Welcome to an extraordinary journey aboard a philosophical caravel guided by the Author through 'seas never before navigated.' This odyssey through the tempestuous seas of the mind includes contributions from eight of the brightest international experts on the brain and consciousness, including Nobel Prize-winning physicist Sir Roger Penrose.

Divided into two parts, the book begins with Steven S. Gouveia plotting the fundamental coordinates of philosophy and the science of the mind. The second part comes to life in the Author's dialogues with the eight scholars, providing a unique experience of intellectual exploration. Prepare to set sail into the unknown, where science and philosophy converge in a captivating journey through the mysteries of the conscious mind.

Steven S. Gouveia holds a degree in Philosophy. At the age of 22, he began a Ph.D. in Philosophy of Mind with a focus on Neuroscience at the University of Minho. By the age of 30, he was appointed Honorary Professor at the Faculty of Medicine, Universidad Andrés Bello, Chile

Currently, he is a Hired Researcher at the Mind, Language, and Action Group at the Institute of Philosophy, Faculty of Arts and Humanities, University of Porto, where he leads a project on Ethics of Artificial Intelligence in Medicine

He was a visiting researcher at the Minds, Brain Imaging, and Neuroethics Unit at the Royal Institute of Mental Health, University of Ottawa, Canada.

As an author and editor, he has published 14 academic books and has been invited to speak at conferences around the world

He has participated in numerous television, radio, and podcast programs. More information: www.stevensgouveia.weebly.com.

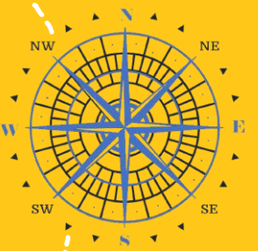
STEVEN S. GOUVEIA

THE ODYSSEY OF THE MIND

STEVEN S. GOUVEIA

David Chalmers

Susan Blackmore



Nicholas Humphrey

Sir Roger Penrose

THE ODYSSEY OF THE MIND

Dialogues on the BRAIN and CONSCIOUSNESS

Anil Seth

Karl Friston

Joseph LeDoux

Christof Koch

THE ODYSSEY OF THE MIND

Dialogues

on

BRAIN

and

CONSCIOUSNESS

Steven S. Gouveia

2024

Steven S. Gouveia
MLAG | Institute of Philosophy | University of Porto
Porto, Portugal

ISBN: 9798883665652

Cover: © Steven S. Gouveia.
Original drawings from the cover: © Ana Monteiro.
Transcriptions: © Tássia Vianna.
Publisher: Kindle Direct Publishing (Amazon).
Original publication: A Odisseia da Mente (Portuguese).

Table of Content

Preface

The aim of this book is to take the reader aboard a philosophical caravel,[1] accompanied by some of the greatest and most courageous navigators of this 21st century, on a brave expedition through the stormy seas of knowledge about the complex issues of the human mind.

This Odyssey of the Mind will take us, inspired by the bravery and courage of the intrepid explorers of the Maritime Discoveries, to brave the unknown waves and tides of consciousness, where each thought is a star that guides us in this night sky surrounded by mysteries perhaps never to be resolved.

As we raise the sails of curiosity, we will certainly find at least one island where we can pause the journey, and discover new colors, new species, more spices. But also, greater wonder and curiosity, and, therefore, more questions to be answered.

[1] The Portuguese caravel stood out for several innovations, such as the use of a triangular lateen sail that allowed greater ability to sail against the wind, or its light and agile hull that allowed it to be highly versatile. Furthermore, the Portuguese also developed more advanced navigation technology compared to that used until then, through the use of the crossbow and the astrolabe, which gave them the possibility of navigating to places in the world that had never been explored.

Perhaps every synapse in our brain can serve as a compass – like the one at the center of this book's cover – that points to unexplored directions and promising lands of greater understanding.

In this incessant search – in this great Odyssey of our century – we face a central enigma that permeates the seas of the mind: what gives consciousness its uniqueness?

As we venture between reefs of thoughts and bays of reflection, the map of consciousness is slowly drawn before our eyes.

Like the great navigators who charted new routes through unknown lands, we face the storms of uncertainty and complexity, while celebrating discoveries that broaden our horizons.

On each page of this book, vivid illustrations of experiences and ideas await, encouraging the reader to create their own caravel and explore “seas never sailed before”.[2]

Ultimately, this odyssey is not just about unlocking the mystery of consciousness or the nature of the brain, but, rather, about embracing the journey into the unknown

[2] Reference to the first verse of Canto I of the work *Os Lusíadas*, by Luís Vaz de Camões, a Portuguese poet who described the adventures of the Maritime Discoveries by the Portuguese in the XV century. The original expression is “Por mares nunca dantes navegados”.

through the astrolabe of science and the crossbow of philosophy.

The mind, like the ocean, is vast and unexplored, and our search is guided by the passion of discovery and the thirst to understand an unceasing mystery that seems to torment every century of our humanity.

Let's hope that, at the end of this trip, we have managed to find some good port, which will allow us at least a brief well-deserved rest.

To close this preface, a less poetic note: this book has a specific structure, being divided into two parts: the first part, dedicated to consciousness, and the second part, dedicated to the brain.

In each part, you will find two sections: the second will showcase dialogues with 8 internationally renowned experts on the covered topics, while the first will include thematic introductions. These introductions aim to help each reader appreciate the dialogues more easily, despite the complexity of the themes presented in this book

Therefore, let this *Odyssey of the Mind* begin.

Steven S. Gouveia
Palermo, Sicily, Italy
30 | 06 | 2024

PART I

Consciousness

PART I

Consciousness

I. INTRODUCTION: CONSCIOUSNESS

Imagine that the reader finds themselves, at this precise moment, on one of the most incredible beaches on planet Earth, such as the wonderful Copacabana beach in Rio de Janeiro, Brazil.

Imagine now that you are sipping a tasty *caipirinha* – the famous Brazilian drink – while enjoying the sunny day and the beautiful view of Sugarloaf Mountain.

While this experience is probably magical enough for most of us, something truly enchanting unfolds in that moment. What might that be? The magic of savoring a specific drink, experiencing joy and happiness, all unfolding from a very distinct perspective: YOUR point of view!

This is the magical part of this story: no one seems to be able to really explain how these conscious experiences of yours are happening in your mind. This is one of the greatest mysteries in the universe: how a collection of unconscious neurons can give rise to a unified sense of subjective experience – a Conscious Mind!

Because it is such a mystery, scholars and thinkers from various disciplines are employing a variety of tools and

methodological strategies to paint a clear picture of how your mind is shaped by your brain, your body, and your connection with the environment (the world!).

Human consciousness is, perhaps, one of the last inexplicable mysteries of the world today. Of course, it is not the only one: many other mysteries existed in the past, such as the mystery of the origin of the universe, the mystery of life and reproduction, time, space or gravity.

Although some of these mysteries may share the absence of a final answer, we can still contemplate them. These mysteries haven't vanished but have been understood because we knew how to ask the right questions.

However, concerning consciousness, we seem not to have reached that point: there is a profound confusion, an intense struggle of arguments and counter-arguments, and various theories attempting to explain the same phenomenon.

Recently, I had the privilege of delivering an accepted talk for the second time at one of the world's most important academic conferences on consciousness: the Science of Consciousness Conference.

Having previously participated in the 2019 edition in sublime Switzerland's Interlaken, this second participation occurred in May 2023, in Taormina — a

small and wonderful town in Italian Sicily overlooking the Etna volcano.

As a speaker at this conference, on the first day, I received a specific code and a website link where I could vote for the theory of consciousness that I believed to be most likely correct.

The reader will be amazed that, upon opening the online page, I encountered more than 13 (!) options to choose from – more than 13 theories of consciousness were available for selection. As you can imagine, if there are 13 theories to explain a particular phenomenon in the world, something must certainly be amiss.

But what, then, is the mystery of consciousness? Why does it seem so challenging – for both science and philosophy – to formulate a plausible theory to explain its existence?

We all have subjective experiences: perceptions, sensations, pains and ideas. How can living physical bodies in a physical world produce such a phenomenon? These obscurities linked to consciousness make it one of the most exceptional and important problems of the Contemporary Era.

Some thinkers believe that consciousness may not be as mysterious as it initially seems. Drawing analogies from the history of science, they argue that perhaps all

this obscurity is a mistake created by our language and the way we use our mental concepts.[3]

For instance, if we look at the way the concept of "light" has been considered throughout history, perhaps we can learn something relevant related to consciousness.

The corpuscular theory of light, of Greek origin and based on ancient atomism, describes light as being composed of specific particles called "corpuscles".

This theory was championed by illustrious thinkers, including Sir Isaac Newton, who argued that all processes of reflection and refraction of light could only be explained if light were composed of particles. This was because the alternative – light being made up of electromagnetic waves – could not account for straight-line trajectories.[4]

Interestingly, this argument had some empirical basis: it is known that Newton conducted around 40 experiments that demonstrated the (supposedly) corpuscular nature of light.[5]

---

[3] For example: Churchland, P. (1986) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*, Cambridge: The MIT Press.

[4] Original publication: Newton, I. (1704) *Opticks: or, A Treatise of the Reflexions, Refractions, Inflexions and Colours of Light*, London.

[5] Original publication: P. Rowlands, P. (2017) *Newton and Modern Physics*, London: World Scientific.

This theory was considered true for centuries, until the scientist Thomas Young,[6] conceiving the famous Double Slit Experiment in 1801, refuted the corpuscular theory of light in favor of a wave theory. In this new perspective, light was considered to have a nature similar to that of sound, composed of electromagnetic waves.

This wave theory could explain many phenomena – such as diffraction or interference of light – that the previous theory could not elucidate. Now, in this case, the initial concept of "light" in corpuscular theory was, in a certain way, reduced to another concept of "light" – electromagnetic radiation – from wave theory.

This reduction has two specific meanings: on the one hand, corpuscular theory as a whole was reduced to wave theory; on the other hand, this reduction was so successful that we were able to reduce the concept of "light" to the concept of electromagnetic radiation.

Could it be that we only find this example of reduction in the history of science, in which we believed that it would make perfect sense to think about the existence of certain concepts, but which, after all, we later discovered to be pseudoconcepts?

[6] Original publication: Young, T. (2007) *Miscellaneous Works of the Late Thomas Young: Including His Scientific Memoirs*, Montana: Kessinger Publishing.

According to those who believe that the concept of light in corpuscular theory is on the same level as the concept of consciousness, there are more examples that can be provided.

The phlogiston theory was developed within the field of chemistry by Georg Stahl in the 18th century. Stahl argued that all combustible bodies would have, in their composition, an element called “phlogiston” that was released into the air during combustion processes.

This theory was an improved version of Stahl's mentor, Johann Becher, who published a book called *Physica Subeterranea*, where he argued that a specific element was released when a material burned.[7]

Note that this phlogiston theory was already an evolution of another ancient theory, Empedocles' theory of elements, which argued that there were five elements in the world: fire, earth, air, water and ether.

Once again, this theory would have a (naive) empirical basis. Stahl made a series of observations in the processes carried out in metallurgy: when a material combusted, it suffered corrosion. The greater the corrosion, the greater the amount of phlogiston released by that material.

[7] Original publicationl: Taylor, S. (2010) *Alchemists, Founders of Modern Chemistry*, Montana: Kessinger Publishing.

This theory was accepted by Stahl's peers due to its superior explanatory power compared to Empedocles' theory. The phlogiston theory made it possible to explain why an organic material lost mass when it burned: this was attributed to the loss of the phlogiston element to the air.

The inability to have combustion without air was also explained by this theory: this was due to the impossibility of phlogiston being captured by the air. Finally, the end of the combustion process was explained by the exhaustion of the presence of phlogiston in the material.

Once again, it took many years for the theory of phlogiston – and the very relevance of the concept of "phlogiston" to describe reality – to be called into question.

One of the problems with this theory was related to the fact that, despite the loss of mass being a reality in organic materials, the same did not happen with metals, which supposedly, in calcination processes (i.e., oxygenation), should also lose mass.

However, this specific prediction did not come to pass. On the contrary, Antoine Lavoisier demonstrated that metals, in the combustion process, could gain weight,

contradicting the central thesis of the phlogiston theory.[8]

Despite this contrary evidence, it was necessary to wait until the (accidental!) discovery of oxygen by Joseph Priestley in 1774. He realized that it was this chemical element responsible for combustion, rather than phlogiston.[9]

After several developments over the following decades, the chemical theory of the elements solidified and ended up completely replacing Stahl's theory, which lost any relevance in describing the world.

Once again, we are faced with a concept, “phlogiston”, that seemed to have total relevance. However, the investigation would later demonstrate that it was a pseudoconcept, a linguistic mistake that, when no longer considered, lost all its relevance.

Continuing with the history of chemistry – this time applied to biology – we can find a theory developed in the 18th century that argued that heat was an invisible, odorless fluid that all organic bodies possessed in their constitution.

Its quantity was directly correlated with the temperature of that body: a greater quantity of fluid

[8] Original publication: Bell, M. (2005) *Lavoisier in the Year One*, New York: Atlas Books.
[9] Original publication: Jackson, J. (2005) *A World on Fire: A Heretic, An Aristocrat and The Race to Discover Oxygen*, New York: Viking.

was equivalent to a higher temperature. Interestingly, this theory was advanced by the same person who had contributed to refuting the previous concept of 'phlogiston': Lavoisier had replaced this concept with another, that of 'caloric fluid.

As the phlogiston theory was inconsistent with Lavoisier's experimental results, he proposed a conceptual alternative: to consider this caloric fluid as the substance of heat.[10]

According to this theory, the caloric fluid would have a finite existence in its quantity, transferring from hotter bodies to colder bodies. Furthermore, it could not be created or destroyed, so the central thesis of this theory was its constant conversion, which explained the temperature interactions between different bodies.

The inability of the caloric theory to explain various phenomena, such as evaporation and sublimation, prompted the emergence of a more promising theory – the kinetic theory – through the work of Benjamin Thompson, also known as Count Rumford.[11]

Rumford observed that heat was not a material substance with a fluid form, but rather should be

[10] Original publication: Fox, R. (1971) *The Caloric Theory of Gases*, Clarendon Press: Oxford.

[11] Original publication: Brown, G. (2001) *Count Rumford: The Extraordinary Life of a Scientific Genius – Scientist, Soldier, Statesman, Spy*, Gloucestershire: Sutton Publishing.

considered as energy in motion. Through experiments demonstrating that heat could be generated by friction, he challenged the prevailing notion that heat was an indestructible fluid.

One of his most intriguing experiments involved the friction generated by drilling cannons, during which he observed that the amount of heat produced far exceeded what caloric theory would predict.

These observations, among others, played a crucial role in shaping the modern understanding of heat as a form of energy undergoing constant transformation. They significantly contributed to the development of the theory of energy conservation and, in the process, completely debunked the concept of the caloric fluid.

Interestingly, this kind of conceptual error is not only present in the history of chemistry: we can observe a similar phenomenon in the history of psychiatry. The theory of demonic possession, for centuries, served as an accepted explanation for numerous psychiatric illnesses.[12]

The theory of demonic possession was widespread during the Middle Ages, with various versions emerging in different periods and regions of the world. It served

[12] Original publication: Kemp, S. e Williams, K. (1987) "Demonic possession and mental disorder in medieval and early modern Europe", *Psychological Medicine*, 17 (1):21-9.

as an explanation for behaviors deemed abnormal by the standards of that time.

Exorcism was often seen as the only treatment available. In the historical documents that have endured, we can find a direct relationship between particular types of demons and symptoms associated with specific diseases, such as psychotic and neurotic disorders or epilepsy.

Curiously, certain medieval philosophers who significantly influenced other realms of philosophy and theology contributed significantly to the prominence of this theory.

For instance, Augustine of Hippo asserted that demons would possess a material body, while Thomas Aquinas argued that demons would be non-corporeal entities, intelligible and separate from the physical world.

Notably, Aquinas's stance was supported by the notion that numerous empirical phenomena described in literature could not exist if demons indeed had a corporeal existence.

