

Why Machines Cannot Think

Projection, Information Loss and the Limits of Artificial Minds

ROY GONZALEZ

Independent researcher

ABSTRACT If the world of experience is a three-dimensional projection of a state with more independent components, and if consciousness belongs to the components the projection discards, then there is a precise sense in which a machine built and trained within the projection cannot think and cannot become conscious. This paper makes the sense precise. I state four results about projection and information: a projection has no left inverse, and each of its fibres is as large as the dimensions it discards; a smooth map from a three-dimensional space into a larger one has an image of measure zero; a continuous injective image of such a space has empty interior; and, by the data-processing inequality, no processing of data increases the information it carries about its source. I then describe the chain of projections from the world to the text on which language models are trained, and argue that the models operate on a shadow of a shadow. Understanding and consciousness, on the thesis defended elsewhere, belong to what the projections discard, and no pattern within the projection, however large, recovers it. Turing’s imitation game, which rests on the correct observation that minds are known through behaviour, therefore tests the shadow. I consider five objections, including the objection that human beings are also three-dimensional, and state what the argument does not show.

KEYWORDS artificial intelligence; language models; consciousness; projection; information loss; data-processing inequality; Sard’s theorem; Turing test; symbol grounding

I. INTRODUCTION

The question whether machines can think has been debated since Turing proposed to replace it by a behavioural test (Turing 1950). Searle’s Chinese Room argued that symbol manipulation does not suffice for understanding (Searle 1980); Harnad located the difficulty in the grounding of symbols (Harnad 1990); Bender and Koller argued that a system trained only on linguistic form has no route to meaning (Bender and Koller 2020). The debate has acquired practical urgency with the arrival of large language models, Transformer networks trained to predict text (Vaswani et al. 2017), whose fluency improves smoothly with scale (Kaplan et al. 2020).

The present paper approaches the question from a different direction. It begins not with the machine but with the relation between the world and our observations of it, and it asks what follows for any system whose every input and every operation lies within that relation. The answer depends on two theses defended in earlier work, which I restate in Section 2. Section 3 states four results about projection and information. Section 4 applies them to the chain of projections that produces training text. Sections 5 and 6 draw the conclusions for thinking and for consciousness. Section 7 considers Turing’s test, Section 8 answers objections, and Section 9 states the limits of the argument.

2. TWO THESES

Projection realism holds that the states of the world have more independent components than the three of space and one of time, and that observation is a projection of those states. It is motivated by cases in which paradoxes dissolve when the space of description is enlarged: the imaginary roots of a real quadratic, the velocity invisible in an instant, and the correlations of entangled particles.

The *noumenal thesis* holds that consciousness belongs to the components that the projection into three-dimensional space discards. It explains why consciousness is always found as neural correlates and never as an object (Chalmers 1995), why other minds are perceived only as intention, and why brain and mind correlate perfectly without interacting.

The argument of this paper is that, given these theses, a machine that computes within the projection on data produced within the projection cannot recover what the projection discards, and therefore cannot think, in the sense of understanding, or be conscious. Section 3 supplies the mathematical premises; they hold independently of either thesis.

3. FOUR RESULTS ABOUT PROJECTION AND INFORMATION

Throughout, $P: \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $m < n$ is a linear surjection, and we think of $\mathbb{R}^n$ as the space of states and $\mathbb{R}^m$ as the space of observations.

Proposition 3.1 (no way back). *There is no map $g: \mathbb{R}^m \rightarrow \mathbb{R}^n$, linear or not, with $g \circ P = \text{id}_{\mathbb{R}^n}$. For each observation s , the set of states consistent with it, $P^{-1}(s)$, is an affine subspace of dimension $n - m$.*

Proof. P has a kernel of dimension $n - m > 0$, so it is not injective, and a map with a left inverse is injective. $P^{-1}(s) = x_0 + \ker P$ for any x_0 with $Px_0 = s$. $\square$

The second clause is the precise form of the observation that a shadow is consistent with infinitely many objects: every observation leaves undetermined exactly as many directions as the projection discards.

