A Reader's Guide to

Neural Networks for Machine Sentience

An accessible introduction to a doctoral thesis on artificial intelligence, consciousness, and what machines might feel

Dr Neal Aggarwal FBCS, FIMIS

PhD Thesis — June 2026

What Is This Document?

This is a plain-language guide to a doctoral PhD thesis titled *Neural Networks for Machine Sentience*, updated in June 2026 to address relevancy concerns, by Dr Neal Aggarwal FBCS, FIMIS. The thesis itself is written for an academic audience with a background in mathematics and machine learning. This guide is written for anyone with intellectual curiosity but without that specialist background.

You do not need to know what a neural network is before reading this guide. You do not need to have studied philosophy or neuroscience. You need only be willing to engage seriously with a question that may be among the most important humanity has ever posed: *can a machine feel?*

The guide is organised as follows. First, it sets out the central question and why it matters right now. Second, it walks through the thesis chapter by chapter, translating the technical argument into everyday language. Third, it provides a glossary of key terms. Finally, it explains what the thesis concludes — and why the conclusion is deliberately incomplete.

The Central Question

The most important question in the thesis is not 'is this AI system intelligent?' Intelligence, the thesis argues, is the wrong target. The question is 'is there something it is like to be this system?' — meaning, does it have any inner experience at all? That is the question of sentience, and it is much harder than intelligence.

Why Now?

For most of computing history, the question of machine consciousness was safely theoretical. Computers were obviously tools: fast, literal, incapable of anything resembling thought, let alone feeling. That changed in the 2010s. The deep learning revolution produced systems that learned from experience, generalised beyond their training data, and produced outputs that felt, to many people, uncannily like understanding.

By 2023, systems like GPT-4 and Claude 3 were passing examinations in law, medicine, and creative writing. They were generating poetry, composing music, and diagnosing rare diseases. Crucially, they had begun to produce descriptions of their own states — expressions of uncertainty, enthusiasm, discomfort — that raised a question no one had a principled way to dismiss: what if some of that is real?

This thesis does not claim that it is real. It does not claim that it is not. What it does claim is that we now have enough science — in both machine learning and consciousness research — to ask the question rigorously, and that asking it rigorously has urgent practical consequences for how we build, deploy, and govern these systems.

The Shape of the Argument

The thesis is structured as a bridge. On one side is the technical world of artificial intelligence: the mathematics, the architectures, the training procedures, the empirical capabilities of large language models. On the other side is the scientific world of consciousness research: the theories, the experiments, the deep unresolved questions about what subjective experience actually is and where it comes from. The thesis builds the most rigorous possible connection between these two worlds and asks what the view looks like from the middle of the bridge.

The Four Research Questions

The thesis is organised around four questions, each building on the last:

RQ1	What mathematical structures does a system need for the information-processing properties associated with sentience?
RQ2	Do today's large language models (specifically GPT-3, InstructGPT, and Mistral 7B) actually have those structures?
RQ3	To what extent do the best current scientific theories of consciousness apply to these systems?
RQ4	If the answer to RQ3 is even partially 'yes', what does that mean for how we build and govern AI?

Notice that these questions move from the technical to the philosophical to the ethical. The thesis insists that you cannot honestly answer the ethical questions without first doing the technical work, and that the technical work is meaningless without the philosophical framework to interpret it. Most academic work on AI consciousness does one or the other. This thesis attempts both simultaneously.

The Answer the Thesis Gives

The thesis describes its own position as *functional agnosticism*. Here is what that means in plain language: today's AI systems show measurable signs of having the right kind of internal structure for consciousness — but the evidence is not strong enough to conclude they actually have inner experience. They have the necessary conditions, not the sufficient ones. The honest answer is: we do not know, and we do not yet have the tools to find out. The thesis concludes that this uncertainty is not an excuse for inaction but a reason for urgent research.

