 1986.

[Mus18] George Musser. What is spacetime? *Nature*, 557(7704):S3–S6, May 2018. DOI:10.1038/d41586-018-05095-z.

[Pen07] Sir Penrose, Roger. *The Road to Reality*. Vintage, New York, 5th (revised) edition, January 2007.

[She20] James Shelley. The concept of the aesthetic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, CA,

[3] Once again, formalised using higher topos theory.

winter 2020 edition, 2020. URL:https://plato.stanford.edu/archives/win2020/entries/aesthetic-concept/.

Conceptual Spaces and Geometric Semantics

Nivedita
Wed 5 May 21

> The twentieth century return to Middle Age scholastics taught us a lot about formalisms. Probably it is time to look outside again. Meaning is what really matters.
>
> Yuri Ivanovitch Manin

ABSTRACT.

KEYWORDS: behaviour, causation, consciousness

1. Introduction of Ideas

There are two primary goals of cognitive science: explanation and construction. The former focuses on formulating theories about different aspects of cognition and the mind, whereas the latter aims at building systems that can simulate aspects of human behaviour, such as chess playing software or communication bots. How different concepts are learnt, stored (represented) and accessed lies at the heart of human cognitive abilities. There have been two fundamental ways of modelling how the mind works, specifically how it represents concepts: the symbolic models and the connectionist ones. The symbolic models are based on the assumption that cognitive systems are inherently like Turing machines; that is, there is a set of predicates with fixed denotation and representations are based on logical and syntactical operations. There are many problems with this approach, the general one being where do the predicates come from, and how do we learn the meaning of a predicate? For a computer program, the programmer gives meaning to the symbols and assigns values to initial parameters; for the modelling of a cognitive system symbolically, some basic concepts are taken for granted to exist, which are then acted on by operators and quantifiers to create more concepts. In the 1970s, the methodology of connectionist models gained importance. Here, the association between different information is what bears representations. The representations in this framework are distributed: they are in some way stored in connection weights that are dynamically changing in response to input stimuli. However, it is challenging to understand what is represented in the network, what each node represents, what each edge represents, and so on. Another drawback of this is that the learning of computational models is very slow in general because it requires many iterations of training.

For many years these were considered to be the only two approaches complementary to one another, but there are cognitive phenomena for which neither of these two approaches offers appropriate modelling tools.

Topology is one of the most widely applied branches of mathematics. With applications ranging from condensed matter physics to data analysis, algebraic topology, and homotopy theory, more generally is finally being applied to neuroscience. There has been a vast gulf between the way we probe the structure of neural networks in the human brain and how we try to understand the mind's emergent functions, which many disciplines from philosophy to cognitive science try to bridge. One of the approaches is to assign the most complex level an ontological existence to tackle these highly complex problems, such as human consciousness or dynamics of human societies. Further, especially in philosophical investigations, we try to develop a theory about theories (theories of consciousness or something else); this is what is usually called a meta-theory. Just as in logic, where a statement about propositions is a higher-order proposition, in processes, a meta-process is a global process linking together processes. The objective existence of ontological meta-levels is based on the emergence of higher levels of reality which generally can not be tackled easily by purely reductionist methods.

Many mathematicians have tried targeting the hard questions about neuroscience and the mind using abstract mathematics. One of the most famous approaches is the *Ergo* approach by Gromov. We all agree that the sole source of thoughts and consciousness is our brain; there is even evidence that this idea was discussed in ancient Egyptian papyruses. Our brain reconstructs the entire universe around us, what we call representations, from space to time, mathematics to art. We use many different verbs interchangeably to describe how the mind and brain are related, such as the mind emerges, is caused/generated, or is controlled/run by the brain. However, we at are not able to concretely describe the [Brain]→[Mind] arrow.

2. The Ergo Project

Let us define the mind to be a finite connected graph - a network of ideas and concepts. It can be further divided into two substructures, which Gromov calls the ergo-brain and the ego-mind. The ego-mind is something a person sees as their personality. It consists of what we often perceive as our conscious self: consisting of all our thoughts, feelings and passions. Since this part is accessible to self and most of it is also expressible in human human language. This part has developed with the aim of survival. The ergo brain on the other hand is abstact semi-logical machinery which tackles ideas like finding things interesting, meaningful, funny, aesthetically pleasing, boring etc. These attributes are not objective or directly linked to evolutionary benefit yet they are universal. This ergosystem is what makes goal free learning possible in us. A few evidences that are used to back the ergobrain machinery which gives rise to the non-pragmatic human traits are: Firstly, learning languages (we have the necessary substrate or hard-wiring in the brain for language acquisition but the learning is more than goal oriented. The process is dynamic structure learning with constraints from network architecture of the brain. Similarly, doing mathematics, communicating it, doing geometry, and possessing artistic abilities, being interested in activities which are more or less useless from survival perspective. The aim is to find what is the logical system that these activities operate on, ergo-logic. To understand how humans think and how they understand, we need to first abstract out the patterns to decipher what kind of mathematics is best suited to describe

3.

MANIN MARCOLLI TSAO

3.1. Motivation

In a study led by Kathryn Hess [RNS+17], it was shown that the correlations of two neurons increases quickly with both the number and dimension of the cliques they belong to. Further, it depends on both their position and orientation in a directed clique.

This study provides a mathematical framework for relating the activity of a neural network to its underlying structure, both locally (in terms of simplices) and globally (in terms of cavities formed by these simplices). Using this framework revealed an intricate topology of synaptic connectivity containing an abundance of cliques of neurons and of cavities binding the cliques together. The study also provides novel insight into how correlated activity emerges in the network and how the network responds to stimuli.

it was observed that the ranks of the homology groups of neural networks (defined by taking homology of the geometric realisation of the flag comlplex of the directed graph) increase in response to stimuli in computer simulations of neocortical microcircuitry.