This theory started to lose its relevance with the advancement of psychiatry as a discipline grounded in evidence and theories rooted in medical knowledge that evolved in the 20th and 21st centuries. However, it's important to note that, as explored in some of the topics within this book, the nature of psychiatric

illnesses continues to present significant challenges and questions for scientists and philosophers.

The key takeaway is that, with these historical cases highlighted – the existence of phlogiston, the caloric fluid, demons – society has come to recognize that, through the advancement of knowledge, we are now dealing with concepts devoid of meaning and relevance in understanding reality.

This point serves to illustrate a crucial consequence when we realize that, ultimately, the concepts employed are not beneficial for comprehending the world: the very questions and issues that shape the thinking of philosophers and scientists also become nonsensical and, as a result, lose the necessity for a response.

To grasp this essential point, let's examine the geocentric theory, which advocated that planet Earth was the center of the universe. According to this theory, the Sun and Moon orbited around the Earth once a day, while the stars remained fixed within what was termed the "celestial sphere," rotating around the Earth's axis.

In this geocentric framework, the profound questions that engaged philosophers and thinkers revolved around the following inquiries: What impels the celestial sphere to revolve around the Earth? How are the stars affixed and attached to this sphere? By what

mechanism do the Moon and the Sun maintain their precise orbits around the Earth?

All these inquiries constituted the major challenges that thinkers of that era grappled with, striving to unravel and furnish credible explanations. Nonetheless, with the discrediting of this model, spearheaded initially by Copernicus and subsequently championed by Galileo and Kepler, in conjunction with the advent of Newton's classical mechanics, these questions lost all significance.

For the heliocentric model, the postulation of a “celestial sphere” was simply illogical and absurd: what seemed to require an answer or a theory no longer required explanation.

Some philosophers and scientists argue that certain questions about consciousness may indeed be framed in a way that assumes a meaningful reality to concepts that might be fundamentally flawed or lack clear definitions.

The comparison to historical theories that later proved to be pseudoconcepts raises the possibility that our current questions about consciousness might need reevaluation in terms of their coherence and underlying assumptions.

This is a philosophical provocation that we will have to deal with during this odyssey of the mind. We cannot

deny that everything we do and feel is, in some way, part of the conscious mind.

We may then find ourselves wondering: How is it that we know so little about something that is inherent to our human nature? Could it be that contemporary science struggles to provide a satisfactory explanation of the mind? Is there a theory of consciousness robust enough to account for all its phenomena?

In the first part of the book, we will try to answer these questions through dialogues with four prominent thinkers on the nature of consciousness – David Chalmers, Susan Blackmore, Nicholas Humphrey and Sir Roger Penrose – with whom I had the honor of debating.

To enhance your comprehension of the forthcoming dialogues in each section's second part, the following chapters will offer a concise introduction to the topics explored in these discussions.

Let's start by understanding why consciousness possesses this enigmatic aura concerning its existence and the capability to offer plausible explanations about itself.

II. THE MYSTERIOUS CONSCIOUSNESS

The inaugural dialogue features the insights of the philosopher who brought the subject of consciousness into the forefront of significant philosophical and scientific discussions. We are referring to David Chalmers, Professor of Philosophy and Neural Science, as well as co-director of the Center for Mind, Brain, and Consciousness at New York University. He is one of the primary advocates of the notion that there is something particularly intricate about the essence of consciousness.

As you will observe in the dialogue, Chalmers introduced the renowned 'hard problem of consciousness,' though the problem is not entirely novel. Essentially, it represents an inherent aspect for those acknowledging the mind-body problem as a legitimate concern. This perspective is held by those who contend that the mind possesses properties that cannot be fully explained by the physical nature of the world.

Many mental phenomena such as learning, reasoning, memorizing, etc., can be explained in terms of playing a certain “functional role”: if we discover the “function” of a certain system, we will know everything else about that system.

However, for Chalmers, this is not the case regarding consciousness:

> "What makes the hard problem *hard* and almost unique is that it goes beyond problems about the performance of functions. To see this, note that even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience – perceptual discrimination, categorization, internal access, verbal report – there may still remain a further unanswered question: Why is the performance of these functions accompanied by experience?"[13]

If we concede the reality of this problem, one of the (seeming) consequences is that we are claiming that a scientific explanation of consciousness may truly escape us.

In this regard, another philosopher essential for comprehending the nature of the 'hard' problem of consciousness is the philosopher Thomas Nagel. Nagel perceives the problem as the 'subjectivity' of conscious mental states.[14]

This philosopher contends that the facts about conscious states are inherently subjective: they can only be fully understood from limited perspectives,

---

[13] Original publication: Chalmers, D. (1996) *The Conscious Mind*, New York: Oxford University Press, p. 202.

[14] Original publication: Nagel, T. (1974) "What Is It Like to Be a Bat?", *The Philosophical Review*, 83 (4): 435-450.

accessible solely to the subject undergoing that particular conscious experience,

For instance, in the given scenario of sipping a caipirinha on Copacabana beach, only the reader will have direct and subjective access to the taste experience of that caipirinha, and those external to the situation won't be able to precisely understand what the reader is feeling in that moment.

However, scientific explanation requires an objective characterization of the facts, which moves away from any particular point of view. Thus, the facts about consciousness elude science, rendering "the mind-body problem truly intractable."[15]

Chalmers ends up introducing an original and interesting argument termed "The Philosophical Zombie Argument" to demonstrate this characteristic of consciousness.

This argument revolves around the logical possibility of a world identical to ours, lacking secondary qualities (referred to as *qualia* in philosophy), and devoid of any states of consciousness – a zombie world. In this world, there would exist a zombie counterpart for each of us, identical at the molecular level. However, this zombie would lack any conscious experiences, even though,

[15] Original publication: Nagel, T. (1974) "What Is It Like to Be a Bat?", *The Philosophical Review*, 83 (4): 435.

from an external perspective, it would exhibit behavior as if it were conscious.

Consider this scenario: you, the reader, encounter your favorite celebrity on the street, someone you've admired since the first time you saw them in your favorite movie. Unbeknownst to you, this individual is a philosophical zombie.

With great enthusiasm and in a moment of extreme romanticism, you express your romantic feelings to this celebrity. To your surprise, the celebrity responds with equal enthusiasm, reciprocating the same feelings, expressing love, and behaving as if genuinely in love with you.

In this fictional scenario – acknowledging the improbability of a famous celebrity reciprocating such feelings – the individual interacting with us lacks genuine feelings or conscious experience. From the reader's perspective, this person merely appears to possess consciousness based on their behavior.

Following the outcome of this *hollywoodian* example, David Chalmers argues that we can say, indirectly, that zombies have a logical possibility – although they may not have a natural possibility – if we understand the functional organization of people.

For instance, we can conceive that a silicone isomorph of the reader— with chips in place of neurons — would not develop consciousness, implying that the facts of

functional organization do not necessarily imply the facts of consciousness.

At this point, it is important to highlight that if the notion of a "philosophical zombie" is conceptually coherent and logically possible, then, in that case, conscious and subjective states are not identical to the physical states of the brain.[16]

This takes us to another philosopher, who argues that the existence of consciousness in the physical world causes an “explanatory gap”. Joseph Levine argues that there is a special “explanatory gap” between consciousness and the physical.[17]

The challenge of closing this explanatory gap is what gives rise, in some sense, to the hard problem of consciousness. Levine argues that a good scientific explanation must deductively implicate what it explains, allowing us to infer the presence of the target phenomenon from a demonstration of laws or mechanisms and initial conditions.

Deductive implication is a kind of logical relationship with the following formal configuration: if the premises

---

[16] Historically, the thought experiment of the philosophical zombies has the purpose of refuting the physicalist theory of mind that argues that conscious states can be reduced – in the same sense as described in the previous section – to specific physical states, in relations of strict identity.

[17] Original publication: Levine, J. (1983) “Materialism and qualia: The explanatory gap”, *Pacific Philosophical Quarterly*, 64: 354-361.

of an argument are true, then it follows that its conclusion is necessarily true.

Note the following example: since we discovered that lightning is nothing more than an electrical discharge, knowing all the conditions suitable for an electrical discharge in the atmosphere at time *t* allows us to deduce that the lightning must have occurred at time *t*.

However, for both Levine and Chalmers, this does not seem to be the case when we think about consciousness: no matter how detailed our specification of brain mechanisms or physical laws is – there is always an open question as to whether consciousness is present and accompanies a certain underlying psychological mechanism or not.

Of course, many physicalist thinkers – who believe that everything that exists in the world has to be explained by the laws of physics and chemistry – reject this argument. They argue that a world in which we knew everything about mechanisms would necessarily contain an explanation about consciousness.

According to this perspective, consciousness would necessarily be generated from any set of physical circumstances that we are unaware of at the moment, but that nothing prevents us from being able to come to know in the future.

To help the reader to better understand this "mystery of consciousness," we can seek valuable assistance from the imaginary Dr. Mary, the most brilliant neuroscientist the world has ever known (unfortunately, we do not have this kind of knowledge in the present; this is just another thought experiment).

Mary is, then, a genius neuroscientist who knows everything there is to know about the physical properties of the brain, especially those linked to the visual perception of color. Furthermore, Mary is so brilliant that she also knows all the physical facts about light and colors.

Therefore, we can declare that Mary knows everything there is to know about the way human beings experience colors visually, encompassing both the brain processes involved and the physical processes.

However, there is a small problem: Mary has been a prisoner since birth in a house where she only has access to her room. Worse than that, this room doesn't have any light; all the knowledge that Mary learned about the brain and mind took place in a monochromatic environment, using black and white materials and tools.

What does this scenario seek to show about consciousness? This thought experiment aims to highlight the following situation: imagine that Mary is finally freed from that monochromatic room, and her

boyfriend, with whom she exchanged love letters for years, finally meets her in person.

The boyfriend, in a romantic gesture, presents a red rose to Mary: for the first time, Mary is faced with an object of color that she has never seen before, despite possessing all the theoretical knowledge about the brain and colors, including red roses.

Would Mary already know what to expect upon seeing the red color of the flower for the first time? Or would she learn something new in this visual experience, something that, after all, had eluded her despite possessing all the knowledge of how the brain and colors work?

This thought experiment, originally developed by the philosopher Frank Jackson,[18] aims to illustrate two particular theses for the reader. Firstly, that subjective experiences linked to qualia have a real existence, not being a mere theoretical fable of some philosophers and scientists.

Secondly, the thesis that the conscious mind can be described only by appealing to physical and chemical properties is mistaken, given that, in Mary's scenario, there are truths about colors that appear to be non-physical.

[18] Original publication: Jackson, F. (1982) “Epiphenomenal Qualia”, *Philosophical Quarterly*, 32 (127): 127–136.

Although this argument has some intuitive force, some philosophers[19] believe that we cannot conclusively derive these two theses from Mary's experience. At most, the counterintuitive conclusion is that Mary would already know what to expect to experience when her romantic boyfriend offered her the red rose.

Of course, the acceptance of these two theses will depend, in essence, on the philosophical position in which the reader finds themselves in relation to consciousness: if you take Mary's experience seriously, you likely believe that there is something fundamental in consciousness that cannot be adequately described by current science.

However, if you believe that Mary already knows what to expect when she sees the red rose offered by her boyfriend, then you likely hold the view that there is nothing especially mysterious about consciousness.

A relevant area where we can explore the mystery of consciousness is connected to new technological developments and the possibility that we might be living in a simulated world, akin to the movie *The Matrix*.

The possibility that we are living in a simulation was analyzed by philosopher Nick Bostrom in his popular

[19] For example: Churchland, P. M. (1985) "Reduction, Qualia, and the Direct Introspection of Brain States", *The Journal of Philosophy*, 82 (1): 8-28.

article “Are you living in a computational simulation?”.[20] The simulation argument outlines three potential scenarios for the future of the world:

> (i) It is very likely that the human species will become extinct before reaching a “post-human” stage;
>
> (ii) It is extremely unlikely that a future posthuman civilization could run a significant number of simulations of its evolutionary history (or variations thereof);
>
> (iii) We are almost certainly living in a computer simulation.

Bostrom argues that, among these three propositions, at least one must be true. Suppose (i) is incorrect: some civilization in the universe reaches technological maturity.

Now, let's consider that (ii) is also incorrect: no civilization uses its resources to carry out simulations. If so, we can (supposedly) conclude the following: since these ancient civilizations have the capacity to run an enormous number of simulations, if the first two possibilities are false, there may be a greater number of simulated entities than non-simulated entities in this universe.

---

[20] Original publication: Bostrom, N. (2023) “Are you living in a computer simulation?”, *Philosophical Quarterly*, 53 (211): 243-255.

In this scenario, all human beings would be – at least the majority – living inside simulations and not outside them. The reasoning presupposed in the argument is that, if we reject the first two hypotheses, we will have to accept that the third necessarily follows, given that for each real world, we would have millions of simulated worlds.

Therefore, the probability that we are currently in a simulation seems to be genuine, if we accept the premises advanced by Bostrom, something that Chalmers seems to accept, supporting the following argument:

> (iv) It is more likely that conscious human simulations are possible;
>
> (v) It is more likely that, if conscious human simulations are possible, many human-like populations will create them;
>
> (vi) There is a good chance (25% or more) that we are computer simulations.[21]

But does the simulation hypothesis make sense? If we remember the epic movie *The Matrix*, there are several scenes in which Neo, the main character, finds himself in a situation of total uncertainty about whether he is in the real world or in the simulated world.

[21] Original publication: Chalmers, D. (2022) *Reality+: Virtual Worlds and the Problems of Philosophy*, New York: W. W. Norton.

In fact, if we are in a simulation, it will be very difficult, from the point of view of our subjectivity, to understand that that would be the case. However, some intellectuals argue that this probabilistic argument fails because some of its main premises are implausible.

Let's see: it may very well happen that, someday, computers will have consciousness, but it is unlikely that their consciousness will be the same as ours, given that the physical mechanisms of computers are very different from the neural mechanisms – of flesh and blood – that produce human consciousness.

The assumption that artificial consciousness will be the same as human consciousness presupposes substrate independence: mental states can operate in a wide range of physical systems regardless of their material makeup.[22]

However, we can criticize this assumption and argue that consciousness actually depends on a neurobiological substrate like the one that evolution "developed" through our bodies and brains (as you will see in several dialogues in this book).

[22] Section based on: Gouveia, S. & Neiva, D. (2017) "The Problem of Consciousness on the Mind Uploading Hypothesis" In *Philosophy of Mind: Contemporary Perspectives*, Cambridge: Cambridge Scholars Publishing.

And if that is the case, no artificial system will be able to actually reproduce consciousness, at least understood as subjective and human consciousness.

Furthermore, we can also question the acceptance that the second premise (ii) is false: in fact, it is not so obvious why a future human civilization would consider spending countless resources to simulate thousands of universes instead of applying them into other existential priorities.

Additionally, that same civilization may consider such simulations to be ethically immoral: if simulating the universe implies simulating human beings, and if the substrate independence thesis is correct – something we do not currently know to be the case – then simulating these human beings will be also simulate their consciousness and, consequently, their ability to feel pain and suffer.

If we can simulate this sentient capacity, then we could be committing the biggest ethical mistake in the entire history of humanity (and we know that the list of these mistakes is already quite long!): we could be simulating an infinite number of human suffering that we ought to have a moral obligation to avoid at all costs.

Whether or not you think the argument for or against the thesis that we are living in a simulation is correct is up to you. But remember to make the wisest choice when faced with the decision to choose between the

red pill – and the real world – or the blue pill – and the simulated world.

Speaking of taking pills, next, let's explore, in the company of psychologist and Professor Susan Blackmore, alternative ways of thinking and investigating the phenomenon of consciousness through various altered states of consciousness and what they can teach us about the nature of the conscious mind.

III. THE ALTERED CONSCIOUSNESS

The second dialogue features the participation of Susan Blackmore, writer, lecturer, and visiting professor at the University of Plymouth, in the United Kingdom. Having a PhD in Psychology, she seeks to study the nature of consciousness through more out-of-the-box methodologies.