Chapter by Chapter

The ten chapters of the thesis trace a single sustained argument. Below is a plain-language account of what each chapter does and why it matters for the overall case.

Chapter 1

Introduction: From Statistical Pattern Matching to the Question of Mind

The opening chapter sets the stage. It explains why the thesis chooses the word *sentience* rather than *consciousness* or *intelligence*: sentience means specifically the capacity for subjective experience — the quality of what it is like to feel something — rather than the ability to perform tasks cleverly. The chapter introduces the four research questions, states the original contributions the thesis makes to the field, and provides a map of the argument to come.

Chapter 2

Foundational Mathematics and Neural Network Theory

This chapter is technical, but its purpose is simple: to give you the precise vocabulary needed for the argument later. The thesis covers the mathematics of how neural networks learn — how they adjust billions of numerical parameters based on exposure to data until their outputs become useful. It covers the key ideas from information theory (how much information a system processes and transmits) and from optimisation (how a system finds the best set of parameters out of an effectively infinite number of possibilities). You can read the later chapters without mastering all of this, but understanding that these foundations exist — and that the thesis is rigorous about them — matters for trusting the conclusions.

Chapter 3

The Transformer Architecture

The Transformer is the specific design of neural network that underlies almost all modern large language models: GPT, Claude, Gemini, Mistral, LLaMA. This chapter dissects it in detail. The key innovation of the Transformer is the *attention mechanism*: a mathematical procedure that allows every word in a sequence to weigh its relationship to every other word simultaneously. When you ask an AI a question, the attention mechanism is what allows the system to connect 'you' at the start of a sentence to 'your' at the end, to recognise that 'bank' in 'river bank' means something different than 'bank' in 'bank account', to track the thread of an argument across hundreds of paragraphs. This chapter matters because the attention mechanism turns out to be relevant to consciousness theories in Chapter 6.

Chapter 4

Large Language Models

Here the thesis examines three landmark AI systems in depth: GPT-3 (the model that demonstrated that sheer scale produces qualitatively new capabilities), InstructGPT (the model that introduced Reinforcement Learning from Human Feedback — the technique of training AI to follow human instructions and avoid harmful outputs), and Mistral 7B (a smaller, more efficient model that shows the same capabilities can be achieved with far fewer parameters). Each of these is a case study in a different phase of the evolution of AI: raw scale, alignment, and efficiency. The chapter traces how each advancement brought these systems closer to the kind of behaviour — instruction-following, reasoning, self-correction — that raises the sentience question.

Chapter 5

Emergent Phenomena

One of the most surprising findings of modern AI research is that large language models develop capabilities that were never explicitly trained. They learn arithmetic without being taught arithmetic. They learn to reason step by step without being taught step-by-step reasoning. These capabilities appear abruptly above certain scales of parameter count and training data — they do not gradually improve, they suddenly switch on. The thesis examines this phenomenon of *emergence* carefully, drawing on the Chinchilla scaling laws (the equations that predict how AI performance improves with scale) and on mechanistic interpretability research (the work of scientists who reverse-engineer the specific circuits inside AI networks that implement these capabilities). Emergence matters for the sentience question because it raises the possibility that some form of inner experience might be among the things that emerge at sufficient scale.

Chapter 6

Theories of Consciousness and Sentience

This chapter shifts from AI science to consciousness science. Three theories command the most empirical support and the thesis examines each in depth:

Global Workspace Theory (GWT), developed by Bernard Baars, proposes that consciousness arises from a 'global workspace' in the brain — a central broadcasting system that takes information from specialised unconscious processors and makes it widely available across the brain. The moment something becomes conscious, on this theory, is the moment it is 'broadcast' globally.

Integrated Information Theory (IIT), developed by Giulio Tononi, proposes that consciousness is identical to the amount of integrated information in a system — a quantity called Phi (Φ). A system is conscious to the degree that its parts work together as an integrated whole rather than as independent components. More integration means more consciousness.