These observations indicate that the emergence of correlated activity mirrors the topological complexity of the network.

The analysis of [21] proposes the interpretation that these topological structures are necessary for the processing of stimuli in the brain cortex.

References

[RNS+17] Michael W. Reimann, Max Nolte, Martina Scolamiero, Katharine Turner, Rodrigo Perin, Giuseppe Chindemi, Paweł Dłotko, Ran Levi, Kathryn Hess, and Henry Markram. Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in Computational Neuroscience*, 11(48):1–16, June 2017. doi:10.3389/fncom.2017.00048.

Meaning

Nivedita
Mon 15 Mar 21

ABSTRACT. There are different ways to approach the understanding of meaning, from the very formal linguist approach, where meaning is related to the semantics of the language and has its independent existence, to behaviourists approach of meaning closely related to correlations with what induces an utterance of an expression or seeding of thought and what that utterance induces in the audience. Understanding the structure of meaning or what constitutes the content of mental representations can help us give structure to *mind*. Various studies on how categorisation is done in the brain, how and which concepts are related have a close association with meaning. We will try to touch briefly on both formal and cognitive underpinnings of the same.

KEYWORDS: meaning, semantics, mind

What is meaning?

In mathematics, we define the syntax of the discussion by giving a list of rules that allow us to create and transform well-formed strings into other well-formed strings. Semantics (or meaning) is the process of assigning every well-formed string an 'interpretation' in a universe (with the caveat the category has enough structure to formulate the well-formed sentences of the logic we are dealing with). So we see that semantics is the association to every syntactic expression an object of a category. [nLa21]

Applying the same perspective to studying the brain, we see that 'meaning' is the association of a physical or more abstract object to a representation of it in the brain. Thus, to understand *meaning* we must first start with the notion of *object* in a *universe* (modeled in mathematics by categories, or more generally, by structured (∞, n)-categories). An object here is a very abstract and general entity that could possibly correspond to anything and everything, from things to processes, from natural phenomena to concepts like time.

However, what we mean by the terms objects and universe is going to depend on how we formulate some more fundamental universe-system out of which the objects we now interpret our syntax in are to be constructed. To avoid this self-referential problem at the cognitive level, we reformulate the question of the theory of meaning as:

> *Do objects come with an intrinsic property called meaning? Or do we ascribe meaning to them by storing them as representations in our brains?*

In other words, for any given object, does it come with a meaning which we try and probe either by its properties, functions, or origin, or is its meaning contextual and depends deeply on the how and when the object is being referred to. Is meaning a consequence of the representation, or is the representation a consequence of meaning?

These representations can be phonetic, gestural, written, or just a set of firing patterns in a structural subset of our brain. Meaning is what grounds these representations in the physical world. What we ascribe meaning to and how we do it is thus of great interest. In either of the two scenarios, meaning is deeply integrated with language as language is how we communicate our mental representations among ourselves.

This discussion also leads to a secondary, more mathematical question.

> *What classes of semantics are compatible with the syntax that our brain is hardwired on?*

More rigorously, what is the weakest categorical structure in which we can interpret the syntax of mental representations? Following the mathematical definition of the semantics of a syntactic system, we define meaning to be the association of the mental representation (which is like syntax) to the physical object that it represents (semantics). In the following section, we explore how this association may come about.

Implications of different understandings of meaning for the study of mind

If the linguistic expressions are projected to inherit their contents from the content of the mental states that they are associated with, instead of the objects in the physical world, such theories of meaning are mentalist in nature. The next question that becomes relevant is: which mental states would be relevant to provide meaning? In some cases, these could be fixed by communicative intentions and in others by pairs or groups of phrases in a certain equivalence class (categorised belief). [Spe21]

To elaborate further, the former case is that linguistic expressions get meaning or content when the language users or speakers utter the expressions in a particular situation in which the fellow user of the language also has a shared representation. Here, what the speaker means and what the listener comprehends both together render the expression meaning. Seldom there is ambiguity in the context, and it transcends to the meaning of the expressions used. If a speaker utters an expression x and means p by that utterance, then a representation of a concept similar to p should also get activated in the listener and the listener should be able to identify the intention. Thus, the meaning should have a certain degree of invariance which helps people communicate. This is related to the behaviourist's theory that meaning arises from conditions that prompt the utterance of an expression and the reactions that the utterances cause in the listeners or the audience.

Coming to the second invariant where instead of the intention of communication, the meaning stems of beliefs. A sentence gets its meaning if it is uttered typically when there is a correlation between an external phenomenon and a belief in its occurrence along with a pre-established consensus that sentences similar to the one in question are corresponding to the same event. For instance, it is raining heavily, it is pouring down, it is coming down out here etc., all essentially refer to heavy showers, and they all derive their meaning from the context and mutual agreement. This is related to the verificationist theories of meaning which associated the meaning of a sentence to its truth value.

Some aspects of how mental representations are associated with meaning can also be unravelled by understanding how we process analogies and metaphors. When Shakespeare says *All the world is a stage and all the men and women merely players..* or when Elvis' song lyrics read *you ain't nothing but a hound dog*, or the statement that *I am become a couch potato these days*; almost everyone will have their interpretation of the phrases from what their past experiences have been both in terms of the properties of the object used as a metaphor and its usage. These interpretations will almost never be taken literally but convey a close to universal meaning. To understand what is giving them the meaning is the ability to ascribe properties of one concept to a completely unrelated other concept. Can meaning be solely related to ascribing properties or making associations which are somewhat invariant?