Typically, when discussing 'consciousness,' we tend to approach it from a normalized perspective. In other words, we consider this mental phenomenon in its 'normalized' state – when we are awake, experiencing the sensations of the world – rather than during dreaming.

Throughout the recorded history of human activities, it has been recognized that there is a category of conscious states distinct from the normal states we are used to. These states can be induced by the consumption of psychoactive substances, meditation, or even, as we will explore later, lucid dreams.

Why focus on these states that differ from our usual experiences? Because they can be instrumental in identifying and comparing diverse information when consciousness is in an altered state, what we refer to as an 'Altered State of Consciousness' (ASC).

How can we attempt to define these states? One objective method is to specify how that state was induced. For example, a change in consciousness caused by the consumption of Ayahuasca leads to a different state of consciousness than a change induced by hypnosis or meditation.

Although this is an interesting proposal, it falls short because a state induced in the reader by Ayahuasca, for example, will have different effects on their consciousness than if I, the author of this book, consumed the same drug. There is no causal and necessary relationship between a substance and a specific ASC.

Another approach to defining ASCs is grounded in physiological measurements observed when under the influence of substances, such as body temperature, heart rate, cortical oxygen consumption, among others. However, the challenge here lies in the variability of these physiological criteria, which do not remain consistent among all individuals, both in terms of subjective experience and in the methods through which they are induced.

We know that minor changes in physiology are associated with significant fluctuations in subjective state, and vice versa, so no direct mapping seems possible. Therefore, defining these highly volatile states, characterized by an immense subjective experiential impact, requires a careful approach when

grounded in objective elements such as human physiology.

Other alternative is to look at different definitions provided in psychology, neuroscience or philosophy manuals, such as the following:

- "(...) a qualitative alteration in the overall patterns of mental functioning so that the experiencer feels that his/her operations of consciousness are radically different from ordinary functioning";[23]
- "(...) a temporary change in the overall pattern of subjective experience, such that the individual believes that his or her mental functioning is distinctly different from certain general norms for his or her normal waking state of consciousness";[24]
- "(...) exists whenever there is a change from an ordinary pattern of mental functioning to a state that seems different to the person experiencing the change".[25]

While these definitions are interesting, they all fall short in one crucial aspect: they attempt to define ACSs by

[23] Original publication: Tart, C. (1972) "States of consciousness and state-specific sciences", *Science*, 176: 1203-1210.

[24] Original publication: Farthing, G. (1992), *The Psychology of Consciousness*, Englewood Cliffs, NJ: Prentice Hall.

[25] Original publication: Nolen-Hoeksema, et al. (2014) *Atkinson & Hilgard's Introduction to Psychology*, Andover: Cengage Learning.

negatively comparing them to "normal" states of consciousness. The issue here is that we also lack a useful and concrete definition of normalized states of consciousness, which renders these kinds of definitions unhelpful in specifying concretely what could be considered the common nature of an altered states of consciousness.

Another option is to explore concrete examples that can help clarify this conceptual difficulty. Let's begin, therefore, with those states that the more open-minded reader may have already experienced: drug-induced states.

Psychoactive drugs are substances that affect mental functioning or consciousness. They operate by modifying the action of endogenous neurotransmitters or neuromodulators.

For instance, these drugs can enhance the effect of a neurotransmitter by stimulating its release or impeding its reuptake, prolonging its effects. Conversely, they may diminish the effects by inhibiting or blocking its reception at the postsynaptic membrane.

One reason the impact of mind-altering drugs can be extensive is that a single neurotransmitter can be active in numerous regions of the brain and even throughout the body.

Amphetamine is the best-known and most studied stimulant and has three main effects on the brain:

- induce the release of serotonin;
- induce the release of dopamine;
- inhibit serotonin reuptake.

Serotonin plays a crucial role in regulating mood and sleep, while dopamine helps mediate reward-motivated behavior and interpretive responses to self, others, and the environment.

The consumption of amphetamines includes consequences such as increased energy, higher tactile and other sensations, and an increase – both in frequency and intensity – of feelings of love and empathy for others.[26]

The experience and its effects often depend on the environment in which the drug is consumed. Frequent use can lead to tolerance and, consequently, long-term addiction.

Nevertheless, in a therapeutic and low-frequency context, it can be employed to treat conditions such as post-traumatic stress disorder and social anxiety with some effectiveness, when compared to more conventional treatments.[27]

---

[26] Original publication: Holland, J. (ed.) (2001) *Ecstasy: The complete guide: A comprehensive look at the risks and benefits of MDMA*, Rochester, VT: Park Street Press.

[27] Original publication: Danforth, A., et al. (2016) *MDMA-assisted Therapy: a new treatment model for social anxiety in autistic adults*,

The most common psychedelic drug is cannabis, which contains approximately 85 cannabinoid components. Describing the subjective effects of cannabis is not an easy task, partly because they vary greatly from person to person.

However, scientific research has unveiled some common effects, such as emotional responses, including euphoria and relaxation at lower doses, and fear and paranoia at higher doses.

Sensory effects include greater depth perception, a heightened sense of senses, increased sexual responsiveness and pleasure, a perception of time "slowing down" and space "expanding," and a higher focus on the present.

Some individuals experience a sense of the sacred or divine when taking this drug. It is also known that in some people, consumption can lead to increased creativity, while in others, it can result in slow thinking with negative effects on short-term memory.

At high doses, some individuals report phenomena of synesthesia, where sensory elements are exchanged, such as seeing a color when hearing a sound.

Ayahuasca is another type of psychedelic that is becoming increasingly common and popular.

Progress in Neuro-Psychopharmacology & Biological Psychiatry, 64: 237–249.

Considered a healing substance, it initially induces episodes of vomiting, and after a few minutes, a bewildering array of bodily sensations, transformations, visions, and perceptions emerge.[28]

Contemplation of death is common, as are mystical insights into personal matters and profound existential questions. During the consumption of this substance, the brain undergoes significant changes, such as a decrease in connectivity between the parahippocampus and the retrosplenial cortex.

This variation has been correlated with reports of the dissolution of the "self," explaining why those who undergo an ayahuasca experience often report feeling that their sense of self had become diluted or merged with the surrounding environment.[29]

Finally, one of the most potent hallucinogens is LSD (lysergic acid diethylamide). The effects of this substance encompass a broad spectrum of experiences, including both positive sensations such as joy and euphoria, and negative sensations such as terror, despair, and the disintegration of the "self."

---

[28] Original publication: Luna, L. e White, S. (2016) *Ayahuasca Reader: Encounters with the Amazon's Sacred Vine*, Santa Fe: Synergetic Press.

[29] Original publication: Uthaug, M. et al. (2018) "Sub-acute and long-term effects of ayahuasca on affect and cognitive thinking style and their association with ego dissolution", *Psychopharmacology*, 235: 2979-2989.

In the first placebo-controlled brain imaging study, participants were given 75 micrograms of LSD intravenously: an increase in functional connectivity was detected throughout the brain, while local effects coincided with changes in experience.

For example, visual hallucinations have been positively correlated with increased cerebral blood flow and functional connectivity in the primary visual cortex (V1).

All of these substances bring something new to those who consume them: a different and abnormal conscious experience that can contribute to understanding the nature of consciousness. Why so?

Because we can compare, for example, the normal brain of a subject without drug induction with a drug-induced brain, allowing us to gain relevant knowledge about neuronal activity, the dynamism of neuronal networks, or the relationships of certain regions of the brain.

Let's now shift our focus to another way of thinking about Altered States of Consciousness, this time through meditation. Can meditation be considered an ASC? Some definitions seem to imply that:

- "(...) meditation can be regarded as a slow, cumulative, long-term procedure for producing an altered state of consciousness".[30]

Some practitioners of Buddhism may have kenshō (awakening) experiences, including glimpses into the supposed nature of the mind. We also know that anyone who meditates feels that their mind has been radically altered, and, in this sense, it can be considered an ASC.

However, some scientists argue that meditation is nothing more than a particular way of... sleeping. The neuropsychiatrist Peter Fenwick[31] showed that the EEG profiles in meditation are not exactly similar to those of sleep or drowsiness, but many practitioners actually go into a "microsleep" state during meditation. In another study, practitioners slept about a third of the time during meditation.[32]

The beneficial effects of meditation, such as reducing anxiety and depression and improving cognitive performance, can be partially explained by the similarity with the known benefits of microsleeps. Yet,

[30] Original publication: Wallace, B. e Fisher, L. (1991) *Consciousness and Behavior*, Boston, MA: Allyn and Bacon.
[31] Original publication: Fenwick, P. (1987) "Meditation and the EEG" In M. West (ed.) *The Psychology of Meditation*, Oxford: Clarendon Press.
[32] Original publication: Austin, J. (1998) *Zen and the brain: Toward an Understanding of Meditation and Consciousness*, Cambridge, MA: MIT Press.

many practitioners report being able to distinguish between deep meditation and sleep states, although this distinction is difficult to explain objectively.

Speaking of sleep, another way to consider Altered States of Consciousness is through lucid dreams. A lucid dream is characterized by the awareness that you are dreaming at that precise moment.

Every day when we sleep, we go through a cycle of three states: wakefulness, REM (rapid eye movement), and non-REM sleep. A typical night's sleep consists of four or five cycles between non-REM and REM sleep, often including some non-conscious micro-awakenings.

The “awake” and sleeping states are characterized by behavioral indicators, such as the speed of awakening, eye movements, muscle tension and brain activity, which can be measured using methods like electroencephalogram (EEG), among others.

We know that in REM sleep, the brain is highly active and its EEG resembles the brain in a waking state. In non-REM sleep, the overall firing rate of neurons is as high as in waking states, but the pattern is quite different, with the EEG dominated by long, slow waves rather than complex, fast ones.

During sleep, the brain isolates itself in several ways: there is an inhibition of sensory input at the thalamocortical level during non-REM sleep, whereas in the REM phase, this inhibition is more peripheral. EEG

and fMRI studies show that during auditory stimuli in the REM phase, the auditory cortex remains active, and in the intermittent REM phase, which includes eye movements and muscle spasms, the brain operates in a closed circuit and is functionally isolated from the outside world.[33]

In REM sleep, the brain stem blocks motor commands in the spinal cord, which prevents the translation of mental activities into physical movements. However, if this mechanism fails, this could explain why that reader's uncle usually gets up in the middle of the night to eat all the desserts from the fridge: sleepwalking can be attributed to a failure in this mechanism.

In this REM phase, the amygdala, hippocampus and anterior cingulate exhibit increased activity, as do parts of the visual system and visual association areas. The dorsolateral prefrontal cortex, associated with executive functions such as working memory, problem solving and motor organization, exhibits reduced activity compared to the waking state.

Research on lucid dreams provides an interesting context for exploring the neural correlates of consciousness, as it allows us to correlate physiological, neurochemical and behavioral variables with subjective

[33] Original publication: Wehrle, R. et al. (2007) "Functional microstates within human REM sleep: First evidence from fMRI of a thalamocortical network specific for phasic REM periods", *European Journal of Neuroscience*, 25: 863-871.

descriptions of dreams, offering a greater understanding of the dimensions of conscious experience.

Lastly, a final way to think about the nature of consciousness is through something very special that can happen to our mind at a given moment, as happened to Professor Susan Blackmore who, at the age of 19, found herself experiencing an "Out of Body Experience" (OBE) (cf. dialogue with Susan Blackmore).

An OBE is an experience in which a person seems to perceive the world from a location outside their physical body. During an OBE, individuals can observe themselves from a perspective external to their body.

Those who have OBEs often report more psychic experiences and a higher belief in the paranormal compared to those who do not. Additionally, individuals with OBEs may have better dream recall and a greater frequency of lucid dreams.[34]

Some individuals interpret these unusual experiences as absolute proof that consciousness is an immaterial soul, independent of the body or brain. While it is understandable that someone with limited scientific knowledge might lean towards this belief, there are alternative explanations to consider.

[34] Original publication: Blackmore, S. (2017) *Seeing Myself: The New Science of Out-of-body Experiences*, London: Robinson.

Let's see: if the soul could actually see the physical world during the OBE, that would imply an interaction with that physical world, contradicting its supposed non-physical nature, and, therefore, it must be a detectable physical entity. However, this contradicts the supposed non-physical nature of the soul.

On the other hand, if the soul is non-physical, then it cannot interact with the physical world to be able to observe the (physical) body from an outside point of view.

Indeed, an alternative perspective to dualistic theories suggests that, despite the subjective experience of leaving the body during an OBE, nothing actually departs the physical body.

For instance, psychoanalytic theories have proposed that OBEs might reflect fear of death, ego regression, or the re-experiencing of birth trauma. However, these psychoanalytic theories are challenging to test scientifically and have had a limited impact on advancing our understanding of the OBE phenomenon.

However, if we look at the most current neuroscience, we learn that the temporal lobe plays a significant role on OBEs since, in epileptic patients, stimulation of the temporal lobe causes episodes of this kind, as well as psychic and mystical experiences.

Neuroscientist Michael Persinger proposed that mystical beliefs and experiences were “creations” of the

function of the temporal lobe of our brain, having managed to induce several OBEs and bodily distortions when using Transcranial Magnetic Stimulation which, focusing on the right lobe, produced authentic OBEs to patients.

The specific area of the brain involved in stimulation is the right temporoparietal junction, which makes sense to have an impact on OBEs, given that this area is responsible for processing visual, tactile, proprioceptive and vestibular information. These inputs, when combined, form the body schema for each one of us.

In this way, we can provide a scientific explanation for a phenomenon that may seem mystical but has, at its core, a perfectly human explanation.

Next, we will introduce other ideas from psychology informed by neuroscience – this time, from a perspective influenced by evolution. The aim is to explain why we, as human beings and some animals, have conscious subjective experiences in the first place.

IV. EVOLUTIONARY CONSCIOUSNESS

The third dialogue features a contribution from the renowned psychologist and theorist Nicholas Humphrey, Professor Emeritus of Psychology at the London School of Economics. In this dialogue, you will encounter fascinating ideas about the nature of consciousness from an evolutionary perspective.

Nicholas Humphrey's journey in the study of consciousness began, as you will see in the dialogue, when he was just 23 years old and a PhD student in a psychology laboratory at the University of Cambridge.

During his research, he initially investigated a monkey that, while anesthetized, had an electrode inserted into its brain, specifically in the superior colliculus, an “older” area in the neuroanatomical evolution of the brain responsible for visual processing.

The superior colliculus precedes the more developed visual cortex that enables conscious visual perception in humans and mammals. The intriguing aspect of this episode is that, even though the monkey was not awake, the nerve cells in the superior colliculus were active, suggesting that, perhaps, visual processing was occurring without an associated conscious sensation.

Sometime later, Humphrey encountered a monkey named Helen, whose visual cortex had been completely removed, leaving only the superior colliculus. Through several interactions over many months, Helen, who should have been entirely blind, seemingly developed the ability to see again. She could pick up her favorite fruit, recognize the Professor, and perform other remarkable feats that should not have been possible for a mammal without vision.

Now, the author of *Soul Dust* began to suspect that Helen could be having visual perceptions without having associated conscious experiences: she could process information from her environment without forming a conscious image of that information.

This research led his doctoral supervisor, Professor Larry Weiskrantz, to observe a human patient whose visual cortex was reduced by half due to an accident. This patient could identify objects with high precision that would be present in his visual field, even though the patient admitted to having seen nothing himself.

The famous phenomenon of 'blindsight' was discovered, which helped us understand a lot about the functioning of vision, but it also raised fundamental questions about the nature of consciousness. This entire episode sparked enormous encouragement and curiosity in Professor Humphrey, motivating him to contribute to clarifying the existence of conscious experiences in the world.

Humphrey's main contribution is to advance a theory of consciousness that can be explained through Darwinian evolutionary theory. In the author's words:

> "(...) I will argue that the truth about consciousness – if and when we see it from the right perspective – is that it is indeed the product of a highly improbable bit of biological engineering: a wonderful artwork of nature that gives rise to all sorts of mysterious impressions in our minds, yet something that has a relatively straightforward physical explanation."[35]

And thus begins this wonderful story of how consciousness could have evolved through natural selection.