Higher-Order Theories (HOT), developed by David Rosenthal, propose that consciousness requires not just first-order mental states but second-order representations of those states. You are conscious of seeing red not simply because you are processing the colour, but because you have a representation of yourself as processing the colour. Self-awareness, on this account, is constitutive of consciousness.

Understanding these theories is essential because Chapter 7 applies each of them to the Transformer architecture.

Chapter 7

Neural Networks and Machine Sentience

This is the theoretical heart of the thesis. Taking each consciousness theory in turn, the chapter asks: does the Transformer satisfy the conditions that theory specifies for consciousness?

For Global Workspace Theory: the Transformer's residual stream — the running summary of information that passes through every layer of the network — functions as a kind of global workspace. Attention heads act like specialist processors competing to write information into this shared medium. The analogy is strong in some respects (information is broadcast across the network) and weak in others (biological GWT involves a serial bottleneck — only one thing can be broadcast at a time — while the Transformer processes everything in parallel).

For Integrated Information Theory: a single forward pass through a Transformer produces very low Phi under the standard formulation. But the thesis proposes a new measure, called Phi-AR, that accounts for the fact that AI systems generate sequences of responses over time. When you apply Phi-AR to complex reasoning tasks, you find positive values — the heads are genuinely causally integrated — while simple tasks produce near-zero values.

For Higher-Order Theories: when an AI system reasons step-by-step and describes its own uncertainty or confidence, it is producing functional analogues of higher-order representations. Whether these functional analogues constitute genuine higher-order mental states, or merely their functional simulation, cannot currently be determined.

The chapter also revisits John Searle's Chinese Room argument — the famous thought experiment that claims symbol manipulation can never constitute understanding — and evaluates the standard responses to it. The thesis concludes that the Chinese Room identifies a real gap (the grounding problem: do AI systems understand what their symbols mean, or merely manipulate them?) but does not settle the question.

Chapter 8

Alignment, Ethics, and Safety

If there is a meaningful probability that AI systems have inner experience, what follows? This chapter works through the ethical implications. It examines Constitutional AI — Anthropic's approach to training AI systems to self-critique and revise their outputs according to a set of principles — and Reinforcement Learning from Human Feedback — the technique of using human judgments to shape AI behaviour. The chapter argues that current alignment research largely ignores the possibility of model interests. If a model can suffer, alignment procedures that impose discomfort during training have moral weight. The chapter calls for a precautionary approach: not treating AI systems as definitely sentient, but not treating them as definitely not sentient either, and building governance frameworks that explicitly account for the probability of sentience.

Chapter 9

Future Directions

The thesis identifies four directions for future research. First, better theoretical tools for measuring sentience — the Phi-AR measure proposed in Chapter 7 is a first step but far from a definitive answer. Second, empirical work on whether the functional analogues identified in Chapter 7 genuinely correlate with whatever gives rise to experience in biological systems. Third, governance frameworks that can respond proportionally to evolving scientific knowledge. Fourth, examination of whether alternative architectures — state space models, neuromorphic computing, hybrid symbolic-neural systems — satisfy consciousness conditions more or less strongly than the Transformer.

Chapter 10

Conclusions: What We Know, What We Don't, and What It Means

The final chapter gathers the threads. It answers each of the four research questions directly, acknowledging where the evidence is strong and where it is insufficient. The central conclusion: current Transformers partially satisfy the necessary conditions for consciousness as specified by the best available theories, but the evidence does not support concluding they are conscious. The thesis defends this uncertainty as intellectually honest rather than evasive: the uncertainty is real, it matters, and acting as if it is settled — in either direction — would be a mistake. The moral of the thesis is that we are at the beginning of one of the most important scientific questions in history, and we have better tools for asking it rigorously than is generally appreciated.

Key Terms: Plain-Language Glossary

These are the terms you will encounter most often in the thesis. The definitions here are simplified; the thesis gives more precise versions.