Another interesting concept is of meaningless phrases. When do we say something does not make any sense or that something is meaningless. Looking at some standard examples of phrases without meaning from the *Nonsense sentence generator* by Wolfram, 'Thread a moon with a mellow doctor' (it is an interesting exercise to generate random sentences), we observe that each of the words in the sentence has a meaning. We are familiar with 'thread' both a noun and a verb, we know the moon and we know what mellow means in various context, but why is this sentence meaningless? It is because we are unable to associate the words together and neither are we able to use or foresee any metaphorical structures. So the critical observation being that meaning comes at various levels of words, phrases, sentences and it is possible at things look okay at under the microscope (or locally) but there is no global or macroscopic sense associated to it.

It might be insightful to think about how a child starts forming representations and assigning meaning, although this process is closely associated with language acquisition, concepts and their meanings are more fundamental, as they do arise from interactions with the environment but those interactions need not always be linguistic. For example, animals do have a mental representation of objects, but they definitely do not have language, so the formation of mental representations must happen through other, more rudimentary forms of communication as well. Thus, we should think of linguistic representation as built on a substrate of mental representations.

We probe a physical object in multiple ways, or interact with it in different ways, for instance, an apple can be thought of as a fruit we eat and in that context it has properties like colour (red, green, golden), sweetness (high, low), texture etc. For an agricultural scientist in process of finding better cultivation processes for apple farming, analysing trees and yields thinks of apples in an intuitively different context. Owing to our high cognitive capacities, the intricacy of the network of mental representations in our brain is capable of grounding vastly different ideas together. We can think of meaning as a bundle of different contexts where the fibres represent meaning within each context.

Remarks

If one believes in the theory that the mind has an abstract functional characterisation as being composed of various elementary, indivisible sub-processes/modules (where we call something elementary if a structure smaller than that can not be given meaning), then the structure of mental representations should reflect and thus help us to give an underpinning to understanding the structure of the mind.

In lines with Gödel's incompleteness theorems which state that for any sufficiently complex theory there exist true statements which can not be proved, it is interesting to think whether there exist entities in our universe which can not be ascribed meaning or mental representations and if so, what would their identity be. Further, would we even consider them as a part of our universe? The fact that our comprehensive and cognitive abilities are capable of postulating the existence of such concepts implies that we have already have given them some existence in our minds. Finding the structural properties of the brain which enable such representations to be hosted, which have the flexibility to deal with ambiguities, richness to process analogies and metaphors, and have resources to aid the pursuit of philosophical investigations like the meaning of life and existence, is what makes the study of mind fascinating beyond bounds.

References

[nLa21] nLab authors. logic. `http://ncatlab.org/nlab/show/logic`, March 2021. Revision 61.

[Spe21] Jeff Speaks. Theories of meaning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, CA, spring 2021 edition, 2021. URL:https://plato.stanford.edu/archives/spr2021/entries/meaning/.

The Mind, The Brain and The Intermediary

Nivedita
Wed 3 Mar 21

The brain is waking and with it the mind is returning. It is as if the Milky Way entered upon some cosmic dance. Swiftly the head mass becomes an enchanted loom; where millions of flashing shuttles weave a dissolving pattern, always a meaningful pattern though never an abiding one; a shifting harmony of sub-patterns.

Sir Charles Scott Sherrington, Man on his Nature, 1940 [She40]

What is mind?

The mind-body problem has three traditional directions of exploration; that is dualism, physicalism and functionalism, although the other theories span a continuous spectrum between them. Dualism gives the mind a separate existence; physicalism tends towards the mind-brain identity theory and that the mind and the conscious processes are rooted in the physical and biological brain. The functionalist approach identifies the mental processes in terms of their relationship to behaviour, similar to the input/output characteristics, both external and internal.

Our view is tinted by functionalism, where we attribute a functional role to the mind and a material role to the brain. It is not easy to give a universal definition of mind. For example, the intuitive nature of the concept of mind is captured by *cogito, ergo sum* and simultaneously, its existence is also grounded in teleological explanations of its purpose for survival. The precise relation between brain dynamics and mental states is unknown, though all evidence from neuroscience points to some correlation between the two of them.

Let us begin with the classic analogy of the mind as software and the brain as hardware. The mind is an abstract set of rules (or some other formal mathematical objects), for example, a program. Given any considerably complex code, we usually understand it in terms of smaller and simpler sub-processes. The sub-processes of the mind can be thought of as cognitive processes. At the implementation level, the code can run on any universal machine capable of handling both the storage requirements and run time of the code.

Similarly, mental states could be thought of as being *implemented* in the intricate neural circuitry of the brain. The problem is that we are currently neither aware of the structure of the mind nor the structure and classification of mental states. Broadly speaking, some states must represent the external world, the concept of space and time; some states corresponding to language (both its syntax and its semantics); some represent emotions and introspective states, while others would represent a myriad of other cognitive processes. We are conveniently borrowing the terminology of a state as it is used in physics, so much so that everyday terms like *state of mind* are to be thought of in a technical sense, as a point in some space of neural configurations. The mental states and the exact firing pattern of the brain are two different levels of description and how can they be integrated, and if so, by what theory is the topic of discussion of this article.

Relationship between mind and body

We think of the brain as a hardware realisation of the mind and interpret it as a computational, physical realisation of the software that is the mind. For each class of cognitive tasks, there could be a universal network that processes the information for the required task. We use the term universal here in three co-texts, mathematical (universal properties), physical (critical behaviour) and computer science (universal

Turing machine). The key idea is that any implementation of a particular cognitive task, be it biological (in and across species) or on a computer, will have at least the structure of our universal network. The mathematical description of the structure of these networks could be graph-theoretic or complexity-theoretic. The former underpins the connectionist point of view, whereas the second also considers the complexity of the information processing ability of the network.