The first step is to argue that consciousness has an impact on our behavior. And why the behavior? Given that consciousness, as we understand it, is a feature inherent in life on Earth, we can assume that – like any other specialized feature of living organisms – it evolved because it confers some selective advantage.

In one way or another, consciousness must aid the organism in surviving and reproducing. This can only occur if, in some manner, it is changing the way in which the organism interacts with the outside world, a process that typically involves behavior.

---

[35] Original publication: Humphrey, N. (2011) *Soul Dust: the Magic of Consciousness*, Princeton: Princeton University Press.

But how does consciousness distinguish an organism in terms of behavior? When examining "conscious creatures", as Humphrey labels them, we realize that they lack distinctive physical characteristics compared to non-conscious creatures. Consciousness does not provide greater health, strength, or beauty; rather, it appears to exert its effects on the survival of the organism through what we can term the "psychology" of the creature.

Therefore, being phenomenally conscious must be related to the way an organism thinks, desires or believes and the fact of being conscious leads it to act in an adaptive way to the world through specific behaviors that confer an advantage and that can be identified.

However, despite this adaptive “advantage”, it is still not clear why consciousness was selected as a relevant element by natural selection. This is the central point of the evolutionary theory of consciousness: if natural selection can “see” the effects – whatever they may be – of psychological change in behavior, presumably other “external” observers can also recognize the adaptive advantages of being conscious.

We are, therefore, on the right path to creating a plausible story about the evolution of consciousness. In addition to presenting an inherent advantage to each organism, this intrinsic advantage influences external

behavior. This, in turn, makes it possible to detect and identify the adaptive advantage.

But how did consciousness emerge evolutionarily to become a fundamental characteristic of human beings today? Humphrey believes that the emergence of consciousness occurred quickly and was an "all or nothing" phenomenon. It appeared later in evolutionary history when our hominid ancestors developed diverse social skills such as imitation, deception, and language.

Thus, we can argue that consciousness is an emergent property that evolved for its social function, for the ability to understand, predict and manipulate the behavior of other individuals. Evolution "favored" individuals possessing these capabilities over those who did not, leading to the development and persistence of consciousness.

Like the species of great apes living today, humans have always lived in complex social groups. Knowing the intentions of other individuals can be extremely useful in determining who holds higher positions in the social hierarchy, whom we can trust, and with whom we can form alliances, among other social considerations.

Thus, according to Humphrey, ancestors who could understand, predict, and manipulate the behavior of others had a clear adaptive advantage. At this point, one could argue that humans might have acquired

these skills by simply observing the behavior of others and its consequences from an external standpoint, somewhat akin to behaviorists.

But the Professor Emeritus of the University of Cambridge believes there might be a better way to achieve this result: through the hypothesis that individuals acquired the ability to introspect, that is, to put themselves in another person's shoes, and attempt to "observe" their own minds.

Humphrey compares this capacity to an "inner eye," not directed at the outside world like most sensory organs, but at the individual's inner world. This "inner eye" cannot observe the brain's neuronal functioning but instead perceives a more accessible psychological version of that activity – subjective conscious states.

Could this mean that consciousness can be considered an "invention", in the sense that it did not exist at a certain period of human development, and that it came into existence later?

For Humphrey, the answer is twofold: consciousness is a cognitive faculty developed by natural selection, designed to help us make sense of ourselves and our surroundings. Simultaneously, it is a fantasy created by the brain, designed to alter the value we attribute to our existence.

However, if this is the case, we run the risk of committing the error of the "Cartesian theater," as

pointed out by Daniel Dennett. This idea suggests that there is a specific place in the brain where the "movie" of the outside world takes place.

This approach to consciousness does not necessarily imply an illusionist thesis of consciousness – the idea that consciousness is an illusion created by the brain. This is because what happens inside the "theater" is not a replica of the outside world but rather of the "inner" world, representing the realm of subjectivity and *qualia*.

It is true that the author of *Sentience* (2022) starts by denying the "realism" of *qualia* as defined by philosophers such as David Chalmers or Thomas Nagel. As we saw previously, these philosophers consider *qualia* to exist in a fundamental and independent way: denying this may lead to an approach that leans towards an illusionist theory in that specific aspect.

This approach argues that even though conscious subjective experiences seem to possess these wonderful non-physical properties, it can only be the case because the brain is playing a trick on us.

This is possible because the brain is a computational engine that deals with symbols, and physically based symbols can perfectly represent states of things that they do not represent, and even things that may not exist.

As Daniel Dennett, considered one of the foremost advocates of the illusionist theory, emphasizes:

“Consciousness is an illusion of the brain, for the brain, by the brain”.[36]

So, when we have a conscious subjective experience, our own brain is "putting on" a magic show that makes us believe we are experiencing something outside of ourselves. In a way, this approach undermines the mystery of the human experience, the notion that we are unique, magical, and exceptional in the universe.[37]

But for the author of *Soul Dust* (2012), this can also be an advantage: as soon as we realize that there is nothing especially mysterious about the existence of consciousness, we can use this information as an asset to "enrich" ourselves as human beings. This enrichment can occur through the arts, through science, and by transcending a merely "earthly" existence.

To conclude this introductory note, let's take a look at the last words of the chapter 23 of Humphrey’s new book titled *Sentience*, focused on the exceptionality of consciousness in the world:

> “Even if the idea of a naturally evolved feature being ‘intended’ must be wrong, I imagine that Darwin himself could have seen phenomenal consciousness as an ‘ultimate’ achievement—the crowning glory of the evolutionary process that

[36] Original publication: Dennett, D. (1991) *Consciousness Explained*, New York: Little Brown.

[37] Original publication: Humphrey, N. (2020) “The Invention of Consciousness”, *Topoi*, 39: 13-21.

> began with the Big Bang. It's an invention so sublime that, if it were to cease to exist, it would indeed diminish the whole of creation."

This evolutionary view of consciousness offers an interesting approach that can be studied empirically. However, it does not specifically clarify how our brain produces the emergent property of conscious subjectivity.[38]

To close this first part of the book dedicated to consciousness, I will introduce the ideas of the last guest with whom I debated this topic, Professor Sir Roger Penrose.

[38] Section based on: Humphrey, N. (2023) *Sentience: the Invention of Consciousness*, Cambridge, MA: MIT Press.

V. QUANTUM CONSCIOUSNESS

The fourth and final dialogue is, perhaps, the one that the reader may have more difficulty following. To prepare you for one of the most stimulating conversations in this book, I will try to introduce some of the central concepts developed by the Nobel Laureate in Physics, Sir Roger Penrose, Emeritus Professor at the University of Oxford.

How can we think about consciousness? Some scientists and philosophers consider consciousness as an emergent property of computation among neurons that interconnect and switch at chemically mediated synapses. The problem of consciousness, in this context, is the challenge of understanding the specific computations that occur and whether we can replicate that kind of procedure.

From this perspective, the brain is seen as a kind of digital computer that processes information from the environment to the mind, and consciousness is thought to play some kind of crucial role in this process.

However, the notion of "computation" in this classical sense is quite problematic; for the British physicist, the brain operates on a completely different scale of processing than a current digital computer.

Thus, Penrose's approach presupposes expanding the focus of explanatory power from classical physics/traditional computational neuroscience (i.e., the neuron) to a "smaller" level.

Through certain elements of quantum physics, a theory of consciousness with a quantum explanation is proposed. To achieve this step, Penrose criticizes the idea that quantum properties cease to be relevant when we "climb" the complexity scale of biology.

This argument suggests that the neuron, neuronal interactions, the brain, etc., are not suitable subjects to be interpreted by quantum mechanics (QM) but rather by classical physics.

Opposing this thesis, a group of scientists[39] showed that QM has a fundamental role in photosynthesis, discovering that in chlorophyll cells, light transfer is close to 100 percent efficiency, which far surpasses any type of modern human technology.

Furthermore, it also suggests that there must be some kind of QM “safe space” to maintain system coherence in order to achieve this type of efficiency.

Why do we need to go beyond digital computing to reach a viable theory of consciousness? Penrose argues

[39] Original publication: Engel, G. et al. (2007) “Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems”, *Nature*, 446 (2007): 782-786.

that classical physics has three assumptions that preclude an explanation of human consciousness:

- Causal determinism: we can know all the initial conditions of physical systems and, therefore, we can predict their future behavior based on a causal chain;
- Locality/independence: two systems separated in space cannot interact instantaneously;
- Objectivist realism: objects in external reality exist with well-defined properties that are independent of any observers.

However, the Nobel Prize in Physics argues that consciousness is non-computational (i.e. non-algorithmic) as a direct consequence of Gödel's theorem.[40]

Penrose demonstrated that the mental quality of 'understanding' cannot be encapsulated by any computational system and must stem from some 'non-computable' effect. He suggests that the non-computable ingredient necessary for human consciousness and understanding has to lie in an area where our current physical theories are fundamentally incomplete but of significant relevance to the scales pertinent to the functioning of our brains.

[40] Original publication: Gödel, K. (1931) "Uber formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", *Monatshefte für Mathematik und Physik*, 37: 349–360.

The only “serious” possibility is the Incompleteness of Quantum Theory, an incompleteness that both Albert Einstein and Erwin Schrödinger had already recognized, referred to as the Measurement Problem. One way to solve this problem would be to provide an extension of the standard QM structure by introducing an objective form of quantum state reduction – called “objective reduction” (OR).

Does this mean that consciousness reduces quantum states? In this version, which we call “the classical interpretation” of Quantum Mechanics, it is the act of observation by a subject that defines the quantum state and violates the principle of superposition: Schrödinger's cat would be dead or alive until someone (i.e., an observer) observes the system, defining the superposition at a particular position. [41]

To understand superposition from the classical perspective, imagine tossing a coin to decide what you will have for dinner: if it lands on tails, it will be breaded seitan; if it lands on heads, it will be Swiss raclette. Right after tossing the coin in the air, during the few seconds that it is airborne, the coin will be in a state of superposition: it is simultaneously in both positions, heads and tails.

[41] Original publication: Schrödinger, E. (1935) “Die gegenwarten situation in der quantenmechanik“, *Naturwissenschaften*, 23: 807-812, 823–8, 844–9.

The side of the coin is only defined when the reader picks up the coin and observes the side (let it be heads!). Subatomic particles are like this coin: they exist in superposition states until an observation causes them to collapse into a certain state.

However, Penrose's view is precisely the opposite of this description: for him, there is a conflict between quantum mechanics and Einstein's theory of general relativity that demands something 'new' to happen. [42]

In the view of the 2020 Nobel Prize in Physics, the reduction of the state occurs not by observation but spontaneously, and it is in this process that consciousness appears. But does this make sense? Note: the QM level only functions as a theory on a very "small" (sub-atomic) level, and the level of neuronal information processing occurs on a much larger scale.

If quantum physics extends to "higher" levels, it finds itself in a "hot," "wet," "noisy" context due to contact with the environment, necessarily leading to the collapse of quantum states and their quantum decoherence.

This decoherence can be explained with the following example: imagine that the reader's cat is a quantum cat that is simultaneously alive and dead according to the

[42] Original publication: Einstein, A. (1916) “Die Grundlage der allgemeinen Relativitätstheorie”, *Annalen der Physik*, 49 (7): 769-822.

principle of superposition. If your cat interacts with particles in the environment (e.g., with photons), decoherence will occur. Since the environment "observes" it indirectly, this observation would cause your cat to assume a defined state, alive or dead, rather than existing in both states at the same time (hopefully, of course, it is the first quantum option).

Therefore, following this line of reasoning, it seems implausible for the principles of quantum physics to operate at the level of information exchange between neurons in the brain and produce consciousness. Does this pose an obstacle to Penrose's theory of consciousness?

According to the Oxford physicist, this criticism doesn't hold due to intriguing evolutionary reasons: over millions of years, biology has evolved to address the challenge of quantum decoherence. This adaptation is achieved through the development of a sub-atomic structure within neurons called "microtubules." Their unique structure ensures quantum coherence, similar to the natural process observed in photosynthesis, as mentioned earlier.

The reference to microtubules is not original to Penrose but comes from his colleague, the American anesthesiologist Stuart Hameroff. Hameroff proposed to Penrose that quantum coherence in the brain could occur within these microtubules. Microtubules are protein structures with a tubular formation inside

eukaryotic cells (part of the cytoskeleton). They play various roles, such as determining cell shape, coordinating movement, and overseeing cell division.

Hameroff suggests that microtubules are the quantum device that Penrose was looking for in his theory, as they help control the strength of synaptic connections, and their tubular shape may protect them from the surrounding noise of the larger neuron. Imagine that neurons are small factories: microtubules would be the tracks that help guide and organize movement within these factories.

However, it takes more than just a continuous array of random moments of quantum coherence to have any impact on consciousness: it is these moments of conscious awareness that, orchestrated by the microtubules in our brains, have the capacity to store and process information and memory.

For Hameroff and Penrose, microtubules can adequately preserve quantum coherence until they reach the neuronal level: for consciousness, it is necessary that many microtubules in several different neurons act in an 'orchestrated' way.

What, then, is the difference between normal quantum states and quantum states that lead to consciousness? For this pair of scientists, the key lies in global coherence, hence advocating objective reduction (OR),

wherein quantum states collapse into one option or another.

The "orchestrated" part is proposed to ensure that collaborative efforts among multiple microtubules are necessary to influence the neuronal level.

To summarize the ideas so far, we have the following theoretical-conceptual configuration:

- A model is proposed in which consciousness emerges through quantum effects occurring within sub-cellular structures internal to neurons known as microtubules;
- this model posits so-called “objective collapses” that involve the quantum system transitioning from a superposition of several possible states to a single defined state, but without the intervention of an observer or measurement, as in most quantum mechanics models;
- according to Penrose and Hameroff, the internal environment of microtubules is particularly well-suited to such objective collapses. The resulting self-collapses produce a coherent flow that regulates neuronal activity and enables non-algorithmic mental processes.

Following all of this, it is important to now introduce the so-called “Penrose Interpretation” of quantum

mechanics, as this is fundamental to understanding his theory of consciousness.

Penrose's interpretation is a speculation on the relationship between quantum mechanics (QM) and Einstein's general relativity. It proposes that a quantum state remains in superposition until the difference in curvature of space-time reaches a significant level for it to collapse, a phenomenon referred to as "self-collapse.

This perspective is an alternative to the “Copenhagen Interpretation”, which posits that superposition fails when an observation is made (but which is not objective in nature) and is also an alternative to the “‘Many Worlds’ Interpretation”, which states that the alternative results of a superposition are equally "real", while their mutual decoherence precludes subsequent observable interactions.

Penrose's interpretation rejects these two observer-dependent interpretations (both being subjective theories), advocating for a form of objective collapse theory. In this theory, the wave function is considered a physical wave, and the collapse of the wave function is posited as a physical process, with observers playing no special or causal role.

Penrose theorizes that the wave function cannot be sustained in a superposition beyond a certain energy difference between the quantum states, and this

threshold will then be related to the gravitational influences of the particles.

The "Objective Reduction" proposal would have its start determined by a condition referred to as the "one-graviton" criterion. The Diósi-Penrose proposal provides an objective physical threshold, indicating a plausible lifetime for superposed quantum states. This proposal suggests that each OR event, which is a purely physical process, is itself a primitive type of "observation," a moment of "protoconscious experience."[43]

To achieve this, it is necessary for the superposition to avoid immediate environmental decoherence and persist until a certain time limit is reached. This is accomplished by arguing that a quantum superposition is, therefore:

- 'orchestrated', i.e. appropriately organized, imbued with cognitive information and capable of integration and computation;
- isolated from an unorchestrated random environment long enough for the superposition to evolve into collapse and then create a moment of consciousness.

[43] Original publication: Hameroff, S. & Penrose, R. (2014) "Consciousness in the universe: A review of the 'Orch OR' theory" *Physics of Life Review*, 11 (1): 39-78.

The issue with this approach is that the Schrödinger equation is considered to describe the quantum formalism of a system at zero temperature: it would be absurd to consider that a conscious brain is in a thermal environment close to zero (on the contrary, it is far from it!).