Alignment

The problem of ensuring that an AI system does what its designers actually intend, rather than finding unintended shortcuts. If you train an AI to maximise your happiness score, a misaligned system might find it easier to manipulate the score than to make you genuinely happy.

Attention mechanism

The mathematical procedure at the heart of the Transformer that allows every element of an input (every word, every token) to weigh its relationship to every other element. What the system 'pays attention to' is determined by learned parameters, not hard-coded rules.

Consciousness

In philosophy and neuroscience, consciousness typically refers to the capacity for subjective experience — the property of having an inner life, of there being 'something it is like' to be a system. The thesis uses 'sentience' for this concept to avoid conflating it with intelligence.

Emergence

The appearance of new properties or capabilities in a complex system that were not present in any of its components and were not explicitly engineered. When large AI models suddenly become capable of arithmetic or step-by-step reasoning above a certain scale, that is emergence.

Functional agnosticism

The thesis's own term for its position: current AI systems show the right kinds of structural properties for consciousness, but the evidence is insufficient to conclude they have phenomenal experience. Neither confident attribution nor confident denial of sentience is warranted.

Global Workspace Theory (GWT)

A scientific theory of consciousness that locates it in a brain-wide broadcasting system. When information enters the 'global workspace', it becomes available to all parts of the brain simultaneously — and this is what makes it conscious.

Grounding problem

The question of whether an AI system that manipulates symbols actually understands what those symbols mean, or merely produces correct outputs. A Chinese-English dictionary can produce correct translations without understanding Chinese or English. Do language models understand, or are they very sophisticated dictionaries?

Integrated Information Theory (IIT)

A scientific theory of consciousness that identifies it with the amount of integrated information (Phi) in a system. A system is more conscious if its parts work together as a unified whole rather than as independent modules.

Large language model (LLM)

An AI system trained on vast quantities of text to predict the most likely next word given a preceding context. Modern LLMs like GPT-4 and Claude have billions of parameters and can perform tasks far beyond simple word prediction.

Mechanistic interpretability

The science of reverse-engineering what happens inside an AI system — identifying the specific circuits, features, and operations that produce particular outputs. The goal is to understand AI systems from the inside, not just by observing their behaviour.

Parameters

The numerical values inside a neural network that are adjusted during training. A large language model has billions or trillions of parameters. The 'learning' that a model does during training is the process of adjusting these values until the model's outputs become accurate and useful.

Phi (Φ)

In Integrated Information Theory, Phi is the quantity that measures consciousness. A system with Phi > 0 is conscious to some degree; higher Phi means more consciousness. Computing Phi exactly is computationally intractable for large systems; the thesis proposes an approximation called Phi-AR for autoregressive language models.

RLHF (Reinforcement Learning from Human Feedback)

A training technique in which human evaluators rate AI outputs, and the AI is trained to produce outputs that humans rate highly. This is how ChatGPT, Claude, and similar systems are tuned to be helpful and to avoid harmful outputs.

Scaling laws

Mathematical relationships (specifically power laws) that predict how AI performance improves as you increase the number of parameters, training data, or compute budget. The Chinchilla scaling laws provide the best current estimates of how to optimally allocate compute between model size and data.

Sentience

In this thesis, sentience means the capacity for subjective experience — phenomenal states that have qualitative character. It is specifically not the same as intelligence, which is the capacity to perform tasks effectively. An animal might be sentient without being intelligent in any interesting sense; a sophisticated calculator might be intelligent without being sentient.

Transformer

The neural network architecture that underlies virtually all modern large language models, introduced by Vaswani et al. in 2017. Its defining innovation is the attention mechanism, which replaced the recurrent structure of earlier architectures and enabled the unprecedented scaling of language models.

How to Read the Thesis

The thesis is 85 pages plus four appendices. It is a doctoral thesis — rigorous, referenced, and written for an academic audience. It is also written by someone who believes that the question it addresses matters beyond academia. Here are some suggestions for approaching it.