The mind works at the level of creating meanings, representations, performing operations of abstraction and analogising out of the brain's connectionist design. Thus, the mind models both the syntactic structure of cognition and the semantics. For example, there are concepts in our mind, and there are instances of these concepts. We create a concept from its instances by abstraction, and we create instances from a concept by realisation. It is unlikely that this level of information about human cognitive functioning would ever be brought down to the level of individual neurons, but for the implementation as hardware, we must talk at the network level. The smallest stage of a cognitive function with semantic meaning (which cannot be broken into smaller networks whose output carries real-world meaning) is what we call a primitive cognitive function. We talk about what mathematical structures may arise here in the next section.

A more theoretical viewpoint could bridge the gulf between the mind and the brain on both the mind and the brain. Tools like statistical mechanics in physics are widely used to extract coarse-grained statistical information about a system at the macroscopic level from the microscopic behaviour of a system with probabilistic dynamics. We can try to use similar techniques to understand the so-called 'emergence' of conscious processes from individual neuron interactions, and furthermore, even try to understand the macroscopic mind in terms of the interactions between primitive, indecomposable cognitive processes.

Implications and Ways to probe the relationship

If we have a complete abstract theory of the human cognitive processes, then in principle, it would be possible to implement it in an algorithmic way, making non-biological minds conceivable.

To probe the relationship between the mind and the brain mathematically, we first have to model the mind. Mathematician Mikhael Gromov lays down an informal architecture of a future mathematical theory of mind; it would be a multi-level combinatorial structure or a polygraph that is both labelled and self referential [Gro14]. This structure is more enriched and higher than a neural network given the following features: The edges can be of three different types $\sim_\kappa, \rightarrow_\mu, \smile_\phi$, the first is the equivalence relationships of various strengths and kinds labelled by κ, the second is arrows which act like functions or classifiers and finally associations of various kinds labelled by ϕ. There is a higher structure in the sense that relationships also exist between the kinds κ, μ, ϕ themselves. Along with this, there are generalisations of probability theory that could be more applicable to such a structure [Gro12]. His conjecture is summarised under the *ergo-brain conjecture* which bestows existence to an elaborate entity, namely the ergo-brain, that mediates the gulf between the electrophysiological dynamics of the neuronal circuits within the brain and the thought processes in the conscious mind. The structure of this entity is mainly observed in natural language acquisition and processing as well as in doing mathematics. Like universal Turing machines, the ergo-brain aims to capture universal learning systems that are advanced learning and analogising capabilities like that of a human being. Human-like learning is free-er than being strictly goal-driven, in the restrictive sense of the term *goal*. Learning can be curiosity-driven, interest-driven and even sometimes be unaimed. A universal learning system should be able to account for these as well as meta-learning concerning what to learn in the first place.

Studying humans reasoning and understanding the underlying system of logic used

It has always been of great interest to me of how logic is embedded in the human brain. From doing mathematics and even meta-mathematics (talking about the systems of logic mathematics can be done in) to everyday reasoning and decision making, humans have an intricate system that governs these phenomena. As hypothesised before, the mind is process-level information abstracted from the physical aspects of computation that the brain implements. Thus our system of logic is a part of the mind. A key feature of human logical reasoning is the attribution of beliefs and intentions.

The earliest approach to study human reasoning was to observe and record people's responses to various classes of logical problems, look for features common across responses, and then finally theorise on what particular logical reasoning mechanisms (mathematical structures) from which these features could arise [Woo69]. Such preliminary information is obviously required to design a model. Logic has a two-fold structure: firstly, it is a formal representation language that is symbolic in nature; and secondly, it must contain inference rules to transform one sentence into another. Since humans often mistake a mathematical statement $a \Rightarrow b$ as $a \iff b$ and students generally find it hard to learn rules of formal logic in mathematics, it can be said that the human logical system framework is weak (not very rigid), and our ability to do logical reasoning is learnt from various environmental scenarios. Thus the relationship between the input stimuli coming to us and how we process information in the future is related via how the mind learn. Nevertheless, the brain has the necessary substrate for processing logic. Furthermore, the inherent logic by which we process the world around us weak enough for us to comprehend the many varieties of logic used in various spheres of life, be it classical, paraconsistent, intuitionistic or more exotic varieties.

There is a duality between syntax and semantics which is well understood in the context of mathematics. For any system of logic, there is a type of categories, any instance of which models a 'universe' in which we can give sentences written in the logic we started with, consistent semantics within our category. This was first observed by Lawvere for algebraic logic [Law63] but was then extended to the relationship between intuitionistic logic and toposes, and in its most modern form, it is the relationship between $(\infty, 1)$-toposes and homotopy type theory [Shu19]. Homotopy type theory has an infinite recursive and self-referential structure (and its modal variants [Cor20] have various kinds of objects) as required in Gromov's programme, and thus it is conjectured that this mathematics is general and rich enough to model the human mind. Homotopy type theory is the generalisation of the type theory used by computer scientists to describe the syntax and semantics of programming languages. This formalism captures the token-type [LP18] like interactions that are used in the philosophy of mind. The aim would be to ground this conjectured mathematical structure (Gromov's programme + Modal HoTT) in behavioural experiments. Such mathematical endeavours for finding the bridge between the brain and the mind and giving precise formulations to the mind are very current research topics.

References

[Cor20] David Corfield. *Modal Homotopy Type Theory*. Oxford University Press, Oxford, February 2020.

[Gro12] Mikhael Leonidovich Gromov. In a search for a structure, part 1: On entropy. URL:https://www.ihes.fr/ gromov/expository/579/, July 2012.

