Essay

The Concept of Generalized Reasoning and Its Underlying Circuitry

Robert Friedman *

Posted Date: 23 December 2024

doi: 10.20944/preprints202411.2410.v3

Keywords: computation; circuitry; neural network; general reasoning; deep learning

 doi:10.20944/preprints202411.2410.v3

Essay

The Concept of Generalized Reasoning and Its Underlying Circuitry

Robert Friedman †

Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA
† Retired

Abstract: The properties and processes of general reasoning are often undefined in literature, leading to a perspective relegated at best to the boundaries of pure philosophy. It is therefore of interest to confine the term as a physical process of information flow instead. This dependence on the elements of matter and energy is the link between the human brain and the phenomenon itself, allowing for the hypothesis that an analogous form is possible of artificial design.

Keywords: computation; circuitry; neural network; general reasoning; deep learning

Definitions

The base term reason derives from a statement offered in explanation or justification [1], while reasoning is the drawing of inference or conclusion through its use [2], and finally, the attribute of generalized refers to applicability of kind to a well-defined group [3].

The properties and processes of general reasoning are undefined in the literature, leading to a perspective relegated at best to the boundaries of pure philosophy [4]. It is therefore of interest to confine the term as a process of information flow instead [5]. This is a dependence of the flow on the elements of matter and energy, a conceptual link between the human brain and the phenomenon itself, allowing for the hypothesis that an analogous form is possible by artificial design. However, it is also possible that the term is not grounded in the physical world and instead merely resembles a concept with sole existence in the Mind, and therefore its applicability is bounded by the constraints of metaphysics and its methods.

Reasoning as an Informational Process

Since advanced information processing is a property of the human brain, its mechanistic basis is in the neurons and their interconnections, a network of a kind of biological cells [6]. This is distinguished from the lesser forms of information processing in biological organisms, such as the regulatory networks of gene and protein expression and their dynamics for a biological code used by evolution and in the development of animals and their body plan [7].

The above perspective resembles a form of biological computation. Therefore, general reasoning as a phenomenon may be restricted in both the theory of physics and the informational sciences. It follows that the generalized scenario requires a "program" or "algorithm" that represents a distinct category of kinds [8], a putative lineage of related kinds of reasoning that are both unique and derived in relationship to any other forms of mental processing. An example of this in a skill of mathematical reasoning [9]. The summation of two integer values is a form of abstract reasoning with application to the physical world. Further, its generalized form applies to other cases of summing integer values, leading to its applicability across the other branches of mathematics and sciences, as in the interpretation of discrete objects across a visual scene. This extends a putative process of a specific arithmetic operation to other instances not confined to that of the Mind, regardless of its final interpretation as a kind of interpolation, or the putative but orthogonal process, of sampling for the extrapolation of knowledge [10].

 doi:10.20944/preprints202411.2410.v3

There is also the question on whether general reasoning is partly or wholly defined as *a priori* in the human brain as a substance or is likewise dependent on an experiential existence, as originates from sensation and the interactions in the physical world. However, it is not biologically plausible, nor is there a known mechanical basis for, programming the developing brain with this attribute *a priori* [6]. Therefore, it may be taken as an assumption that general reasoning emerges from experience and programming instead of from an uninitialized neural network. Furthermore, it may be hypothesized that the informational elements of this activity forms patterns [11] that are expressed by a notation grounded in past knowledge of mathematics and computation, such as a subnetwork, a circuit, or an algorithm [5]. It follows that this form of biological computation is interconvertible with a computational framework, whether the translational process is a tractable and bounded computation or not. Another open question is whether, or to what degree, does this process depend upon a modular design [12], since any category of generalized reasoning must have a physical basis for its occurrence, a mechanism based on the conventional elements of matter and its composition from a set of smaller discrete set of elements (atomic perspective of reality).

Reasoning as Circuitry in Computation

If this knowledge of biological computation is robustly and relatively equivalent to that of an artificial form, as can be represented by a human engineering and artificial design [11], then the artificial neural network serves as the analogous form of the biological kind [5].

Recent work in machine interpretability of transformer circuitry [13] shows examples of information processing in this artificial setting and includes a learning process that can lead to a general form of an algorithm and its putative computation, as in the skill of generalized reasoning, as observed in a large language model [14]. An example is shown where two circuits are formed in parallel, a process that is a type of "grokking" [15], that leads to the formation of circuitry dedicated to this higher form of information processing and search of the algorithmic space. The study further contends that the generalized circuit spanned particular but local layers across the neural network, but it is undetermined as to whether this circuitry is confined by locality, and therefore a gradual building upon of the increasingly larger features of the network, or that the circuit can span without interruption across the network layers, and therefore bridge and "bind" the lower order features with that of the higher order features [16].