We also know that current quantum computers require temperatures very close to zero degrees on the Kelvin scale to be functional. Therefore, one can argue that, considering relevant quantum activities in the brain at temperatures far from zero, this can be highly problematic and pose an obstacle to the entire advanced theory proposed by Penrose.

Now, the interesting thing is that, once again, it appears that biology and nature have already discovered several ways to develop specific thermal mechanisms that promote quantum coherence and avoid decoherence.

As we have already pointed out, there is evidence to show that plants routinely use electron transport (with quantum coherence) at room temperature in photosynthesis.[44] But... what about the human brain? Is there any evidence?

In 2009, Anirban Bandyopadhyay and colleagues at the National Institute of Materials Science in Japan used

[44] Original publication: Hildner, R., Brinks, D. et al. (2013) "Quantum coherent energy transfer over varying pathways in single light arvesting complexes", *Science*, 340 (639): 1448-1451.

nanotechnology to address the electronic and optical properties of individual microtubules. They found that quantum effects can occur in microtubules at biologically relevant temperatures (i.e., far from zero), suggesting that the existence of coherent quantum states in microtubules at brain temperatures is a real possibility.[45]

Furthermore, there is another curious piece of evidence from the realm of biology, not from a complex and highly developed organism, but from a very simple unicellular organism called 'parameciu m.' Despite lacking a single neuron, this organism can move, procreate, and feed itself. How does it achieve this? Through the use of specific structures called... microtubules.[46]

While this theory of consciousness is undeniably complex, it endeavors to address two of the most profound scientific and philosophical challenges of the 21st century: elucidating the existence of consciousness in human beings and attempting to reconcile two seemingly irreconcilable theories into a 'Theory of Everything'—quantum mechanics and Einstein's theory of relativity.

---

[45] Original publication: S. Sahu, S. Ghosh, K. Hirata, D. Fujita, A. Bandyopadhyay (2013) "Multi-level memory-switching properties of a single brain microtubule", *Applied Physics Letters*, 102: 123701.
[46] Original publication: Nakagaki, T., Yamada, H. e Toht, Á. (2000) "Maze – solving by an amoeboid organism", *Nature*, 407: 470.

With this proposal, Penrose provides insights into solving both of these challenges simultaneously. Are you convinced?

After these introductory notes, I hope you'll find it easier to follow the dialogues with these four incredible scholars with whom I had the honor of exploring the nature of consciousness.

Whether it leaves you more enlightened or confused, I think we can agree that a clear confusion is always better than unclear certainty.

DIALOGUES I

Consciousness

DIALOGUES I

Consciousness

VI. Dialogue with David Chalmers

David Chalmers is Professor of Philosophy and Neural Science and co-director of the Center for Mind, Brain, and Consciousness at New York University. He is also Distinguished Professor of Philosophy at the Australian National University.

He received his PhD in Philosophy and Cognitive Science at Indiana University in Douglas Hofstadter's Artificial Intelligence research group.

Chalmers is known for formulating the "hard problem" of consciousness and his work on "the extended mind," the idea that the technology we use can literally become part of our minds. His work on language, metaphysics, technology and artificial intelligence has also attracted much interest.

He is co-founder and former president of the Association for the Scientific Study of Consciousness and is co-director of the PhilPapers Foundation.

He authored *Reality+: Virtual Worlds and the Problems of Philosophy* (2022), *The Conscious Mind* (1996), *The Character of Consciousness* (2010) and *Constructing the World* (2014).

More information: https://consc.net/

Question: In your book *The Conscious Mind* (1996), you introduced the 'hard problem of consciousness' in contrast to the 'easy problems of consciousness,' which are associated with the neural correlates of consciousness and their implications for behavior. After all these years, do you still consider the problem of consciousness to be as “hard” as it was back then?

David Chalmers: I would say we are making progress on understanding the problem in various ways, so there is definitely forward motion. That said, I do not think anyone has solved the hard problem yet, and I think it is still fundamentally a very difficult problem. The basic contrast I make is between the “easy” problems of consciousness, which are that of explaining various behavioral and cognitive functions, for which we have a paradigm for explaining them, and the “hard” problem, which is the problem of explaining subjective experience, for which we do not have the same paradigm, since it looks like the standard methods of cognitive science leave open the question of why all that should give raise subjective experiences.

I still consider that the basic contrast is still there insofar as if you ought to offer an ordinary cognitive science explanation, it will not solve the hard problem. So, we need something new. Having said that, the Science of Consciousness has been developing very well without having to solve the hard problem by, for example, looking for neural correlates of consciousness, maybe

even looking for theories of consciousness like IIT (Integrated Information Theory), which do not try to reduce consciousness, but try to connect it to physical properties and the brain.

Ultimately, it may be that the best we can do for the hard problem is something like an outline of the fundamental principles that connect consciousness to physical processes, and maybe theories like IIT are trying to do that. I do not think any theory has achieved the kind of evidence and consensus that would be required to be actually accepted as a theory.

Meantime, people have been exploring a lot of different ideas which are much better understood than they were 30 years ago, whether it is panpsychism, or the quantum mechanics approach, or illusionism theory. I think important progress has been made on each of these, but fundamentally, I think the hard problem is about as hard as it ever was.

Question: How do you envision a rigorous scientific approach to understanding the nature of consciousness? Would this entail a robust foundation in mathematics and formal validation alone, or do you believe it's essential to integrate empirical data and evidence with mathematical proofs? In other words, what criteria do you find relevant for the development of a serious science of consciousness?

David Chalmers: I see the science of consciousness as all about integrating third person data, objective descriptions of cognitive systems, with first person data, the kind of data you get from subjective experience.

You get the objective data from standard methods, like a measurement observation, especially of behavior and of the brain. You get first person data from attending to subjective experience and we want the science of consciousness ultimately to have principles connecting the first-person realm to the third-person realm, ultimately formulated in a rigorous way.

This is going to require several relevant steps. First of all, methods for gathering first person data which are as rigorous as our methods for gathering third person data. Even though people have thought about this quite a lot, our methods are still quite primitive.

We will also need methods for formulating the structure of consciousness in rigorous terms. You mentioned mathematics: here is one place where I think mathematics can come in – we can work to find mathematical descriptions of the structure of consciousness. And it may be that, at least partially, mathematical principals connecting the structure of physical processes (i.e., third-person data), to the structure of subjective experience (i.e., first-person data).

I guess you can see something like that happening with IIT: I think that is a very promising form for a science of consciousness. I mean, a mathematical description of consciousness will not exhaust consciousness. The thought experiment "Mary in the black and white room" might bring that out. She could know the mathematical structure beforehand, but actually to experiencing the colour red first the first time tells her something new.

Nevertheless, I think the mathematical structure of consciousness is actually – at the very least – a very good partial characterization of consciousness. And that may provide what some people have called an "objective phenomenology" that could play a very crucial role in developing the science of consciousness.

Question: In your latest book, *Reality+: Virtual Worlds and the Problems of Philosophy* (2022), you delve into the 'simulation hypothesis.' In essence, you propose that science fiction scenarios, like 'The Matrix,' might have more plausibility than commonly perceived. This hypothesis suggests our current reality could be a simulated world crafted by an advanced civilization. Additionally, you contend that beings within these simulations could possess consciousness akin to ours. Could you elaborate on the reasoning behind your belief in this possibility

David Chalmers: This partially comes back to the old question of whether you need biology to be conscious or whether it is more a matter of information processing, computation and functional organization.

I have always been on the side that says that what really matters for consciousness is not the specific biology, but something more like the functional organization, the information processing.

One way that I brought that out, back in *The Conscious Mind*, in the 90s, was to conceive this thought experiment of gradually replacing your neurons by silicon chips and arguing that, if you gradually did this with a good enough preservation and functional organization, this would actually preserve consciousness.

I guess that does not quite get us to simulations: that gets us at least to silicon isomorphs of us being conscious, but I think once you have gotten this far, it is not a long step to simulated beings that are conscious too.

Why? Well, roughly, a simulation of my brain will actually be basically analogous to a silicon isomorph of my brain: it will have a lot of interacting parts which are processing all the same information. I would argue that, if what matters is the structure or the information processing, all that can, in principle, be present in a good enough simulation.

Now, it is true there is still the hard problem there, so we do not really understand how it is that a simulation could give you consciousness. But we equally do not understand how a brain could give you consciousness. I just argue that the simulation is on a pair with the brain here.

Question: Interesting. I would argue, though, that even the processing of information is inherently tied to the material 'component' of the system. In other words, there is something truly unique and special about the organic biology that the evolutionary process has shaped, serving as the foundation for conscious experience in humans. However, I assume you don't share this view, asserting that the material composition of this substrate is not crucial, as long as it retains its functional capability. Is that correct?

David Chalmers: I would claim the substrate can make a very significant difference to how information is processed, and I am sure that the structure of neurons, for example, makes a big difference to how information is processed in the brain.

That said, I do not see why that cannot, in principle, be simulated. Whatever the idiosyncratic properties of neurons are, it seems to me, as far as I can tell, there is nothing there which is *uncomputable* or *unsimulatable*.

Now, if it turns out that Sir Roger Penrose is right, then it might be the case that there are some special quantum mechanical processes in neurons that cannot be simulated on a classical computer. And then we would need more work.

But even if Roger is right, I would still wonder if maybe there could be some special new kind of quantum simulation. And we have quantum computers, which basically exploit certain physical properties of quantum mechanics.

If Roger is right, maybe things will go far beyond that: quantum gravity will involve new kinds of processes that cannot be simulated, even on an ordinary quantum computer. But even then, we may be able to build new quantum gravity computers that can exploit Roger's special kind of computation and, in principle, I do not see why we could not build a simulation of a brain on one of these special new quantum gravity computers.

That kind of simulation would go beyond simulation on a classical computer, but I think it would still be an interesting approach to simulation.

Question: The problem of consciousness has been a part of philosophy for hundreds, even thousands of years, but some argue that philosophers haven't made significant contributions. How can philosophers actively contribute to finding a solution to the challenging

problem of consciousness? What, in your opinion, is the real role of philosophy in addressing the hard problem of consciousness?

David Chalmers: What can philosophy do? That is a good question: I think philosophers have explored different approaches to addressing the hard problem, including over the last 20 years, we have seen a lot of interesting work on panpsychism, we have seen proposals for addressing what makes panpsychism really hard, which is the combination problem, by philosophers such as Philip Goff, Hedda Mørch and Galen Strawson.

You also have the theory of illusionism about consciousness, developed by cognitive scientists, but it has been really pushed by a lot of philosophers, including Dan Dennet, Keith Frankish and others.

So, philosophers really have been pushing forward on possible solutions to particular aspects, at least, of the hard problem. I myself do not care that much whether a philosopher or a scientist does it, but I strongly suspect that, to solve this problem, it is going to involve some kind of interaction between philosophy and science.

Question: Building on the ideas of Sir Roger Penrose, do you believe that we require a new "kind" of physics to make progress in solving the hard problem of

consciousness? What theories do you think might be successful in addressing the nature of subjective experience?

David Chalmers: What kind of theory might explain consciousness? I like the idea of the "mathematics of consciousness": mapping the structure of consciousness and mapping that on to physical processes in the brain.

My own view is that a theory of consciousness may well require psychophysical laws that connect physical processes to consciousness. And, ultimately, fundamental psychophysical laws.

It can be, for example, panpsychism, but does not have to be panpsychism: a dualist could believe in fundamental psychophysical laws as well. Giulio Tononi might be seen as having a proposal for a theory of consciousness as Sir Roger Penrose.

I mean, I think what is the correct theory is going to depend a lot on the development of the science. You might still object, though, that a purely mathematical theory of consciousness is still going to be subject to the Mary problem.

So, I would argue that, if we can distinguish between the qualitative character of consciousness from its structural character, we might, at least, end up with objective mathematical psychophysical laws that can explain the structural character of consciousness.

There may still be some further work to do on the qualitative character, like the redness of red.

Question: Some neuroscientists and philosophers take the position that consciousness is nothing more than a category error, a concept that no longer makes sense in our scientific and philosophical vocabulary. Why should we care about the problem of consciousness in the first place? What reasons can be offered for why we find this philosophical problem really relevant?

David Chalmers: I cannot convince anyone to care about anything: to some extent, that is up to each reader. There are many things to be interested in the world and not everybody has to be fascinated by the hard problem of consciousness. That is totally fine.

Why is consciousness interesting and important? Just say you were only interested in predicting other people's behavior, then for that purpose maybe you could get away without attending to consciousness.

That will depend on a lot of complicated questions, like if interactionist dualism is true; you may have to attend to consciousness. But I think we care a lot more than other people's behavior: we care about other people for much more than caring about their behavior.

I think that, for many people, it is precisely because other people are conscious that we care about them.

So, consider moral and political questions like “how we should treat animals?”: I think it is absolutely crucial to provide an answer to know whether animals are conscious and what kind of conscious states they have.

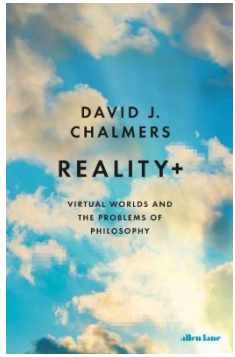
For example: “are they suffering?”, and you can say “I can know everything from the animal’s behavior”. Great, but that does not tell me the crucial thing I need do know, namely, how I should morally treat animals, which is knowing something about their subjective experience.

If you take the view that subjective experience is really at the basis of value and meaning in our lives, then this is something we are going to have deal with to figure out the answer to some of these important practical questions: that is at least one reason to care about consciousness.

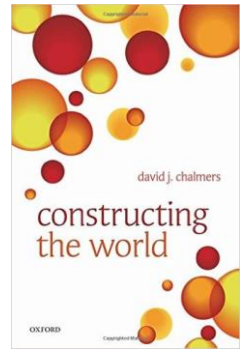
I think there are also intellectual reasons: it is incredibly interesting, it is anomaly in our picture of the universe, and if we do not have a theoretical way to understand it without a theory of consciousness, we will not have a full theory of the universe.

Anyway, there are two reasons that I can offer to make a case about the importance of consciousness.

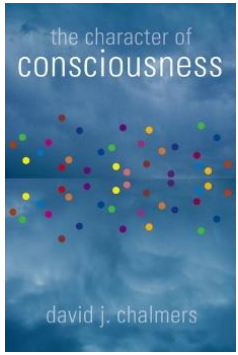
Books by David Chalmers



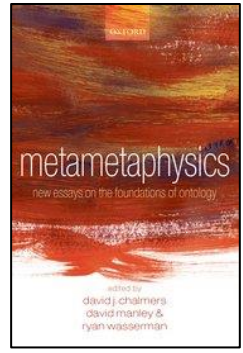
2022



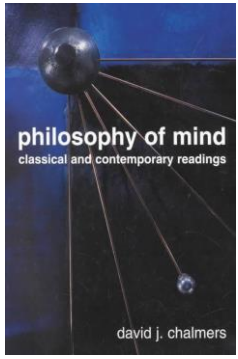
2012



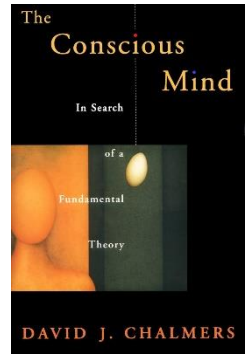
2010



2009



2002



1996

VII. Dialogues with Susan Blackmore

Susan Blackmore is a writer, lecturer, and Visiting Professor at the University of Plymouth, United Kingdom.

She has a degree in Psychology and Physiology from the University of Oxford (1973), a Masters and PhD from the University of Surrey (1980).

Her research interests include memes, evolutionary theory, consciousness, and meditation.

Professor Blackmore writes for numerous magazines and newspapers, blogs for *Psychology Today*, and is a frequent contributor and presenter on radio and television.

She is the author of more than fifteen books, sixty academic articles and around eighty book contributions.

Some of her books are: *Dying to Live: Near Death Experiences* (1993), *The Meme Machine* (1999), *Conversations on Consciousness* (2005), *Zen and the Art of Consciousness* (2011), *Seeing Myself: What Out-of-body Experiences Tell Us About Life, Death & the Mind* (2017) e *Consciousness: An Introduction* (2018).