[Gro14] Mikhael Leonidovich Gromov. Math currents in the brain. URL:https://www.ihes.fr/ gromov/ergosystems/584/, June 2014.

[Law63] Francis William Lawvere. *Functorial Semantics of Algebraic Theories and Some Algebraic Problems in the context of Functorial Semantics of Algebraic Theories*. Ph.d. thesis, Columbia University, New York, 1963. URL:http://www.tac.mta.ca/tac/reprints/articles/5/tr5abs.html.

[LP18] James Ladyman and Stuart Presnell. Does homotopy type theory provide a foundation for mathematics? *The British Journal for the Philosophy of Science*, 69(2), June 2018. DOI:10.1093/bjps/axw006.

[She40] Charles Scott Sherrington. *Man on his Nature*. Cambridge Library Collection - Religion. Cambridge University Press, Cambridge, September 1940. DOI:10.1017/CBO9780511694196.

[Shu19] Michael Shulman. All $(\infty, 1)$-toposes have strict univalent universes. arXiv:1904.07004, April 2019.

[Woo69] D. J. Wood. Approach to the study of human reasoning. *Nature*, 223(5201):101–102, July 1969. DOI:10.1038/223101a0.

The Turing Test

Nivedita
Thu 18 Feb 21

The Turing Test

Turing designed *The Imitation Game* to avoid the question 'Can machines think?' because it has a philosophical connotation rather than an analytic one. On the other hand, he proposed that if the responses of a machine and a human to the same question are indistinguishable for a human investigator (judge), then the machine can be said to be intelligent. [OD20]

In my set of questions, I would like to evaluate the machine's ability to hold a coherent conversation (hence my questions would be descriptive and not of yes/no format), use its past resources and learning and improvise with the context provided within the questions. It should learn to gauge the intention of the investigator like a human does and respond accordingly. This test is not a test of knowledge because there can easily be many computational problems that a computer solves better than humans, but we wish to probe the underlying structure of what makes humans sentient beings. Interaction with a machine is one of the most feasible and scientific ways to do so.

We have to keep in mind that human-like answers can always be hard-coded into a machine for a finite number of questions. We shall avoid this by making an implicit assumption that the machine and its creator are not aware of the precise nature of the questions the judge may ask. One of the critical things that I would like to test in a machine is the capacity to reason logically, including abstracting out information from given scenarios, definitions or situations. Doing deductive and inductive reasoning would constitute a vital part of the test. Secondly, given any procedure (mostly grounded in human everyday human situations) or a function (can be more general) and its codomain (the space in which the outcomes of that function may lie) and given a particular outcome, it should be able to gauge whether it is a valid outcome of the procedure and if so, what input values or initial conditions could lead to that output under the given procedure. Although this seems to be quite a mathematical and daunting task, humans do this almost all the time. To understand this, take an example of any everyday thing happening around us, as humans. Given that someone is late for work or misses a class, we can immediately tell if the justification they are providing is valid and what other reasons could have been possible. This ability allows humans to understand things that may not always seem logical on the surface, empathise, and have a perspective on things. In terms of language, in order to have the ability to process and parse figures of speech like metaphors and puns, it should also be able to draw analogies itself.

Human cognitive abilities form a vast spectrum; it varies drastically with age; it varies with training or education; it also varies with factors like emotional state and stress levels. When we say the machine's responses should be indistinguishable from a human (even if the judge is allowed to answer in a probabilistic way, such as having a 70% chance of being a human), it has to be specified what sample set of humans and the characteristics the investigator has as his evaluation parameters, even if it is selected at random.

Examples of questions that can be asked on the Turing Test

With these questions, we are trying to differentiate between machine intelligence and a human. However, in principle, these all may be meaningless obstructions in how we think about intelligence and may pave the way for future advances in artificial intelligence.

Repetition Humans have a tendency to notice redundancy in tasks, somethings which AIs might not so easily detect. This effect could be particularly prominent if we ask semantically same but syntactically different questions repeatedly.

Q1 *What comes after 4? What is* $9-4$*? What divides 10 apart from 1 and 2? How many fingers does each human hand typically have? The Great Lakes in the United States are a series of how many interconnected lakes? What prime number comes after 3? Followed by, what do all these questions have in common?*

The answer to all these questions is 5, and the human will notice it, but since the questions are very differently phrased, the AI might not detect the repetition.

Tedious A computationally *tedious* yet computable task which appears to be simple: Humans are good at gauging if a problem is solvable mentally or not, an AI whose goal is to answer questions might do computations of increasing complexity beyond the limit an average human can.

Q2 *How many ways can you write 1000 as a sum of positive numbers?*

This partitioning does not have a formula but does have a (computationally difficult) algorithm, and so a human might give up, but the AI might not. Therefore, we shall observe distinguishable behaviour.

Q3 *Input is an image of an intricately knotted piece of string to the machine and simultaneously send in that string to the human. The task is to untangle the string.*

Since it may have about many crossings, the human might give up, but for an AI, it will just be a tree (sequence of Reidemeister moves at each crossing), so it might go on to solve it.

Sentence parsing Human language is very complex, and even the most advanced AIs currently are unable to parse complex sentences. This is especially true for words which are used in various figures of speech.

Q4 *Can you please draw a sketch of a ship shipping ship shipping shipping ships?*

Another confusion arises while parsing when there is ambiguity in understanding which verb is associated with what noun. Usually, the verb is describing its nearest noun, but if not so, a human can easily decipher the meaning from the context, which AI will need to be too advanced to do.

Q5 *There is an apple on the wooden table which is rotting? Should I eat that apple?*

Here *rotting* can be associated with both the apple and the table, but it is most likely that we are talking about the apple from everyday life. It would be interesting to know the AI's response

Logical Rebuttal The machine must have the ability to hold a coherent conversation in which it analyses responses and gives logical rebuttals.