The above example shows the blurring of the boundary between the neural network approaches in engineering and the alternative practices of neurosymbolic ones [16]. However, the underlying process can be stated as originating and emerging from an unstructured neural network as opposed to any *a prior* design as exemplified by a neurosymbolic approach. This suggests that a neural network is a basis for forming and emerging of higher order concepts, and that they are not defined as *a priori*. A hypothesis to fully test this concept is difficult because it depends on reachability, that any experiment is a robust measure of circuit design. This concept, and that of the others in the above sections, refer to the elements of higher cognition and to the recognition of information for the processes of advanced computation. Therefore, it can be said that the priors of cognition are a neural network that is "programmed" for construction of the informational pathways and patterns that serve as the intermediate basis for advanced computation [11].

A Critique of a Pure Reasoning Process

The pillars of knowledge of the natural world are derived from a theoretical foundation at the suggestion of the Cartesians and reinforced by the practices of experience and experimentation. Its construction is lacking in permanence, a consequence of the corrosive power of dogma and the pressure in a loss of utility from the erosive forces of misrepresentation and misinterpretation of ideas. The twin virtues of elegance and beauty have further served as a stray guide in a search for the truer paths, such as revealed in the “bitter lesson” [17], a reminder of the problems in navigating idealized scientific practices and the central role of data in the machine learning methodology. It suggests data and learning as the core properties of interest, a kind of a complex manifold, and against the pitfalls of hand-crafted design or an over-indulgence for model tinkering. Instead, the

 doi:10.20944/preprints202411.2410.v3

general algorithm for learning with a viable optimal search space are central in finding the patterns of interest in the data.

Hindsight shows the importance of assembling a large quantity and high quality of data samples in building a foundational model and in developing principles for a materialist epistemology based on the mechanistic understanding of knowledge generation. In this perspective, knowledge is represented by a physical process involving the particles of information flow and its cost of movement across space and time. This process is also a mathematical expectation that describes the optimization of data as a manifold and its geometry. In essence, an optimal manifold of data may be described as a set of connections and paths resulting from the compression of data into a more concise state, an optimal arrangement for effective algorithmic processing. The optimality is reflected in the restrictive forces of information flow, a process dependent on the concept of entropy and informational complexity, guides for the potentiality in data compressibility and therefore dependence on algorithms to find the shortest possible paths in traversing the data and its manifold.

This compressibility serves as a guide for forming both hypotheses and a stronger form in the expression of a formal expectation which is central to the formulation of theory. One outcome is that data compressibility should result in an object that is simpler in its geometrical configuration as a manifold. Another is the optimization process which is expected to compress the navigable paths across its features and find efficient algorithmic designs for lowering of traversal costs, such as observed in the resulting linearity in the relationships of facial features among individuals in neuroscience [18] and recapitulated in the engineering sciences [19]. It follows that any circuitry described in this context derives from the processes of compression and that of an evolutionary process. This is also a connectionist agenda for building a manifold from the data, leading to the formation of circuits that may exist in parallel to one another while some of these events defy common predictions on their occurrence [14]. The formation of algorithms in totality is the outcome of these processes and reflective of a form of cognition that is observable in deep learning, leading to models with a deeply entangled parameter space and dependence on empiricism for robust insight.

The newly formed circuitry is a mirror into the optimization process and its unexpected influence on the dynamics of information flow [20]. It leads to a predictiveness on the occurrence of formations of algorithms and optimality in the structuring of the data. Cognition is therefore better defined as a manifold that results from a learning procedure and its attributes for enabling a search that leads to an optimal set of paths in a system.

This is also descriptive of a logical design and potential for computation of a manifold as a system. It follows that the data is therefore expected to contain *a priori* the primitive elements for the formation of new circuits. The totality of the data is representative of these elements as an implicit coding scheme in contrast to any explicit and direct coding of it in the manifold of the data itself.

Conclusion

These definitions and the definition of processes with a mechanical basis are suitable for hypothesizing and constraining the space of possibilities of experiments in the quest of validation of any theory of generalized reasoning. It also suggests that the problem of advancing the skill of it is not an unreasonable proposition, given the assumption it exists *a priori* in human cognition. Moreover, there is supporting evidence for it in the mechanisms of the binding of the lower to the higher order concepts, and a reminder of the categorical nature of generalized reasoning and its kinds, as it must depend upon the physics of information flow, as any of these advanced informational processes involve physical motion, and not depend on an abstract and undefined set of traits, but instead tractable to a formalization and definability as a physical process.