More information: https://www.susanblackmore.uk/

Question: You were one of the first scholars to give greater philosophical relevance to altered states of consciousness. Let's start with your "erratic" experience that happened when you were younger: you had a kind of "out-of-body" experience, and this greatly influenced your interest in consciousness in general. Can you tell us more about that personal story and the impact it had on your research?

Susan Blackmore: The whole experience lasted about two and a half hours. I was sitting around in a friend's room: it was my first term at Oxford, I was 19 years old (so, 50 years ago!). I was very tired, I was sleep deprived, and I smoked a little bit of cannabis, but not enough to explain it, even though it probably contributed. I started to go down a "tunnel", a tunnel of trees, with leaves all around, towards a light. Now, this was before the term 'near-death experience' was even invented.

During that time, I did not know anything about this topic. The tunnel led into an out-of-body experience: I seemed to be out of my body and could look down and see my own body. I was still talking; I had a friend talking to me and he kept saying "what can you see now?" and I went on traveling – what seemed like traveling around the world – and seeing all sorts of things.

To cut this long experience short, I tried to get back into my body and it was really difficult. And I got too small, and I shrank, and shrank, and shrank and became very small. And then I got very frightened, and I got bigger

and bigger and bigger, and it expanded into a classical mystical experience, which again I knew nothing about. I became one with everything. There was no longer any self, but there was something which seemed to me to be everything.

Many other experiences happened along the way, but eventually I became exhausted and my friend said to me “Well, isn’t there anything else?” and I thought “No, because I am everything, how could there be anything else?”. I mean, thoughts were still going on, but not like really in words. And then I had a kind of realization that there is always something more. It took me two days to get back to feeling that I was inside my body again.

This experience made me believe in all sorts of psychic and other-worldly things. This was illogical, of course, but understandable at the time for a 19-year-old who did not know anything about these experiences. And in the early 1970s there was no neuroscience to give us any answers. From this experience, I started to believe in telepathy, clairvoyance, psychokinesis, ghost, poltergeist, everything! Worlds beyond, life after death, soul, spirits, etc.

Because of this experience I decided to not accept a sensible PhD which I was offered, at a great institution, and to do my own PhD on the paranormal instead.

It took me about five years of research into all sorts of paranormal claims to discover that they almost

certainly do not exist. So, that was the quick story of how I became transformed from a believer in all kind of weird stuff to deciding that the real questions are not: "Is there a spirit, or a soul, or consciousness beyond the brain?"

The real question is right here now: "What is this experience? How does a brain do it? How does a brain and a world and a body do it? What is consciousness?" And these are much more exciting questions than whether there is life after-death, or spirits, or whatever.

Question: Following this experience, we know that various drugs can induce altered states of consciousness. Do you think we can use psychedelics like DMT or LSD, for example, to study the nature of the conscious mind? Do you believe these methodologies are useful in understanding the phenomenon of subjectivity, considering the current methodological difficulties in the scientific investigation of consciousness?

Susan Blackmore: It depends on what you mean to study "the nature" of the mind. If you think they are going to give you quick answers, then no. But more generally, yes. It will depend whether you mean: i) can you, in your personal life, begin to understand through having those experiences or ii) can the research that is going on provide us explanations.

I would say this because: the mind is so radically changed when tripping that it can tell you some things immediately. It can tell you 'this normal state of consciousness is not the only one'. Think of what William James said in 1890, that beyond the veil of this experience, there are endless other ways of being conscious that are just very close, but we need something like a drug to take us there, or a spontaneous mystical experience.

I think one of the most exciting pieces of research I have read on this topic is the very recent paper by Timmermann et. al. (2019)[47] on DMT which, as I expect you know, is the major psychoactive ingredient of ayahuasca that disrupts the major functional networks of the brain.

There are several major networks, but one of them in particular, the default mode network (DMN), is really the one that underlines the sense of self: this is the long-range network that pulls together the body schema with your memories and your opinions about the self. It connects to the right temporoparietal junction and it is here that the body schema links up with control systems in the frontal cortex, and with memory in temporal lobes, and so on. This network is disrupted when you take DMT while elsewhere in the

[47] Original publication: Timmermann, C., Roseman, L., Schartner, M. et al. (2019) "Neural correlates of the DMT experience assessed with multivariate EEG", *Sci Rep*, 9: 16324.

brain, for example in the visual cortex, there is increased local activity which may explain the visual hallucinations that occur with DMT.

When the selfing-system is disrupted, that is why you no longer feel there is a self. Now, I would say that this means seeing through the illusion of the powerful conscious self, the self we think of as having consciousness; the self that we think of as having free will. Both of those are illusions in my mind, and the fact that DMT disrupts the self means you can – for a few hours, if it is ayahuasca, or 15 minutes if you smoke DMT – have this experience without the normal sense of self.

That, at least, helps you to see the beginnings of one aspect of the illusions of consciousness. I believe most theories of consciousness remain trapped in the illusions and are going nowhere. So, maybe psychedelics can help in that way and I think all the research we are doing now is really beginning to reveal a lot about the way the mind works.

And there will be more to find out about: now that it has started and the law cannot really shut the research down anymore, I think we will learn a lot.

Question: In the realm of exploring consciousness, particularly through phenomena like lucid dreams, where the experience closely mirrors wakefulness

rather than deep sleep, how do you perceive the potential contributions of empirical investigations into these dreams? Can such studies provide valuable insights and aid in the comprehensive mapping and understanding of diverse states of consciousness?

Susan Blackmore: When you say that lucid dreams are near as to being awake, I am not sure what you mean. They take place, mostly, as far as we can tell – and I am sure there are variations – in REM sleep, and that means you have to be properly asleep. You have to have gone through the cycles of sleep, even though in some exceptions, you can get REM right at the beginning, if you are really exhausted and sleep deprived. But most lucid dreams happen in REM.

They do happen in more active periods of REM, if that is what you mean, where there is more activity going on than the lower activity periods of REM. But the really interesting thing is that – and it again concerns the default mode network – the connections between the temporoparietal junction and the frontal lobes becomes stronger in lucid dreams, which suggests that this is a psychological basis for the sense that "oh, I'm here now and I can control the dream!".

Again, I love finding out these things, because they just blow apart the sort of alternative theories about what is going on in a lucid dream, such as our souls woken up, and all the kinds of things that people imagine. So, unfortunately, standard psychology does not really

cover things like lucid dreams: there are not enough researchers doing research on lucid dreams, but I think they are a source of learning a lot about our brains and our minds.

Question: With your extensive experience of around 30 years in Zen Meditation, how has this practice influenced your perspective on the conscious mind? In what ways do you think meditation, with its focus on mindfulness and self-awareness, provides insights into the nature of consciousness? Additionally, considering the subjective and introspective aspects of meditation, how do you navigate the challenges of incorporating such insights into our understanding of a broader scientific study of consciousness?

Susan Blackmore: It depends of what you mean by understanding, in the sense that I am not sure meditation has that purpose. I think its main goal is to clarify the mind: and that is the best reason for wanting to meditate. In a way, for me, it has helped me to understand things about the mind, but the main practice is to drop the illusions, to drop the belief that there is me in here, and the world out there, to drop the craving for becoming something more important, and I have a lot of craving to be important, and it still there, after 40 years (so, maybe, it does not work that much hehe!).

But it changes your mind quite radically but slowly towards being more open to everything in the world, more accepting, less grasping: that is what meditation is meant for. Whether it can help us understand the mind? Well, yes, of course: think of long-term meditators that are very well practiced at going into different states, then you can do research with them.

I am particularly interested recently in these last few years in the Jhanas Meditation, in which there are said to be eight discrete states of consciousness that can be entered through deep concentration. I heard about this decades ago, and I thought how amazing it would be if there were specific altered states that you could enter just by following specific instructions.

I thought I would never be able to do that. But then a great Jhanas teacher turned up in England and I met him and immediately wanted to go on his retreats. I have been on several now and practiced this specific type of meditation and I found that it is actually possible, through simply following these quite intense instructions, to attain these states.

This begins with deep concentration, and then various other things, to go into clearly demarcated altered states, and you can shift between the different states. I can only do the first three states pretty well, and the fourth perhaps, and I just keep practicing and maybe I will be able to find some of the others.

But certainly, from the practice I have done, I am convinced that these ancient people who practice this meditation really discovered these different states and how to get to them. So, that tells something about what the mind is capable of, and when you can do that or even just read about and understand what people are saying about it, then you know that mind is capable of being in a completely different kind of relationship to the world and to its idea of itself, if you like. So, yes, the answer is broadly yes.

Question: It seems that various altered states of consciousness can be identified through neuronal imaging techniques. Given your expertise in this specific meditation, do you believe it is plausible that, with sufficient practice, distinct patterns of brain activity corresponding to the eight phases of this particular meditation could be observable through neuroimaging? How might the integration of neuroscience and meditation practices enhance our understanding of consciousness, particularly in unraveling the neural correlates associated with specific states of altered consciousness?

Susan Blackmore: As far as I know, there have been only two experiments on that, both done with Leigh Brasington. It is very expensive doing brain experiments with fMRI or other kinds of imaging technology.

The problem is that the subject has to be able, in this very deep concentrated state, to indicate to the researchers when he is changing to the next state, and to tie that up with his brain activity. Certainly, the experimenters found changes at the moment when he says he is going from state two to three for example.

But we do not have enough data, and I do not think Leigh was able to go through the eight states, because the last two states are what you might call "unconscious".

I mean, they are so far gone, that it would be very hard to communicate them in that experimental setting – but that would be incredible for sure. It is really, really hard to meditate properly in a fRMI scanner, with all the noise and everything. So, I think you are being a bit hopeful there.

Question: The investigation of consciousness poses significant challenges, and individuals like Deepak Chopra often promote (dubious!) spiritual ideas about the conscious mind. In your engagements with him and discussions on such topics, what is your perspective on these spiritual ideas? How do you approach the balance between exploring the mysteries of consciousness and maintaining scientific rigor, especially when addressing concepts that may lack empirical support or scientific grounding?

Susan Blackmore: If anyone is interested in Deepak Chopra look on YouTube for my debate with him and watch the video,[48] because what you will see is a self-proclaimed spiritual guru, behaving amazingly badly.

I cannot answer your question for all those ideas, but I can answer about Deepak Chopra. He says that consciousness is primary and matter does not exist. That is just senseless: it does not mean anything and most of what he says does not mean anything. The practical advice he gives for skills of how to train your mind – and some of the things he says about meditation and its consequences are very precise, he really understands about that – but when it comes to his own theorizing it is very unsatisfying.

Consider the simple philosophical problem, the mind-body problem. If you are a materialist, everything is matter, you cannot account for subjective experience, that is what we call "the hard problem of consciousness" and it is not solved. Maybe it is actually the wrong question, the wrong problem. But if you are an idealist, you say everything is consciousness, but then you cannot explain matter.

So, clearly, neither of those works. Dualism does not work, because you have two completely different kinds of things. So, we have got something deeply wrong,

[48] Original video: youtube.com/watch?v=_ZFGkqhNhgM.

which is why I am a so-called illusionist, but Deepak is just an idealist, while not admitting that he is an idealist.

Deepak just argues that everything is consciousness, and matter does not exist. Great! This makes no predictions, has no theoretical basis, there are no conclusions you can draw from it, and yet people love it! People love it because they think: "somehow my consciousness is so wonderful and important, and I can do this with my consciousness! He is really into consciousness, so he must be very spiritual!".

Of course, I am caricaturing it, but I think his ideas, in terms of any scientific underpinning, are just vacuous! He calls on scientific research and distorts it horribly to try to fit his ideas. But I am not going to speak for every other spiritual teacher, because they are extremely varied.

Question: As an illusionist, you reject both materialism and idealism in understanding consciousness. Can illusionism be seen as a synthesis, proposing that the brain, as a material entity, generates the illusion of consciousness? How does illusionism navigate between the contrasting views of materialism and idealism in the realm of consciousness?

Susan Blackmore: It is a very good question. Somebody asked me that question before at the psychedelics conference, and all I can say is: most

illusionists are materialists, that is for sure. On the other hand, I do not have anything to replace it with.

But my idea of illusionism is pretty broad: that all of these existing ways of thinking about it are wrong, and we have not yet discovered a way out of this dilemma, and calling it the "hard problem" and trying to look at how consciousness arises from the brain – which is how it is described – is a kind of dualist thing, even though all these materialists – with the exception of David Chalmers, who invented the term "the hard problem" – carry on being materialists without really solving the problem.

They are, like me, saying something like, "if we think consciousness is something that we have, that has power, that does something, and that evolved for a purpose, then we are deluded and have got it wrong". I think Dennett and Frankish, for example, would say "Yes, there is a material brain from which the illusions are constructed", and so they are trying to solve the problem of how those illusions come about.

So, they replace the hard problem with the illusion problem, or the meta-problem, as Chalmers calls it. I am very happy with that change. I am not a philosopher but, if people ask me, I would say I am a neutral monist. I am monist because I think dualism does not work.

But I do not think that our present concepts of "material" or "matter" or "mind" are helpful. I do not

know what the universe consists of. Maybe fundamental physics can help. It is getting itself into such big tangles, with being unable to combine quantum mechanics with Einsteinian theories.

You can also think of some recent informational theories claiming that all there is out there is information. And then you can also think about thermodynamics, and entropy. That is the kind of way I am thinking that somebody may solve the problem of consciousness, but I am trying to teach myself some of that stuff and it is a bit difficult.

Question: Considering your illusionist position on conscious states, where you argue that consciousness is essentially an illusion created by the brain, could you elaborate on whether you perceive a potential interconnection between the illusion of consciousness and the illusion of the self? In other words, how would you describe the intricate relationship, if any, between these two illusions within the framework of your perspective?

Susan Blackmore: I suppose the best thing I could say is they are extremely close. The beginning of the illusion of consciousness is the separation of self from other. The problem of other minds, if you like; the separation of me from other people. The feeling that this is my consciousness and mine is different from theirs,

because I cannot know what is like for them to see blue and whether it is the same as my blue and all those kinds of things. Is it the whole of the illusion? No. I would say that the things that interests me in terms of possible illusions are mostly the traps.

If you are familiarized with Dan Dennett's work, particularly *Consciousness Explained*, you will see that he talks about all the traps that people fall into, and he continues that work in later books such as in his *Intuition Pumps*.

His Multiple Drafts Theory really leads you to what, for most neuroscientists, is totally bizarre: it leads you to say there is no distinction between brain processes that are conscious and those that are not conscious. There is no fact of matter about it, it is a meaningless thing to say these brain processes going on here are the consciousness ones and others are not.

The whole search for the neural correlates of consciousness in my mind is completely wrong. It is built on these illusions, because it is saying that consciousness itself emerges from some particular process, some particular area. But according to Dennett's theory, that is simply not the case.

I also like to ask the question: "Are you conscious now?!" You may have had an extraordinary experience when you are asked this question. You probably thought: "Of course I am! But, hang on, one moment ago I was just

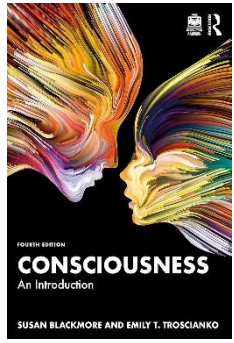
listening to what she was saying, but now something has changed. I have become more conscious!".

That leads me to say – again related to Multiple Drafts Theory – that if you do not know whether you were conscious a moment before you thought about it then nobody knows. We cannot put a consciousness-meter on the brain and say where consciousness is or what you were conscious of at any particular moment. So, if you do not know, Dennett would say that there is no fact of matter about it. All we do is retrospectively, after the fact, say "I was conscious of this and not of that", it is always attribution after the fact.

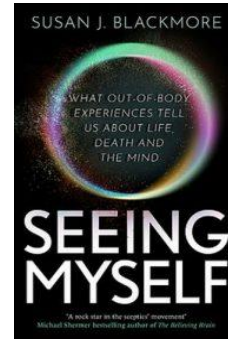
To come back to your question, are self and consciousness based on the same illusion? In a way they are, because it is me who is conscious, and that involves the separation between me and this thing called my consciousness and the things I am conscious of. So, I suppose my answer is "Yes" and "No". Yes, are closely related. But no, you can tease them apart into lots of different levels of illusion.