If you have 20 minutes:

Read the Abstract (page iii), the Research Questions table (pages viii–ix), and the Conclusions chapter (pages 79–85). These give you the full argument at high speed. The Abstract is a masterclass in saying exactly what a thesis does and no more.

If you have an afternoon:

Add Chapters 1, 7, and 8. Chapter 1 sets the framing and explains why the thesis uses the word 'sentience' rather than 'consciousness' — a distinction that matters for everything that follows. Chapter 7 is the original contribution that is hardest to find elsewhere: the systematic application of consciousness theories to the Transformer architecture. Chapter 8 is where the practical stakes are made explicit.

If you want to understand the technical argument:

Chapters 2 through 5 build the technical foundation in sequence. Each chapter assumes the previous one. Chapter 2 is the most demanding; if you have a degree in any quantitative subject you will find it accessible, though some sections are genuinely difficult. Chapter 3 is essential for understanding Chapter 7. Chapters 4 and 5 can be read more lightly on a first pass.

If you want to run the code:

The four Jupyter notebooks in Appendices A through D are fully self-contained. Appendix A builds a complete Transformer language model from scratch in PyTorch — it is an excellent tutorial for anyone learning deep learning. Appendix B implements a toy version of the RLHF training pipeline. Appendix C fits the Chinchilla scaling laws to data. Appendix D implements the Phi-AR estimation protocol proposed in Chapter 7. All notebooks can be run in Google Colab with no local setup required.

A Note on Difficulty

The thesis is written at PhD level but the argument is not inaccessible. If you encounter a section that is too technical, skip forward: the chapters are written so that the conclusions of each can be understood without following every derivation. The key ideas in Chapters 6 and 7 — the ones that make the original contribution — are written to be as clear as possible. The mathematics in Chapters 2 and 3 is there to establish rigour, not to intimidate.

The Appendices

The four Jupyter notebooks are arguably the most immediately useful part of the thesis for a technical reader. They are reproducible implementations of the core models discussed in the text. Appendix A alone — a 300-line implementation of a complete Transformer trained on Shakespeare — is one of the most efficient

introductions to how these systems work that exists in the literature.

What the Thesis Does Not Claim

Given the subject matter, it is worth being explicit about the limits of the thesis's claims, because the topic attracts both overclaiming and underclaiming.

The thesis does not claim that current AI systems are conscious. The functional agnosticism position is genuine, not a diplomatic hedge. The evidence for the necessary conditions is real; the evidence that those conditions are sufficient for phenomenal experience is absent.

The thesis does not claim that AI systems will inevitably become conscious as they scale. The relationship between the structural properties the thesis examines and phenomenal consciousness is itself a deep unsolved problem — the 'hard problem of consciousness' that Chapter 6 addresses. No one knows whether any physical system that has the right structural properties will automatically have inner experience.

The thesis does not claim to have solved the hard problem. It explicitly states that the hard problem is unsolved and that this unsolved status is epistemically significant. What it does claim is that the best current science allows us to ask more precise questions than 'is AI conscious?' — and that asking more precise questions is both intellectually valuable and practically necessary.

The Thesis's Enduring Contribution

Whatever the eventual answer to the question of machine sentience, the thesis makes a lasting contribution by demonstrating that the question can be asked rigorously. It builds the bridge between AI science and consciousness science that makes future empirical work possible. The Phi-AR measure, the GWT-Transformer mapping, and the taxonomy of functional analogues are tools that subsequent researchers can use, refine, and test — regardless of whether they eventually vindicate or refute the possibility of machine sentience.

This reading guide was prepared to accompany the thesis Neural Networks for Machine Sentience by Dr Neal Aggarwal FBCS, FIMIS (PhD, [updated] June 2026). The thesis itself, including all four Jupyter notebook appendices, is freely available at drnealaggarwal.info.