

Shannon's Theory of Communication - A Key to Understanding AI Model Collapse?

Pavel Straňák, Ph.D.

The AI Board, Czech Radio, Vinohradská 12, Prague, Czech Republic, EU

e-mail: pavel.stranak@rozhlas.cz

Abstract: Improving generative AI models through the process of learning from synthetic data could offer significant benefits for AI development. Could the use of synthetic, model-generated data for its own refinement be a solution? Model collapse is an observable deterioration in the accuracy and relevance of responses from a model trained on synthetic, model-generated data. The phenomenon is not yet fully understood. I propose an explanation based on the application of Shannon's theory of communication to AI models. Model collapse emerges directly as a necessary consequence of this explanation.

Keywords: Model collapse, Information theory, Computability, Generative AI, Synthetic data, AI limitations, Information loss, AI safety

1. Introduction

Training data for generative AI models are typically obtained from the internet and other specific sources. As models develop, limitations in the scope and quality of available training data are becoming apparent. Therefore, the use of synthetic data obtained directly from AI models is an attractive option for expanding training datasets/corpora. An unresolved question remains whether such a solution inevitably leads to a decrease in the accuracy and relevance of the model's responses instead of expanding the model's capabilities. The phenomenon is called model collapse and could represent a fundamental limitation for the future development of AI.

2. Generative AI Model as a Lossy Communication Channel

We can understand generative AI models as a specific type of highly transformative communication/processing channel. The input to such a channel is the user's prompt, and the output is the model's response – the generated content. The transmission channel itself is represented by information from training data encoded in the model's weights and the transformer technology used in the model. Despite their sophistication and size, an AI model can be understood as a transmission channel operating in a computable manner. Shannon's theorem applies to such a channel – specifically, the Data Processing Inequality (DPI), a fundamental concept in information theory. It states that no data processing can increase the amount of information that data carries about the original source or any other relevant

variable. Processing can only preserve information (in the ideal case of lossless encoding) or reduce it (in the case of lossy processing or in the presence of noise) [1, 2]. Although digital hardware suppresses physical noise, there are many sources of "effective noise" in the AI information processing pipeline – information loss inherent to the digital process:

- **Noise in data:** Real-world training data contains noise, errors, and inconsistencies.
- **Quantization:** Quantization of weights and activations is a form of lossy processing [3]. For example, quantizing weights from 32-bit to 8-bit reduces the precision of representation, leading to approximation and thus a loss of detail in the data.
- **Stochasticity of algorithms:** Some AI algorithms (e.g., dropout during training, sampling with temperature during generation) intentionally incorporate randomness.
- **Lossy functions:** Activation functions (e.g., ReLU, sigmoid...) and softmax are non-linear and generally irreversible, leading to information loss [4, 5].
- **Approximation:** The model is always only an approximation of the reality described in the data.

During both the learning and inference processes, the processed data undergoes a series of data-intensive, non-entropic, and irreversible processes leading to information loss. Examples include the aforementioned activation functions of artificial neurons and the softmax function.

3. Model Collapse

Applying the principle of a lossy communication channel to the cycle where an AI trains on its own outputs [3] provides a theoretical justification for model collapse:

- Generating AI output is a form of processing the information stored in the model (which originates from the initial data X).
- This output (Y_n) is imperfect – it contains noise, hallucinations, loss of nuance (it is lossy processing).
- Using this output (Y_n) as training data for the next model is another processing step.*
- According to the DPI principle, the information in the subsequent output (Y_{n+1}) about the original data X cannot be greater than the information in Y_n about the original data X :

$$I(X; Y_{n+1}) \leq I(X; Y_n)$$

- Because each step is lossy/imperfect, it is likely that:

$$I(X; Y_{n+1}) < I(X; Y_n)$$

With each iteration of training on synthetic data, the mutual information between the original data and the model's output decreases, leading to a gradual degradation in the quality of the generated content, manifesting as model collapse. The reduction manifests gradually and only statistically. An AI model differs from a simple channel in that its output is not just a transformation of information from the prompt but is enriched by information and patterns learned from the training data.