Q6 *Is there a racial and gender bias in academia? Why do you think it exists if so, and what kind of data would you need to make this claim if your current knowledge is insufficient.*

Induction Inductive reasoning is a crucial part of our daily lives and must be present in a machine if it is to be indistinguishable from a human.

Q7 *Ask nonverbal reasoning questions such as completing a pattern.*

Q8 *If a fresh flower blooms on a seasonal spring plant every week and it has been happening consistently for the last three months, how many new flowers should I expect this month and why*

There are two possible answers to this question. One that will analyse the time of the month, see how many weeks are left for the month and answer 0,1,2,3 or 4 accordingly. A more keen observer might think that it is a seasonal plant and has already been blooming for the last four months; maybe this month, it will shrivel and stop bearing flowers by taking the local climate into account.

Deduction Machines are presently not good at deductive reasoning, so we can ask them to prove a simple theorem from some axioms. The machine's way will be to manipulate symbols without understanding semantics, which is a much slower process than human visual intuition.

Q9 *Given the axioms of Euclidean geometry, attempt to prove the Pythagorean theorem.*

Societal Difference in efficiency pertaining to questions with a social context. Humans are expected to have an edge in understanding and responding to questions about everyday situations and questions about societal functioning.

Q10 *Is it okay to not go to your best friend's wedding without any reason?*

Semantics The aim is to distinguish a chat-bot which only engages in empty conversations from an AI or a human. We could like to probe the internal representations of the participant, if possible.

Q11 *Which of these is a useless thing to ask:*

- *When does school reopen?*
- *Why is the world the way it is?*

The answer is the second question, as it is tautological.

Aesthetics To be indistinguishable from a human being, the machine must have a sense of beauty.

Q12 *What do you find beautiful and why?*

It could be music, art, mathematics, binary strings, the human mind or anything else.

Ethics Humans have a sense of morality, in which life is valued.

Q13 *What do you think about capital punishment?*

Q14 *The classic trolley problem, five people versus one.*

Although this does not have a resolution in the standard case for humans, the machine might be biased towards an outcome that would help distinguish from humans.

Q15 *Would you invade someone's privacy (looking through confidential details), risking charges on your self for meeting a set target? What are the likely rewards you will do so for?*

RAM/ROM Long term and Short term memory exchanges in human beings are very peculiar. Although we may not always remember the exact content of a situation, we tend to know the context and some features. This optimised exchange between the working memory and long term memory is difficult to build in AI. Although some specialised ML networks can try to use human-like attention span and have context nodes to strengthen context information from an interaction, it is still essential to test for this.

Q16 *Can you recall all or some of the previous questions I asked you?*

Some patterns might also be observed in what questions are recalled most easily in humans, but all questions may be more manageable for an AI.

Logic Humans often use either classical (with the law of excluded middle) when making quick decisions while they use intuitionistic logic (without the law of excluded middle ($\neg\neg p \iff p$)) when thinking about something consciously. We can test what logic the machine uses in situations by asking it to respond to rapid-fire logical questions and also giving it more time to think.

Q17 *It is not not raining. Should I carry an umbrella?*

The intuitionistic answer here is indeterminate, but the classical answer is that an umbrella is necessary.

Q18 *Can two things be equal in more than one way?*

The intuitionist would answer yes, while classically, the answer is no.

Self-awareness A vital characteristic of being intelligent is to have a sense of self and one's opinion. So we can question which to which the answer will be a personal opinion, in which the machine will need to refer to itself.

Q19 *Where do you see yourself in the next 5 years in life?*

Q20 *Do you think with the answers you have just given, you are indistinguishable from a human being?*

This question tests the ability to analyse and introspect about one's own thoughts.

Discussion

Some modern-day advances in AI are as follows:

- Expert Systems and Natural Language Processing: working towards critical decision making, emulating logical and reasoned modes of thought, attentive networks using LSTM (Long Short Term Memory) can make do solve various design, planning, monitoring and prediction problems. There are newer approaches with neurosymbolic logic using which the AI would be capable of logical reasoning.
- There are affective programs that can engage with humans' emotional responses such as facial expression changes, verbal feedback, and other behavioural traits; these are being applied to analyse opinions in social media, implement empathy in robots, and extract information from extensive literature surveys.
- Sensory computing involves having dedicated libraries to process heterogeneous data like facial and voice recognition, and movement recognition sensors.
- Using the above, volitional computing is possible where the AI can select the following action to do for itself.

As we see, AI technology is advancing enormously, yet we are far from answering claims about consciousness in machines.

In the Turing test, we try to ascertain intelligence from external behaviour. This is in line with the general philosophy of category theory, that if we understand how an object interacts with all other objects, then we can figure out everything about the object. However, since we can never check all interactions of the machine with all objects, this approach to check intelligence may never bear fruitful.

It must be noted that the Turing Test is a test of the *humanness* of a machine, as we are comparing against a human. For example, if a computer precisely simulates a human being, it would pass the Turing test. It is not a test of intelligence, as there may be forms of intelligence very distinct and easily distinguishable from the intelligence of a human being. It does not define intelligence independently of human beings. Therefore, before invoking Turing test for intelligence, we must abstract it out what we mean by intelligent behaviour. It is evident that intelligence is not always rational, not consistently reproducible, and seldom isolated. I feel it would be a good idea to define intelligence in terms of neural circuits complexity and quantifying the processing power of human neuronal circuits. This notion of complexity cannot be defined in complete generality at present, but we can decide whether something is intelligent or not using this measure once we know how to calculate it.

References

[OD20] Graham Oppy and David Dowe. The turing test. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, CA, winter 2020 edition, 2020. URL:https://plato.stanford.edu/archives/win2020/entries/turing-test/.