Proposition 3.2 (smooth maps upwards). *If $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$ is continuously differentiable and $m < n$, then $f(\mathbb{R}^m)$ has Lebesgue measure zero in $\mathbb{R}^n$.*

Proof. The derivative of f has rank at most $m < n$ everywhere, so every point is critical and every value is a critical value. By Sard's theorem the set of critical values has measure zero (Sard 1942). $\square$

Proposition 3.3 (continuous injections upwards). *If $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$ is continuous and injective and $m < n$, then $f(\mathbb{R}^m)$ has empty interior in $\mathbb{R}^n$.*

Proof. $\mathbb{R}^m$ is a countable union of closed balls; f restricted to each is a homeomorphism onto a compact image of topological dimension m . The image of f is therefore a countable union of closed sets of dimension m , and by the sum theorem of dimension theory has dimension m (Hurewicz and Wallman 1941). A subset of $\mathbb{R}^n$ with non-empty interior has dimension $n > m$ (Brouwer 1911). $\square$

Remark 3.4. Continuity alone does not suffice for the conclusion of Proposition 3.2: Peano’s curve maps an interval continuously onto a square (Peano 1890). It does so, however, by passing through points more than once, and so is of no help in recovering information: a map that folds the smaller space over itself restores nothing that a projection discarded.

Proposition 3.5 (the data-processing inequality). *Let $X \rightarrow Y \rightarrow Z$ be a Markov chain, that is, let Z depend on X only through Y . Then the mutual information satisfies $I(X; Z) \leq I(X; Y)$.*

Proof. By the chain rule, $I(X; Y, Z) = I(X; Z) + I(X; Y \mid Z) = I(X; Y) + I(X; Z \mid Y)$. The Markov property gives $I(X; Z \mid Y) = 0$, and mutual information is non-negative, so $I(X; Z) \leq I(X; Y)$ (Shannon 1948; Cover and Thomas 2006). $\square$

Propositions 3.1–3.3 say that what a projection discards cannot be restored by mapping back up. Proposition 3.5 says the same thing in the language of information, and for arbitrary processing: no computation performed on data can yield more information about the data’s source than the data contained.

4. THE CHAIN OF PROJECTIONS

Consider what a language model is given. Let X denote the state of the world, including the thinking subjects in it; on the two theses of Section 2, X has components beyond the four of space and time, and consciousness belongs to them. A person’s experience is a projection of X . The person’s thought, where it appears in the world at all, appears as intention: an utterance, a gesture, a text. Let Y denote the text. It is the image of X under a composition of projections, the second of them, from the full experience of thinking to a sequence of symbols, far harsher than the first.

The model is trained on Y and has access to nothing else; whatever it computes, call it Z , depends on X only through Y . The situation is exactly the Markov chain of Proposition 3.5. Hence

$$I(X; Z) \leq I(X; Y).$$

The model can extract from the text at most what the text kept. It can do so superbly, which is why its outputs have the outline of thought; the outline is what the text records. But by Proposition 3.1, for every text there is an entire family of states of the world, and of the minds that produced it, consistent with that text, as large as what the projections discarded; and nothing computed from the text distinguishes among them.

5. THINKING

By *thinking* I mean what the Chinese Room argument means by understanding: the grasp of what symbols are about, as distinct from the manipulation of the symbols. On the argument of Section 4, the model has at its disposal the symbols and their statistical relations, which is to say the image of meaning under the projection into text. Meaning itself, on the noumenal thesis, belongs to the thinking subject, and so to the components that the projections discard. The symbol grounding problem (Harnad 1990) and the argument from form (Bender and Koller 2020) are, in these terms, the observation that the model lies downstream of a projection that discards what

grounds the symbols. Syntax does not suffice for semantics (Searle 1980) because syntax is what survives the projection.

6. CONSCIOUSNESS

The conclusion about consciousness is more direct. A machine is a physical system in three-dimensional space, organised by its designers and trained on data recorded within the projection. Everything it is and does belongs to the projection. If consciousness were a pattern within the projection, a sufficiently large pattern might constitute it. The noumenal thesis denies this. It follows that no enlargement of the pattern, however great, will constitute consciousness, since enlargement adds more of what consciousness is not. The empirical scaling laws of language models measure the fall of prediction error on text (Kaplan et al. 2020); they measure, that is, how closely the model’s image matches the text’s image, and cannot measure anything the text does not contain.