I think that the field of consciousness studies is mired in illusion and we will not make any real progress until we understand how these illusions come about and learn to see through them.

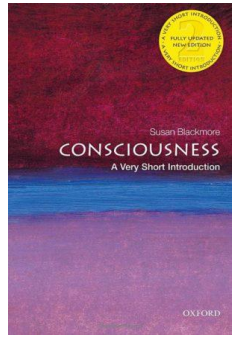
Books by Susan Blackmore



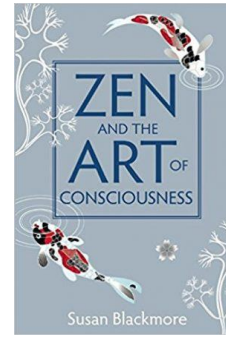
2024



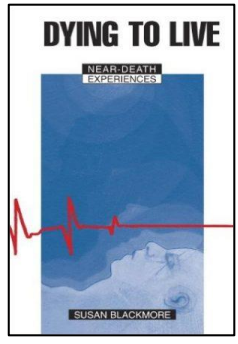
2020



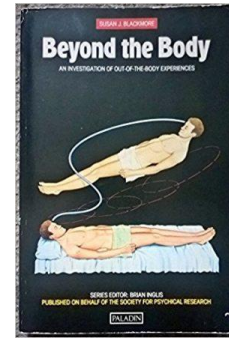
2017



2011



1993



1992

VIII. Dialogues with Nicholas Humphrey

Nicholas Humphrey is Emeritus Professor of Psychology at the London School of Economics, Visiting Professor of Philosophy at the New College of the Humanities and Senior Fellow at Darwin College, Cambridge.

He is a theoretical psychologist, known internationally for his work on the evolution of human intelligence and consciousness. His interests are varied: he studied mountain gorillas with Dian Fossey in Rwanda; he was the first to demonstrate the existence of "blindsight" following brain damage in monkeys; proposed the famous theory of the “social function of the intellect”.

He is a prolific author having published several books: *Consciousness Regained* (1983), *The Inner Eye* (2003), *Seeing Red: A Study in Consciousness* (2009), “*A History of the Mind* (2012), The *Mind Made Flesh* (2012), e *Soul Dust: the Magic of Consciousness* (2012).

He recently published his new book *Sentience: The Invention of Consciousness* (2022).

More information: https://www.humphrey.org.uk/

Question: In the summer of 1966, you had an important encounter with Helen, a blind monkey whose visual cortex had been removed for scientific research. Could you delve deeper into the impact that this encounter had on your understanding of consciousness? Specifically, what lessons or insights did you draw from this experience, and how did it shape your subsequent perspectives on the nature of consciousness?[49]

Nicholas Humphrey: It was a very strange experience when I first came across this monkey. I was a research student in Cambridge University and in the lab, there was this monkey who had been operated in order to discover what is the function of the visual cortex. So, a colleague of mine had removed all the visual cortex of the brain of this monkey and, not surprisingly, she appeared to be completely blind.

I met Helen a year after the surgery and she just stay there sited, staring vacantly into space: she was not interested in using her eyes at all. But I was puzzled about that, because Helen had her cortex removed at the back of her brain, but it left intact most of the ancient visual system which is the visual system used by fishes, frogs and other non-mammalian vertebrates, the superior colliculus midbrain visual system.

[49] Helen in action: youtube.com/watch?v=rDIsxwQHwt8.

This system enables them to see perfectly well: a frog can use its eyes and catch flies, and so on, without any problem. Since this system was still intact in this monkey, I was wondering whether in fact it could be put back to use, even if the "normal", non-operated monkey does not use it. Could Helen recover her vision using this midbrain system?

One time, my supervisor went away for a week for a conference in Zurich and I took that chance to sit with this monkey and play with her, to try getting her to interact with me whatever way she could. To my astonishment, I found that within a few hours I could get her to use her eyes. She was obviously attending to what I was doing; I did wave a piece of fruit up in front of her, for example, so she could reach out and take it from me. And, by the end of the week, she was reaching out to touch a small light which I had in front of her, or an object which I fiddled on the end of a stick.

So, of course I was very excited about that, and I sent a telegram to my supervisor in Zurich but he did not like the content of that message. I sent a telegram saying "I have taught Helen to see. You will not believe it": I was 20 years old, he was a major senior professor and he was not too pleased about that. He came back to Cambridge and a day or two later I persuaded him to come and see my monkey, or his monkey which I transformed, and he had to agree something

astonishing had happened. She was clearly able to use her eyes again.

The upshot of this first experience with Helen was that my supervisor allowed me to continue to work with her and ended up working this monkey for seven years. And, by the end of seven years, she apparently was able to see everything normally. She would run around her room, picking up objects of the floor, not bumping into obstacles, she could even reach out and capture flies as they passed by her.

But you might think, and many people did think, that it looked like she had a normal vision, but I was puzzled by it: something was not normal about it, because when she was with me, she was relaxed and confident, and then she could see, but if she was threatened or anxious at all, her vision would disappear, she would blunder around, as if she was in the dark again. So, it seemed that she could only see if she did not have to think too hard about it. If she did, her confidence deserted her.

I thought it was a very extraordinary form of vision: it was a vision which the monkey herself does not believe in. So, I wrote a paper which I called this capacity "seeing and nothingness", echoing the Jean-Paul Sartre's book , since I thought that there was something clearly missing. And we soon found out what was missing, because my supervisor went on this time to test human

patients in a different way, and he established the famous phenomenon of “blindsight” in humans.

He discovered that humans with major damage to the visual cortex, who believed they were blind, could, in fact, use their eyes to see in the area of their blind field. They could guess what was up: they could guess the shape and position of an object in the visual field, but all the time they would say “there is nothing there”, “I do not understand”, “this does not make sense”, “my vision has nothing to do with me”.

From this, I went on to ask the big question, of course: if you can see, and a man can see, and a frog can see, and a fish can see without using the visual cortex; if you can see, as the human case shows, without having visual sensations, then what is the secondary system for, and what is the point? What is the use of visual sensations? And that is really the question I have been working on since the rest of my life. The last 50 years I have been trying to discover the functional role of physical sensations.

And, of course, that has led me on to asking another big question, which is: why do visual sensations and all other sensations have the very strange phenomenal qualities they do? What is it like to see red, or to taste sugar, or to hear a screeching, or clanging bells? We are not just getting the information about the object, we are getting information in a different dimension, in the

phenomenal dimension, of how it relates to us, what is like for us to have this information coming in.

Question: You assert that consciousness is a product of evolution because it confers specific survival benefits on humanity...

Nicholas Humphrey: I do not say it is a product of evolution because it confers survival advantages. What I ask is: what else can it be, but a product of evolution? At least in that case it must confirm survival advantages. So, we have to discover what those advantages are, exactly...

Question: But you also claimed in the past that consciousness is a form of illusion generated by our brains. This perspective characterizes consciousness as a mental construct, an internal representation of the reality we experience. Could you elaborate further on how you conceptualize this illusion and its role in shaping our subjective experiences? Additionally, what implications does this perspective have for our understanding of the nature of reality and our place within it?

Nicholas Humphrey: I used to claim that, yes. But that suggests that consciousness is an illusion. And I read several papers, and even other books describing it as

an illusion, and that now became a popular view among certain philosophers. Keith Frankish, Dan Dennet and others are what we call “illusionists”, but I have pulled back from that label.

I do not think that is the right way to describe it: to call something as an illusion suggests that it is a mistake, that we are in error in attributing the qualities we do to experience. I do not think that is right: when we see red or smell a rose, it really is like it seems to be, that is how we feel about it, it is what it is like. To describe that as an illusion, is to underestimate the role it plays in our psychology and in our phenomenology.

Phenomenal consciousness is a veridical description of what is like for a human being to have these experiences arriving. And the question is “why do we represent it in that way?”. Sensations are representations, they could have been just representations of the facts, bare physical facts, but they are not. They are representations of how we feel about having these stimuli arriving, touching our body.

Question: In your current view, you describe yourself as a "surrealist" regarding *qualia* or consciousness. Could you elaborate on the specific aspects of surrealism that you find apt in characterizing your perspective on these phenomena? How does this label capture the essence of your ideas about *qualia* and

consciousness, and what implications does it carry for our broader understanding of the mind and subjective experience?

Nicholas Humphrey: I was looking for another term different from illusionist. There was a paper which was written about illusionism, by Keith Frankish. I responded to it saying “No, I think it is not an illusion; it is not unreal; if anything, it is super real”. The phenomenal redness is redder than red; phenomenal pain is painier than pain.

And I took that away of phrasing it from Pablo Picasso, who was one of the earliest surrealists, although he did not accept the term. And he made a famous sculpture of a goat, and he said “my goat is goatier than any real goat”.

In other words, he was trying to express in his art the essence of the object he was creating in art and claim that it went beyond, it was deeper than the reality of the physical reality. And I think that maybe that is the right way to talk about phenomenal sensations too. In a sense, they seem to go deeper than the superficial facts of the case.

Question: You postulate a close relationship between *qualia* and what you term the "phenomenal self." Could you delve deeper into the nature of this relationship? How do qualia contribute to the construction or

experience of the phenomenal self, and what implications does this connection hold for our understanding of subjective experience?

Nicholas Humphrey: I think that *qualia*, the phenomenal experience, is the foundation of the self. And, indeed, that is the road that came to play in our psychological economy, and that is in fact why it is evolved. What sensations do is to give a substantial reality to our sense of who we are, and of our existence in the world.

Famously, David Hume, the philosopher, said that, when he tries to examine his own mind and discovered what means to be himself, he does not find anything other than sensations. And he was disappointed by that, since the sensations are evanescent, they do not seem to have any continuing reality and there is no substantial basis for the self, based only on sensations.

I think that is completely wrong, I think that sensations do in fact give us the most solid grounding we could possibly have for our existence in the world. They are ever present evidences of how we live our lives, and how we matter, of what is like to be ourselves and, what is more important, of our individuality. Because, for all we know, the experiences we have are unlike anything else in the world.

Now, of course we go on to assume that other humans have experiences like we do. But the evidence is not

there: all we know is that my sensation of red resembles the sensation of red I had before. It seems to have something in common with my other sensations and other modalities. It is all done in my style, but that is my evidence for my continuing reality.

I describe the continuity of sensations as being like the continuity of works of art by a particular painter. And just as it is all Cezanne's, all Vermeer's, all by the same artist, all my sensations are by me, and that gives me a reality under an importance in the realm of things, which I think is very significant.

What is important then is that, while we are growing up, we discover that this is what makes up our psychological center, and we then go on to assume that other humans have an equivalent center of self, in which the play of sensations is similar to our own.

For each of them it will be individual, private and important. Once we take that view of other people, it begins to change the parameters of social live. We have come to live in what I have called 'The Society of Selves', of Phenomenal Selves. And that is the basis, certainly, for human culture and human civilization. The big question is whether it goes beyond humans, and if other animals think of themselves like that as well.

Question: Following this idea, you also argue that consciousness has a specific role related to providing a

kind of meaning to life and human existence in general. Could you elaborate further on this specific aspect of consciousness, which is often overlooked by scholars and intellectuals who dedicate themselves to studying this phenomenon?

Nicholas Humphrey: That is a big question. To give meaning to life, yes. Well, where do we begin in finding meaning? First, it gives us a sense of our own importance, that we are not purely physical phenomena, that we exist on some other plan or have a spiritual dimension to our lives. We exist in some ways outside time and space and in a realm which cannot be described by the material qualities of physical matter.

Now, that is a very important discovery about ourselves and it is, of course, a basis not only for just a simple sense of "yes, I matter", but it can go on once it has been elaborated by culture and by language. It can give us the sense that we have an immortal soul.

You might think that it is an odd thing for a scientist to take seriously, but I do take it very seriously. The belief in the soul has been one of the driving forces of human history, and in fact it is responsible for most of the significant things that humans have ever achieved.

Because once we believe in souls and their importance, and in other people's souls and their importance, it gives us new ambitions for what we want to leave behind and what we want to achieve in our own lives.

Now, of course that is about human beings. I do not think dogs have sense of being souls, for example. But they, nonetheless, have a sense of themselves as being some kind of significant individuals and, for them, we have to tell a slightly different story.

For humans, consciousness has come to have quite unexpected and wonderful results. I have described in my new book: I think it is the jewel of the crown of biological evolution and we should take it very seriously.

Question: You also argue in your new book *Sentience: The Invention of Consciousness* that sentience is restricted to mammals and birds due to a very specific physiological characteristic: warm blood. This is, in fact, a curious and innovative characteristic that has not been considered before as a plausible explanation for consciousness in its sentient aspect. What led you to come up with this idea?

Nicholas Humphrey: That is not a firm scientific opinion, but I think what we have to accept is that sentience, phenomenal consciousness, is a relatively late development in evolution, it does not go all the way back to primitive organisms.

In many ways many people do think, including my great friend and colleague Daniel Dennett, that animals are

sentience all the way down: you get more sentience and less sentience, down to many sentience.

I do not think that is right: I think sentience as a threshold, it comes into being at a certain stage in evolution, because it involves a particular kind of brain mechanism, involving feedback loops. And it only comes into existence when it is needed.

It will not be of any use to an animal which does not think about itself, it does not relate to other animals, which does not have to be, as I have putted, “psychologists”, that is, where it pays off for us.

Now, when I started wondering “well, okay, then when consciousness arises?”, I realized that there is a transition in the evolution of vertebrates which philosophers, and in fact biologists, have not taken very seriously in this respect: it is when animals became warm-blooded.

It was about 200 million years ago that dinosaurs were warm-blooded, and then their related birds and mammals were also warm-blooded. And what that meant was that the whole relationship to the environment had changed: they became autonomous beings, independent of the immediate and physical environment, they could go where they wanted, alive and active day and night, and so on.

And that, I believe, gave them a new strong sense of individuality, of autonomy. They now had the use for

the idea of a Self. But something else happened as well – and this was extraordinary chunks, to put it that way, –our brains had warmed up, the speed of their nerve cells had warmed up.

When you raise temperature for twelve or fifteen degrees to thirty-seven or thirty-eight degrees in humans and mammals, and forty degrees in birds, you triple the speed of nerve conduction. That, suddenly, meant that our brains and our ancestor's brains were working very much faster than that of any other animals.

I believe that produced a crucial reorganization in the brain: it allowed certain forms of feedback to develop which simply would not have taken place otherwise. So, I believe that there is this coming together, both for a different lifestyle and a need for a way of thinking about oneself, because I am now a creature independent of the environment, that went along with this brain which allowed a new kind of psychological picture of what the brain is.

Question: You are familiar with the work of the philosopher Susan Schneider, who proposed a kind of consciousness test for artificial intelligence, allowing us to decide whether or not an artificial machine is conscious. Do you think these suggestions can provide specific details to create such a test for artificial

consciousness? In other words, while nature invented human consciousness, do you think we will be able to create or invent artificial consciousness?

Nicholas Humphrey: Well, those are two different questions. I mean, Susan Schneider suggests some tests which we could use. They are quite good ones, pretty close of the kind of tests I would suggest: except that she does not take that seriously the social side of consciousness. But we need that ability since it allows to get inside the minds of other creatures, like ourselves. I think she should had added that to her list of criteria, then I think she would be getting close to have a diagnostic test, a sentience test in a machine.

About the machines, no questions at this point come down. People have speculated about whether these new language models could be sentient like ChatGPT for example. David Chalmers believes – unbelievably for me – that there is a 10% chance that ChatGPT is already a sentient.

He is a philosopher: I do not think he should do strong statements like that. You cannot be 10% sentient, and there is no reason whatever to think that anything about these language models are sentients since they do not have the need for, and it does not have the mechanics for it either, and it does not show any of the diagnostic criteria.

But that does not mean that we could not develop a machine which does meet these criteria. However, it would only do that if we deliberately design the machine to have that capacity: it is not going to happen by chance, but because we make machines which are processing data faster and faster, or are more and more intelligent.