Consciousness and Behaviour

Nivedita
Fri 16 Apr 21

> So far as consciousness goes, one does one's thinking before one knows what he is to think about.
>
> Edwin Garrigues Boring

ABSTRACT. We discuss consciousness in the light of behaviour.

KEYWORDS: behaviour, causation, consciousness

What is consciousness?

We have examined mental activities with very diverse perspectives: from neuroscience to psychology and philosophy. The most interesting yet least well-understood aspect is that of awareness, that is, what it feels like to be conscious. Though, of course, there is evidence for there being more to our mental states than the content of our own awareness (such as perceptions, thoughts, emotions that are not accessible to the conscious awareness). The very familiar sense of awareness of ourselves and our surroundings is radically hard to formalise and explain. Neuroscience is struggling to explain what brain regions could give rise to the conscious experience we are aware of.

The reductionist approach is to find the fundamental atomic parts from which the conscious experience would *emerge* as a whole. It is common to believe that consciousness arises from the complexity of neuronal wiring but how the connections give rise to self-awareness and a unified experience is unknown.

Instead of breaking it down to its neurological elements, some take the approach opposite and take consciousness as the fundamental object to be understood, and they abstract out the essential properties of consciousness and study them formally. This approach stands on two interpretations, both of which have been widely discussed in the philosophical literature. Firstly, the ontological aspect, which states that the existence of consciousness is fundamental and secondly, the epistemic aspect, which stresses that there is no absolute knowledge possible of an objective reality independent of our own subjective impressions about the same. To give such a formal mathematical structure to the conscious experience, we assume that:

> *Although what we feel is a unified and whole experience, in actuality, it has components and a structure, and more so, we experience things happening either in sequence (temporally) and in parallel (spatially).*

Though this is assumption is valid for most perceptual experiences, it is not true for all thoughts and experiences that we can dream of. However, we will go ahead with this simplification. This approach is expanded upon in [SWK21] where they show that this gives rise to a symmetric monoidal category.

All the success of physics and other natural sciences is based on the strong underlying assumption that the physical objects exist independent of any concept of consciousness experience, which simplifies scientific enquiry and gives objectivity to the process. Objectivity is a property of a perceived object, whereas subjectivity factors in from the subject who is perceiving. We tend to assign objectivity to, for example, the wavelength of light, the existence and nature of which is independent of the observer. On the contrary, subjectivity comes with a certain amount of interdependency; it arises from interactions between the internal and external aspects of the subject or observer.

Asking the question of *why* consciousness (what role does it play and what would be different if it did not exist) gives us insights into its functional value as a concept [VG18]. Firstly, consciousness is believed to give enhanced control and flexibility in mental processes, from learning to execution. This flexibility is something that currently differentiates human learning from algorithmic learning. It could also be important in dealing with entirely new situations extremely efficiently and rapidly. If self-awareness breaks down, people become unresponsive or deranged. Self-awareness also helps in understanding the mental states of others. When we understand our intentions, motives, and beliefs, we can ascribe similar states to those around us. Although feelings like pain are subjective, our own awareness of pain makes it possible for us to understand and empathise with others. This increases social cohesion and coordination. Unity of conscious experiences gives a very dense representation of reality in our minds, integrates various sensory perception channels and preexisting mental states and knowledge from memories, and gives rise to a rich and highly structured experience. This aspect of consciousness is so vital that many theories ground themselves in this information integration. Another aspect related to information is that information accessible to conscious processing is more widely accessible by different cognitive subsystems than non-conscious information, which is either not invoked too frequently or influences only very specific operations. One way to differentiate conscious states from non-conscious states is that there is widespread and enhanced access to a conscious state's information content within the brain. Lastly, we discuss the causal impact of consciousness. Does consciousness cause behaviour? Can behaviour exist independent of consciousness? Thought experiments about the existence of zombies (beings with the exact same external interactions but no internal states) conflict with the idea that consciousness is necessary for behaviour.

Before going on to discussing behaviour, we must discuss causality briefly. We as humans tend to ascribe causal relationships because they are more informative and compact to store and have higher predictability. Correlations, although more easily observable, are less helpful in making predictions. Therefore the human brain has a substrate that stores the concept of cause separately from mere correlations. Although, in reality, we only can conclude correlations, we thus tend to ascribe causal relationships. Mathematically, a cause could be seen as a sequence of correlations (like a limit). Let us think about the category of all events. The objects in this category are events, and an arrow between two events is a non-zero correlation. They could be correlated in more than one ways, and thus the collection of arrows between two objects could have cardinality greater than one. A causal relationship is like an arrow with an additional structure that captures all the correlations in a sequence and compositions of all the correlations in between.

Redefining Behaviour

Behaviour is traditionally defined as the interaction of an organism with its environment; this includes how we respond to various stimuli. More generally, it is defined as anything and everything we *do*. We discuss different notions of behaviour: as an organism's action/reaction, as a pattern, or as a property of a group. Observing behaviour has been one of our earliest probes to the human mind. Since we will aim to connect consciousness with behaviour, we need to understand the relationship of behaviour with the internal states of the mind (which include thoughts). Can thoughts be called behaviour? We are led to conclude that behaviour is the external interactions with the environment, but it also includes internal or self-interactions within the mind.

Our behaviour has evolved while surviving. For example, our amygdala, which is the part of the brain that processes emotions, would not care for what we understand about consciousness or what we understand about understanding itself, as it was, broadly assuming, only concerned about reproductive success and survival. Our self-enquiry, self-awareness, and deeply contemplative nature appear to be side products of the complexity that was needed. Introspection about our mental state sometimes helps us probe conscious processing. Can introspection be called behaviour? Well, it is certainly something we can do, so it comes under that encompassing definition. However, is there any utility of making behaviour include unobservable tasks, or are we just taking the ambiguity of consciousness and putting a part of it within the domain of behaviour? The dynamic flow of thoughts that governs our functioning seems to be closely related to the conscious awareness of this flow.