This closes the gap for expectations on efficiency among the artificial and natural forms of cognition, with the data of an artificial neural network as potentially compressible in reference to its parameterization of the data and the consequent search for the shortest manifold, a feat by a sufficiently capable deep learning method. This description is less a process of any force of sorcery and instead a reflection on cognition as a process of information flow across a set of paths in a

 doi:10.20944/preprints202411.2410.v3

manifold – not dependent on any *a priori* neurosymbolic design that lacks in a tractability and explanatory power.

Further, the misattribution of a traditional definition of cognition with a concept of human agency and related phenomena may stem from an inherent resistance on an inherent belief in oneself as central to the universe and its workings, rather than a more material view as an emergent phenomenon from a collection of particles and their motion in forming an abstractive system of cognitive processes; however, the concepts of agency, pursuit of vanity, and "free will" emerge from the deeply entwined processes of the human brain, where the probable cause of this entwinement is likely in a dependence on the necessities for maintenance of sociality and constant measurement of status and its dynamics across a group of conspecifics. Perhaps these concepts reflect a dependence on "social mirrors" for robust social functioning [21]. These phenotypes are high in maintenance cost for the individual, but consistent with the other costly displays as observed across the domains of social and non-social life forms [22].

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Merriam-Webster Dictionary (an Encyclopedia Britannica Company: Chicago, IL, USA). Available online: https://www.merriam-webster.com/dictionary/reason (accessed on 28 November 2024).
2. Merriam-Webster Dictionary (an Encyclopedia Britannica Company: Chicago, IL, USA). Available online: https://www.merriam-webster.com/dictionary/reasoning (accessed on 28 November 2024).
3. Merriam-Webster Dictionary (an Encyclopedia Britannica Company: Chicago, IL, USA). Available online: https://www.merriam-webster.com/dictionary/general (accessed on 28 November 2024).
4. Friedman, R. Cognition as a Mechanical Process. NeuroSci 2021, 2, 141-150.
5. Friedman, R. A Perspective on Information Optimality in a Neural Circuit and Other Biological Systems. Signals 2022, 3, 410-427.
6. Friedman R. Themes of advanced information processing in the primate brain. AIMS Neuroscience 2020, 7, 373-388.
7. Davidson, E.H., Erwin, D.H. Gene regulatory networks and the evolution of animal body plans. Science 2006, 311, 796-800.
8. Hennig, W. Grundzüge einer Theorie der Phylogenetischen Systematik; Deutscher Zentralverlag: Berlin, Germany, 1950.
9. Marcus, G.F. The algebraic mind: Integrating connectionism and cognitive science. MIT press, 2003.
10. Balestriero, R.; Pesenti, J.; LeCun, Y. Learning in High Dimension Always Amounts to Extrapolation. arXiv 2021, arXiv: 2110.09485.
11. Friedman, R. Tokenization in the Theory of Knowledge. Encyclopedia 2023, 3, 380-386.
12. Friedman, R. Higher Cognition: A Mechanical Perspective. Encyclopedia 2022, 2, 1503-1516.
13. Nanda, N., Chan, L., Lieberum, T., Smith, J., Steinhardt, J. Progress measures for grokking via mechanistic interpretability. arXiv 2023, arXiv: 2301.05217.
14. Yang, S., Gribovskaya, E., Kassner, N., Geva, M. and Riedel, S. Do Large Language Models Latently Perform Multi-Hop Reasoning? arXiv 2024, arXiv: 2402.16837.
15. Power, A., Burda, Y., Edwards, H., Babuschkin, I. and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv 2022, arXiv: 2201.02177.
16. Chughtai, B., Chan, L. and Nanda, N. A toy model of universality: Reverse engineering how networks learn group operations. In International Conference on Machine Learning (pp. 6243-6267). PMLR, 2023.
17. Sutton, R. The bitter lesson. Incomplete Ideas (blog) 2019, 13, 38 (https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf).
18. Chang, L., Tsao, D.Y. The code for facial identity in the primate brain. Cell 2017, 169, 1013-1028.

 doi:10.20944/preprints202411.2410.v3

19. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A. The building blocks of interpretability. Distill 2018, 3, e10 (https://distill.pub/2018/building-blocks).
20. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., Carter, S. Zoom in: An introduction to circuits. Distill 2020, 5, e00024-001 (https://distill.pub/2020/circuits/zoom-in/).
21. Bonini, L., Rotunno, C., Arcuri, E., Gallese, V. Mirror neurons 30 years later: implications and applications. Trends in Cognitive Sciences 2022, 26, 767-781.
22. Zahavi, A., Zahavi, A. The handicap principle: A missing piece of Darwin's puzzle. Oxford University Press, 1999.