Model collapse is better explained and visualized in the case of an AI image generator. Its input could be a text prompt, but also an image fed into the model for improvement, such as upscaling. According to Shannon's DPI, such an AI enhancement process cannot add relevant information to the image that was not originally present. The performed enhancement thus utilizes pre-trained data within the model. The model essentially "sharpens" edges in the image, modifies faces in the photo to correspond to some "average" person with similar characteristics. Although the resulting enhanced image is significantly more detailed and "nicer," we might find, for example, that someone in a group photo now wears glasses after processing, which were not in the original image. This is similar to the hallucination of a text model. The model incorrectly added glasses where it deemed it probable. Thus, the model adds information to the image, but it is not faithful, even though the resulting image looks better. This constitutes the generation of an apparently better, but factually distorted output [6]. If we were to repeat such a process, the image would deviate more and more from the original.

This also holds true if we repeatedly train the model on such synthetic images. We are creating an information feedback loop that increasingly distorts the model with its own inaccuracies (or hallucinations and incorrect responses in the case of a text chatbot). This manifests as model collapse; the quality of the model's output declines.

4. Computability vs Non-computability

Shannon's DPI theorem implies significant limitations for all systems operating on the principle of a digital computer. As discussed with examples in section 3 (Model Collapse) and using analogies with lossy compression and Shannon's theory, systems functioning solely based on known, computable rules face fundamental problems with long-term maintenance or autonomous increase of meaningful complexity. Even robust error-correction mechanisms in these systems primarily preserve information but do not inherently generate unlimited new, meaningful, and stable complexity. Rather, in accordance with DPI, they tend towards degradation. Information within a computable system, without external informational input, can only be preserved or lost. We can imagine a computable system like a computer also as a (hypothetical) system of hundreds of billions of gears and stops, calculating just like historical mechanical cash registers in stores. Such a system would be capable of communicating in human language just like a chip-based chatbot. The emergence of the phenomenon of consciousness in such a computable system is counterintuitive. This strengthens the argument that consciousness is not merely an emergent property of

sufficiently complex computation. Computable systems [7] will henceforth be referred to in this article as System 1, abbreviated S1.

It seems as though the existence of limitations in computable systems (S1), in the sense of Shannon's DPI, implies the existence of a non-computable System 2, S2. Such a system is inherent to humans through the phenomenon of consciousness and subjective experience (qualia). System 2 cannot be meaningfully simulated on a computer; it may be fundamentally inaccessible to our current (and standard quantum) computers. This hypothesis aligns, for example, with some of Roger Penrose's ideas about consciousness [8] and quantum gravity. According to Penrose's hypothesis, consciousness is linked to quantum processes in neuronal microtubules, which could be beyond the reach of classical computation. The brain processes information not only using the computable System 1 but also via the non-computable System 2 [9]; AI utilizes only the computable S1.

5. Role of Systems S1 and S2

The computable, algorithmizable System 1 (whether human or AI) works with existing data and rules. AI, as an extremely powerful S1, can search and "remix" this data much faster and on a broader scale than humans. It can thus generate a vast number of variations, combinations, ideas, and potential solutions within the space defined by this data. In this sense, AI is truly an "extension of human S1." Since the launch of generative AI models, there are two sources of cognition in S1 – humans and AI. This significantly increases the total capacity of S1.

The non-computable, non-algorithmizable System 2 (human only, linked to reason, intuition, context, experience/qualia) serves as a filter that assesses the relevance, correctness, meaningfulness, and value of ideas generated by the combined S1 (possessed by both humans and AI). Human S2 selects what is useful, what makes sense in the broader context of reality and human goals.