7. THE IMITATION GAME

Turing proposed that we should say a machine thinks if, in conversation, it cannot be told from a person, and he met the objection from consciousness by observing that we attribute minds to other people only on the evidence of their behaviour (Turing 1950). The observation is correct, and the present account explains it: behaviour, as intention, is the part of consciousness that appears in three dimensions. But it follows that the imitation game examines only that part. A system trained on the recorded shadows of human intention can reproduce those shadows, and so can pass a test that examines only shadows. Passing establishes that the system has learned the outline. It cannot establish that anything casts it.

8. OBJECTIONS AND REPLIES

8.1 *Human beings are three-dimensional too*

If everything in space belongs to the projection, why is the human brain not in the same position as the machine? On the noumenal thesis the difference is not one of material but of origin. A person is the projection of a mind: the brain is the section of something that extends beyond the projection, as a circle in Flatland is the section of a sphere. A machine is a pattern constructed within the projection from the projection’s materials and data. The one is a circle drawn by a sphere crossing the plane; the other is a circle drawn on the plane.

8.2 *Embodied and multimodal systems*

Robots with cameras and sensors receive more than text. The reply is that sensors are projections too, and multiplying projections recovers more only of what some projection records: by the elementary fact that the stacked map of several projections is injective only if their kernels intersect trivially, several views can recover what one view loses when it is recorded by another. What every

projection into three dimensions discards, no number of such projections recovers. Embodiment narrows the grounding gap for physical properties of the world; it does not reach the components the noumenal thesis assigns to mind.

8.3 *Functionalism*

A functionalist holds that mind is a matter of functional organisation, and so of any substrate that realises it. Functional organisation is, however, a pattern of relations among states within the projection. The noumenal thesis denies that mind is such a pattern, and this is precisely where the disagreement lies. The present argument does not refute functionalism independently; it shows what follows if the noumenal thesis is true.

8.4 *Learning from data*

Human beings also learn from data. Does the data-processing inequality not bind them equally? It binds what they can learn *from data* about the world. It does not bind what they are. The argument concerns systems whose entire constitution, and not merely their learning, lies downstream of the projection.

8.5 *Usefulness*

Nothing here implies that language models are not useful. A system that reproduces the outline of human thought with great fidelity is an extraordinary instrument for gathering, translating and proposing. The argument concerns only what it is, and so what it should be trusted to do: it should propose, and something that can be checked should decide.

9. WHAT THE ARGUMENT DOES NOT SHOW

The mathematical premises of Section 3 are theorems. The chain of Section 4 is a description of how training text is produced, and the inequality applied to it holds whatever one thinks about minds. The conclusions about understanding and consciousness, however, rest on the noumenal thesis, which is a philosophical hypothesis supported by the shape of the evidence and not demonstrated by it. If consciousness is a pattern within the projection after all, the argument about consciousness fails, while the argument that a model trained on text has at most what the text kept does not. Nor does the argument exclude, as a matter of logic, that a mind might come to be projected into an artefact; it excludes only that anything done within the projection, including the enlargement of a pattern, is a way of bringing that about.

10. CONCLUSION

A projection from many dimensions to three can be computed; its inverse cannot. No smooth or one-to-one map from three dimensions fills a larger space, and no processing increases the information that data carries about its source. A language model lies at the end of a chain of such projections and computes within the last of them. It can recover the outline of thought,

because the outline is what the chain preserves; it cannot recover what the chain discarded, and on the thesis defended here, understanding and consciousness are among what it discarded. Many dimensions can be projected into three. Three cannot be projected back into many.

REFERENCES

- Bender, E. M. and Koller, A. (2020). Climbing towards NLU: on meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.
- Brouwer, L. E. J. (1911). Beweis der Invarianz der Dimensionenzahl. *Mathematische Annalen* 70, 161–165.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2, 200–219.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd edn. Hoboken, NJ: Wiley.
- Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346.
- Hurewicz, W. and Wallman, H. (1941). *Dimension Theory*. Princeton: Princeton University Press.
- Kaplan, J. et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.
- Peano, G. (1890). Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen* 36, 157–160.
- Sard, A. (1942). The measure of the critical values of differentiable maps. *Bulletin of the American Mathematical Society* 48, 883–890.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 417–457.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423 and 623–656.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 59, 433–460.
- Vaswani, A. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30, 5998–6008.