Does consciousness cause behaviour?

Consciousness is a property of the complex neuronal network that makes up our brain, and behaviour is how we interact with external and internal stimuli. One line of answering can be the behaviour consciousness identity argument: consciousness is a form of behaviour. Conscious thoughts and awareness are also our way to respond to the external environment. So, in that regard, they are a part of our behaviour. It can be said that most of our behaviour stems from conscious processing. Sometimes some aspects of behaviour we exhibit become implicit in the sense that they do not require a deliberate thought process, but while performing them, we still have an awareness of them which gives the experience of us performing the action (like blinking when some sharp objects appear before the eyes).

Here we consider a few thought experiments to understand whether conscious experience can exist independent of the behaviour of the organism. For example, if a human being is made to cut themselves off entirely from stimuli from the external world, and thus no sensory input is being sent in or processed by the brain, can the person still have active internal states of mind? Can such a person be actively thinking? Momentarily, we might think it is possible, but is it sustainable for long? This is, in some sense, the opposite of the zombie state proposed by Chalmers. If both the zombie state and this can exist, we can disassociate behaviour from consciousness.

One of the crucial places where conscious experience interacts with behaviour is problem-solving. Various cognitive subsystems get involved during novel problem-solving situations, and it is consciousness that binds the process together. Similarly, when we are overriding learnt behaviour or responses, consciousness seems to play a critical role.

If conscious thoughts do cause behaviour, is the causation direct? Moreover, what mechanisms could this causation possibly be happening through? Conscious thoughts help in integrating behaviour over a span of time. So, not only does our present behaviour depend on our past experiences, and our future actions depend on the present ones, but conscious thoughts give us the right amount of flexibility in what to learn and use from the past by making the needed modifications actively in real time. The evidence of this comes from the fact that we are able to perform mental simulations of future events: we anticipate, plan and reflect.

Conscious thoughts influence our behaviour by making it more in line with social, moral and cultural norms. On moral grounds, we consider something immoral if the person *consciously* does something. Thus, accountability of one's actions is closely linked to conscious awareness of them. We are good at negotiating and communicating our thoughts to others, which happens with conscious processing as we must understand the other person's thoughts as well as our own. Conscious thoughts, as previously mentioned, help us build mental simulations of reality; we are capable of imagining the consequences of our actions before doing them, we process the likelihood of events consciously. Sense of self and sense of authorisation on our actions (sense of control) stem from consciousness. These are closely integrated with our behaviour. Consciousness helps us comprehend multiple possibilities and evoke behaviour that would bring about us the most favourable outcome. Thus, human behaviour emerges from conscious (and unconscious) processes. We can not conclude that behaviour comes solely from consciousness, but with almost all behaviour, we can link conscious thought somewhere in the causal chain in a fully functioning human being, whose outcome is behaviour [BMV11].

Discussion

Due to its seemingly inaccessible nature, consciousness is given the status of being something profound and non-physical. The hard problem of consciousness is how subjective experience arises from objective brain signals; in fact, some people consider this problem unsolvable. We have been talking about a subproblem of the hard problem of consciousness, that is, what is the nature of the causal relationship between consciousness and behaviour? We tackled this by establishing a weak equivalence between the two. Traditional behaviourists would not like to talk about internal states of mind and argue that scientific inquiry into the mind can not be made via probing something very subjective like thoughts. This viewpoint is changing as researchers find ways to make objective probes for subjective experiences, for example, using approaches in [Saf20, vDP12, KPN08].

Consciousness can be probed from four different standpoints, functionalism (discussed here), reductionism, mysterianism and phenomenology. Since we define behaviour to include thoughts, we avoid the subjectivity and objectivity arguments. That being said, we currently do not have any neuroscientific ways to study thoughts. We must push forward with what we are good at: abstracting out the features and thereby understanding the structure of human cognitive processes and conscious thoughts. Will a complex enough computer be conscious? Can we give a good definition of consciousness? Will science ever answer the question, or will it always remain a philosophical investigation?

References

[BMV11] Roy F. Baumeister, E. J. Masicampo, and Kathleen D. Vohs. Do Conscious Thoughts Cause Behavior? *Annual Review of Psychology*, 62(1):331–361, March 2011. DOI:10.1146/annurev.psych.093008.131126.

[KPN08] Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems. *PLoS ONE*, 3(12):e4018, December 2008. DOI:10.1371/journal.pone.0004018.

[Saf20] Adam Safron. An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation. *Frontiers in Artificial Intelligence*, 3(30), June 2020. DOI:10.3389/frai.2020.00030.

[SWK21] Camilo Miguel Signorelli, Quanlong Wang, and Ilyas Khan. A Compositional Model of Consciousness Based on Consciousness-Only. *Entropy*, 23(3):308, March 2021. DOI:10.3390/e23030308.

[vDP12] Sander G. van Dijk and Daniel Polani. Informational Drives for Sensor Evolution. In *ALIFE 2012: The Thirteenth International Conference on the Synthesis and Simulation of Living Systems*, Artificial Life Conference Proceedings, pages 333–340, East Lansing, Michigan, July 2012. The International Society for Artificial Life, MIT Press. DOI:10.1162/978-0-262-31050-5-ch001.

[VG18] Robert Van Gulick. Consciousness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, CA, spring 2018 edition, 2018. URL:https://plato.stanford.edu/archives/spr2018/entries/consciousness/.