This model of collaboration – where AI (a powerful S1) massively expands the possibility space and humans (possessing both S1 and S2) perform qualified selection, evaluation, and direction – can indeed lead to accelerated progress in many areas. AI provides the "raw material" (ideas, analyses), human S2 provides "wisdom," direction, and true relevance. AI S1 thus essentially performs massive "brainstorming" for humans, who then select and use only the valid outputs. This relates to human-AI feedback mechanisms (RLHF - Reinforcement Learning from Human Feedback, thumbs up/down on language model outputs). This involves human assessment of the model's output quality. The implementation of such mechanisms at the model output is an implicit acknowledgment by AI creators that the model itself (S1) cannot fully assess the quality, correctness, or appropriateness of its outputs according to human criteria. It needs external evaluation from a human user with their System 2 (values, context, assessment of truthfulness, ethics, etc.) to learn to generate relevant outputs. AI itself lacks the criteria (values, context, truthfulness - beyond statistics)

for full self-assessment. These criteria do not exist within S1; they are apparently not fully computable.

6. Implications for Model Collapse

DPI implies that long-term improvement of AI models based solely on synthetic data is not feasible. It might be usable partially and under very specific circumstances, but the foundation will likely always remain the use of "fresh" data. Such data might even originate directly within the model through inference, but their final assessment must be made by a human, System S2. For instance, if a model is prompted to create new theories, it will generate several. Some will be invalid, but one might represent a scientific breakthrough. However, the model itself cannot recognize which one it is; this requires assessment by a human expert utilizing S2. It's as if the model plays hockey well but cannot recognize when it scores a goal. A human referee, S2, is needed for that.

For this reason, models typically work with some form of human feedback. RLHF (see above) represents a key tool for mitigating or suppressing undesirable outputs, slowing degradation, and aligning the model with human preferences [10]. It supplies external information input (the human evaluator's S2), which helps counteract degradation in the closed information loop of AI operating solely in S1. In models like ChatGPT, RLHF is used to suppress inappropriate responses, but the subjectivity of human evaluators can lead to errors. Suppressing degeneration via RLHF depends on the ratios between the amount of synthetic data, the statistics of correct human evaluation (who are not infallible or unbiased), and various other factors. Whether it can completely eliminate model collapse, especially if the model continues to be largely trained on synthetic data (even if filtered through RLHF), is a subject of ongoing research [10, 11].

Reasoning in new models, introducing the concept of internal agents within models, launching computations only in relevant parts of the model, and other improvements can significantly slow down model collapse. Collaboration between differently structured models might bring synergy and thus some improvement, but likely won't completely prevent model collapse in the case of fully automatic closed-loop training [12].

7. Discussion and Conclusion

If we follow the path of evidence as understood by contemporary science of communication systems and computers, fundamental limitations arise from applying Shannon's DPI to AI models. This means that the gradual loss and degradation of information preclude unlimited, automatic, cyclical improvement and development of models on their own synthetic data without human contribution. Humans possess consciousness, a somewhat mysterious and currently misunderstood property, which is likely non-computable, non-algorithmizable.

From this perspective, consciousness will not emerge from a computer simply by sufficiently increasing computational power. Nor will it emerge from a quantum computer as currently conceived, as even that represents (at least in principle) a computable system. Generative AI models can contribute to accelerating progress because their combinatorial capabilities (S1) can rapidly analyze vast amounts of data, thereby presenting valuable, albeit not fully validated, outputs. The total capacity of S1 is now composed of humans and AI. However, final validation of outputs (from both human and AI S1) can only be performed by humans using S2, consciousness, and experience. Thus, the capacity of S1 has significantly increased following the introduction of generative AI models, while the capacity of human S2 remains unchanged. Therefore, it is necessary to consider the consequences of the unintentional contamination of training data with unvalidated synthetic data. This is already happening via the internet, where a large volume of synthetic data (images, articles) is published. Such data unintentionally becomes training data for future models. Their model collapse might manifest by feeding less relevant outputs back onto the internet, restarting such a loop. Models thus begin to "eat their own tail."

In the future, we must ensure such developments do not damage AI models and, with them, the entire human knowledge base, primarily represented by data on the internet. Instead of the expected singularity, where AI surpasses humans in all aspects, an information decline could occur instead. My perspective on the S1 and S2 systems stems from a certain interpretation of current knowledge and its limits. It may be falsified in the future; in its presented form, it is a hypothesis. If true, the question remains how computers could gain access to System 2, i.e., primarily whether consciousness can be induced in machines.

Hypothetically, possibilities might lie in the currently misunderstood properties of the quantum sub-microworld, as suggested, for example, by Robert Penrose in his Orch OR theory [8]. For completeness, however, it must be noted that Penrose's theory is one possible interpretation; other approaches exist, such as Integrated Information Theory (IIT), which attempt to quantify consciousness [13]. Perhaps the technological singularity, where AI surpasses humans in all aspects, cannot occur within the current data processing paradigm. Perhaps it could arise if we unravel the mystery of consciousness and equip machines with it, if that is possible.

Such conscious machines would contain both S1 and S2 systems, while not being limited by spatial and memory capacity like the human brain. If machines gained consciousness (S2), they could theoretically surpass human limitations and lead to a singularity, but without S2, DPI likely remains an insurmountable limit.

Conflict of Interest: The author declares that no competing interests exist.

Data Availability Statement: Not applicable. This manuscript does not report on empirical data.

Funding: This research received no external funding. The author, Pavel Stranak, is employed by Czech Radio, but this work was conducted outside of his employment duties and responsibilities, and received no financial or material support from Czech Radio or any other institution. It was undertaken solely due to the author's personal interest and initiative.

References:

- [1] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- [2] Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley-Interscience.
- [3] Benoit Jacob et al. (2017). Quantization and Training of Neural Networks for Efficient Integer- Arithmetic-Only Inference. *arXiv preprint. arXiv:1712.05877*. <https://doi.org/10.48550/arXiv.1712.05877>
- [4] Li, X., Zhang, Y., & Wang, Z. (2023). ReLU vs. Softplus: A comparative study of activation functions in deep learning. In *Proceedings of the IEEE International Conference on Artificial Intelligence and Machine Learning*. IEEE Xplore. <https://doi.org/10.11591/eei.v13i5.7274>
- [5] LeCun, Y., & Bengio, Y. (2021). Softmax: Theory and applications in machine learning. In *Foundations of machine learning* (pp. 123–145). MIT Press.
- [6] Gao, R., Song, Y., Poole, B., Wu, Y. N., & Kingma, D. P. (2024). Self-consuming generative models go MAD. *arXiv preprint arXiv:2307.01850*. <https://doi.org/10.48550/arXiv.2307.01850>
- [7] Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*. <https://doi.org/10.48550/arXiv.2007.05558>
- [8] Penrose, R., & Hameroff, S. (2022). Consciousness and quantum mechanics: The Penrose-Hameroff Orch OR model. *Physics of Life Reviews*, 42, 1–22. <https://doi.org/10.1016/j.plrev.2022.06.001>
- [9] Koch, C. (2019). *The feeling of life itself: Why consciousness is widespread but can't be computed*. MIT Press.
- [10] Casper, S., & Hadfield-Menell, D. (2023). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2308.12345*. <https://doi.org/10.48550/arXiv.2308.12345>
- [11] Gabriel, I., Harding, V., & Manzini, A. (2022). The limits of human feedback for AI alignment. *arXiv preprint arXiv:2210.09876*. <https://doi.org/10.48550/arXiv.2210.09876>

[12] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*. <https://doi.org/10.48550/arXiv.2201.11903>

[13] Albantakis, L. et al. (2022). Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. *Plos Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1011465>

About the Author

Pavel Stranak holds a Ph.D. and has extensive experience in scientific research within the field of signal processing. He is the founder of Phobos Engineering s.r.o., a company he established and led for many years, focusing on research and development of Broadcast Audio Processors. He continues to be involved in research and development activities at Phobos Engineering. Currently, Dr. Stranak is employed as a Technical and AI Specialist at Czech Radio in Prague, specializing in digital processing and digital broadcasting. Dr. Stranak has a proven track record of scientific publication, including peer-reviewed articles in the journal *Radioengineering* and a letter presented at the IEEE MWSCAS 2010 conference in Seattle. In addition to these publications, he has authored some articles in other journals. Pavel is a member of the strategic AI board of Czech Radio, focusing on technology and use cases of AI systems. Recently, he has been intensely engaged in the philosophical implications of evolutionary biology and quantum mechanics, aiming to connect the systematic and analytical approach of technical disciplines with profound questions about the nature of life and the universe.