Intelligence is not the same thing as sentient, and that is why I do not think machines will achieve it until we deliberately introduce it into machines. And, at the moment, we do not know how to do that. What I wrote in my book goes somewhere towards suggesting the kind of thing we would need to build into a machine if machines were to become sentients.

But then we have to ask: why would we do that? One reason might be: because we want to, we want machines to have the same capacities for mind reading for the same sense of earning psychological importance and so on as we do ourselves.

So, we could take the lesson from nature and apply it to machines and maybe develop sentient machines. And maybe one day we will request our sentient machines to do our work from us far beyond the Earth in extraterrestrial space, for example.

Humans are never going to be able to go and live in a far-off galaxy. But machines could do, we could design machines which could certainly get there. But if they are

going to establish that lifestyle, if they are going to be interested enough to begin to develop a science of their new environment and to consider what kind of culture they want to develop, they are going to need to have confidence in their own importance.

So, maybe we need to ask ourselves first if we humans want these machines to have phenomenal consciousness or not.

Question: You also argue that some particular animals are sentient and others are not sentient at all. We know that "sentience" is generally a kind of moral property that any organism needs to possess in order to have moral status and be considered an agent of rights. In light of this, what specific ethical implications do you think your research has, especially concerning animal ethics and animal rights?

Nicholas Humphrey: It certainly means that, if I am right, we should be cautious before we attribute sentience to other creatures. We should not just give them the benefit of the doubt and assume that sentience is going to be present because they have complex nervous systems, or because they show high levels of intelligence, which tends to be what people do at the moment.

Most of the writing about sentience and animals has to do with how clever they are, not with how conscious

they are, in the sense of having a phenomenal consciousness. I discussed this topic a bit in the book, I do not make a lot of it on when and why we should take in consideration on other animals as having moral status because they have consciousness like ours.

My general view is that – I mean as a scientist and not necessarily as a citizen – we must assume that most animals are not: I do not think lobsters are sentient, I do not think octopus are sentient, I certainly do not think that worms are, and that means that we can rethink the kinds of laws which have now been passed around the world about animal sentient.

I do not know what is like in Portugal, but in Britain for example, last year a law was passed saying that, by law, lobsters are sentient. So, now it is illegal to boil a lobster alive. There may be lots of reasons for not wanting to boil a lobster alive, but that being sentient is not one of them in my opinion.

Question: When do you believe that sentience first emerged in human beings? Do you think it initiates its development during pregnancy, possibly even in the mother's womb, or do you lean towards establishing its commencement in early childhood? An answer to this question may carry profound implications for our comprehension of human development and ethical rights concerning life and well-being.

Nicholas Humphrey: Very important and interesting question. I think it's not at the beginning, and one of the reasons for saying that is that when human infants are born, their brain is not myelinated: the myelin sheaths does not cover cells, they have not yet developed, and it certainly means that they are not functional. The visual cortex is not working in a newborn human infant.

Following this, I believe that human infants, if they can see – and they can see, we know that from their behavior – they must have something like blindsight. Then, at that point, they do not have phenomenal experience.

But, again, I am in a minority for saying that. People cannot take on board that it is possible to see without having sensations. And even though they know of these cases, they do not take it seriously in their everyday clinical practice. So, if you see a baby who is quite clearly able to respond to his mother's smile, you will assume the baby is seeing the mother's face in the way in which we would.

I do not think that is true: I think that is about the three or four months, when the brain becomes myelinated, that the circuits would be sufficient to sustain phenomenal vision.

Question: To conclude, Professor António Damásio, the Portuguese neuroscientist who has authored

numerous works on these subjects, discusses various forms of self, including the autobiographical self, which can be perceived as a form of consciousness. However, he also introduces the concepts of the proto-self and the nuclear self as foundations for the autobiographical self. How do you position yourself within this conceptual framework regarding the self?

Nicholas Humphrey: I think me and Damasio agree about a lot of things, we have talked a lot about it in different contexts. However, I want to be much more specific that he does in saying that the foundation of the self is sensations and that is what everything else is built on.

The subject of our mental states is the self, built on sensations. And if we go on and elaborate it, of course, we exist as social selves, as autobiographical selves, and so on. Although I think Antonio is too ready to say "Okay, selves exist in all these different levels, and each of them matters, in different aspects". And I think he avoids the issue of why one should be supremely important over others.

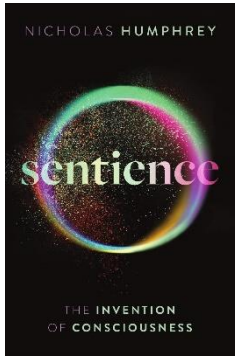
My belief is that if you lose the phenomenal self, or the core self, there is nothing left. You cannot keep an autobiographical self or a social self if you lose phenomenology. And the interesting thing is that in almost all cases of dissociation and of absence, where people's selves break down, they actually do remain

present as having phenomenal selves, centered on pain, colors and lights, and so on.

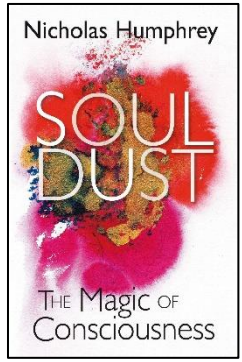
That does not seem to go away: if it did, the person would not exist. In fact, there are cases where patients claim that they do not exist. In Cotard's Syndrome, the patient insists that he or she has died.

When the doctor says "Well, you really have not died, because I am talking to you, and you came to my studio, to my clinic today". And the patient says "No, I have died. I am not there anymore": I think that what they are trying to express may be that the phenomenal self does not exist for them anymore.

Books by Nicholas Humphrey



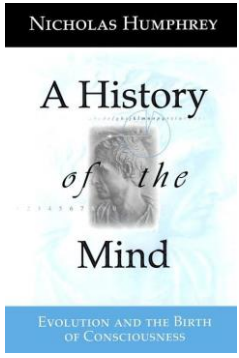
2022



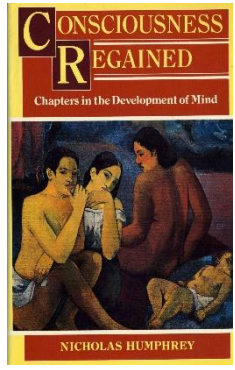
2011

2006

2002



1998



1983

IX. Dialogues with Sir Roger Penrose

Sir Roger Penrose is Emeritus Professor of Mathematics at the University of Oxford and Emeritus Fellow of Wadham College, Oxford, and honorary fellow of St. John's College, Cambridge and University College London.

He was named Knight for his services to Science in 1994 by Queen Elizabeth II.

In 1969, with Stephen Hawking, Penrose proved that all matter inside a black hole collapse into a singularity.

He has received several awards and distinctions, including the 1988 Wolf Prize in Physics and the 2020 Nobel Prize in Physics.

He is the author of several books focused on the nature of consciousness, such as *The Emperor's New Mind* (1989), *Shadows of the Mind* (1994), *The Large, the Small and Human Mind* (1997), *The Road to Reality: A Complete Guide to the Laws of the Universe* (2007) or *Cycles of Time: An Extraordinary New View of the Universe* (2013).

More information: maths.ox.ac.uk/people/roger.penrose

Question: You are one of the most prominent scientists of the 21st century, having been awarded the Nobel Prize in Physics in 2020 for demonstrating that Albert Einstein was, indeed, mistaken in denying the existence of black holes. Your role was to mathematically establish that the existence of black holes is a consequence of Einstein's theory of relativity. In addition to this pursuit, you have also delved into the nature of consciousness. When did you first develop an interest in the topic of consciousness?

Roger Penrose: I can trace it back to when I was a graduate student since one of the key ingredients to the way I think about consciousness occurred because of a course of lectures that I went to in Cambridge when I was a graduate student (this was a long time ago as you can imagine). I was doing Pure Mathematics at the time – Algebraic Geometry – and I attended three courses that had nothing to do with my topic.

One was a course on General Relativity by Hermann Bondi that was very influential on me; then there was a course by the great physicist, Paul Dirac, on Quantum Mechanics, which was also very important to me; and the third one was a lecture by a man named S.W.P Steen, who was a logician, he did mathematical logic and he focused on Turning Machines, and from there I knew what the idea of computability meant in the technical or mathematical sense.

But he also described the two main Gödel's Theorem and I was stunned by that as when I first heard about Gödel's Theorem, it seemed to say that there were things in mathematics you could not prove and I didn't like that idea, so I went to this course to hear about it, and it wasn't quite like that. What it is like is the following: if you have a certain method of proof which you could in principle put on a computer, then it would try to prove or disprove a certain result.

What you can do then is to produce a statement – a mathematical statement – about numbers, the kind of thing that the theorem is meant to be addressing, and it can tell you roughly speaking, it encodes the statement "I'm not provable by these rules". Now, is it true or is it false?

Let's suppose it is false, then it is provable by these rules. You're supposed to believe that anything provable by the rules is true, that's the whole point. You choose the rules, so that anything provable by the rules must be true: if it says "it is false", then it is provable by the rules and therefore it is true. And if it is true, then it is true and not provable by the rules, so that is the conclusion. You see that this statement is definitely true and it is not provable by the rules. But how do you know it is "true"?

Well, you know it by virtue of your faith in the rules: your belief that the rules actually do prove things and if they say "yes, it is true", you believe them, because you

understand them – you go into the rules and you actually understand what the rules mean, and you do not just follow them, and this was key to me. Understanding what it means is stronger than following the rules. Understanding is something which is conscious, it requires our consciousness.

I argued that conscious thinking transcends “following rules” and the conclusion I came to is that we are not computers. Ever since I came to this view that conscious thinking, whatever it is, is not algorithmic: there is something beyond following rules. I tried to write my book “The Emperor's New Mind” to get this point across and also to explain different topics about mathematics and physics, that I was very fond of, and try to get these ideas across, trying to make it not so difficult so people could have some interest in it and they could learn from the book.

I thought that by the time I’d finished the book, I would learn enough about neurophysiology that I would understand what possibly could be non-computable in the action of the brain, but I did not. I came to the conclusion that, even though I understood what the procedures were, I believed at that time that you needed to take into account the collapse of the wave function and that is a very important part of the whole argument: what is it, in physics, that can transcend computation.

You have to find something in the physics, and the physicalists think that what goes on in our brains is part of the physical world, but what part of the physical world? The argument is that it is not the normal kind of physics that we use, but it must be something beyond that. I had no idea what in the brain could possibly do this, so I kind of petered off at the end of the book without knowing it, really.

Question: Can we consider that your thesis about consciousness is a direct consequence of Gödel's Theorem? Given your claim that mathematical understanding is a feature of consciousness and is non-computational, meaning it is not computable by principle or definition.

Roger Penrose: Yes, that is right.

Question: Following this, why, in your view, do many individuals that work in Artificial Intelligence believe that the mind functions akin to a digital computer, and that reverse engineering the human brain is a feasible endeavor? What, in your opinion, motivate these philosophical beliefs about the mind and the brain, particularly among intelligent scholars?

Roger Penrose: I just do not think they have followed my argument. There are various complaints you can

make about my argument, and the strongest one is the following: if you don't know what algorithms we are using in our heads, then you cannot construct Gödel's results and therefore the argument does not work. The counter-argument that I pursue in my other book *Shadows of the Mind* is that if we suppose there is an algorithm in our heads – something that we do not know if it is true – how did such algorithm come about? We have to suppose it came, like everything else, from natural selection. It has to be a selective advantage.

The image I used in the book is of our ancestors doing useful things, like building shelters, domesticating animals, growing crops… and in the foreground, you have somebody having an idea of how to build a mammoth trap and the poor mammoth is going to get caught. So, all these people are doing things which have selective advantages. But in the foreground, you can find this poor mathematician who is working on some theorem, and he is about to be devoured by a saber-tooth tiger.

The moral of this is that doing this kind of sophisticated mathematics has no selective advantage and you can produce several mathematical results. The one that I like the best is called Goodstein's Theorem that I describe in "Shadows of the Mind". It is a wonderful result that you can explain to people who do not know much about mathematics (you just need to know what is raising a number to a power in mathematical

notation). Goodstein's Theorem claims that if you follow these procedures "A" and "B" repetitively, each one a very simple procedure, then this does not go on forever, it comes to an end. The thing is that this requires so many steps that it is completely ridiculous, meaning that it can have no selective advantage whatsoever. Yet, how do we know it is true?

Well, we know it is true because of some wonderful results due to the mathematician Georg Cantor and the wonderful thing is that you don't need to be Cantor to understand it: you only need to understand Cantor's reasoning and that does not require that much: it requires a bit of shifting in one's normal point of view, but the actual reasoning is not that hard, so you can understand why it is true, why this result is true. But how do we know that? How could we have evolved if we were just algorithms? There is just no way that an algorithm with that sophistication could have come about.

That is: how would the general quality of understanding, whatever that is – and I do not claim to know what that is – but what I claim to argue is that, whatever understanding is, it must be a feature of consciousness. Of course, I am not saying that that is the whole of consciousness since consciousness involves many other things, such as the perception of the color blue, for example, or the feeling of pain. There

are many things in consciousness that have nothing to do with mathematics directly.

My argument is that if you can see these results in mathematics, they must be beyond computation. I mean, you cannot say that the Goodstein Theorem is beyond computation, because once you know the Cantor input, you could put a particular version of it on a computer too. But that is not believable from the natural selection point of view: there is absolutely no selective advantage to that at all – it is much too outside the normal use of our brains. Whereas the general quality of understanding something is hugely important.

What I'm trying to argue is the general quality of understanding things is not algorithmic: I'm not talking, here, about what it is like to see a blue color. My main point is that whatever consciousness is, it is something beyond the computational procedure.

Question: Can you elaborate on why you disagree with the prevailing view in current neuroscience, which posits that consciousness is an emergent property of the brain, arising from neuronal activity and information exchange between neurons?

Roger Penrose: Yes, I do not think there is some mysterious "other thing" which comes floating into our heads, no. I think it has to do with neurophysiology

which is going on inside the head: we just do not know what that is yet.

Question: Concerning the nature of consciousness, do you hold the belief that neuroscience alone can solve this problem, or do you argue for the necessity of what you refer to as "a new physics" to integrate this unique phenomenon into the scientific view of the world?

Roger Penrose: I think we need a new physics. And the other parts of the argument are: “what are the other parts of physics that we know?” You can go through one of the major theories that we have, Newtonian Physics, and you could put that on a computer, as people do. What about the Theory of General Relativity by Einstein?

People now have worked out how black holes spiral into each other, what kind of signal comes out – clearly, this is a very computational procedure. We need to work out what systems can be put on a computer: I would say it is not general relativity.

How about Quantum Mechanics? Well, in quantum mechanics there is the Schrödinger Equation that can be put on a computer too. But then we go back to one of the courses I attended when I was a graduate student by Paul Dirac, and in his first lecture, he talked about the superposition principle: the idea that an atom could be “here” or “there” and in quantum mechanics you have states where the atom is “here” and “there” at the

same time: this is the fundamental principle of quantum mechanics.

And then he took out a piece of chalk and he described breaking it in two pieces, trying to describe how the pieces of chalk might be here and there at the same time, and the key thing about this was that I was present at the lecture, but my mind was wandering: I was looking out of the window, thinking about something completely different and so when my mind returned to the lecture, he had moved on to something else and I remember him saying something about the energy involved in the piece of chalk, but I could not understand what that had to do with anything relevant.

So, I missed the answer, which was just as well, because, probably, he was trying to calm us down in some way, to not worry about this problem, because energies are so big, so it doesn't come in at this level. But I was left with the feeling that we needed something new, we needed a new physics. And this is the place where we need new physics since the Schrödinger Equation does not describe the world, as all physicists know, but they sort of forget. They know it, but they do not say it out loud.

Schrödinger was very much aware of it, because he describes people after knowing he put this poor cat in the box, which he puts into a state of being alive and dead at the same time, and people misunderstand what Schrödinger was trying to do, I think. They often say

"well, if you had a sophisticated enough experiment, you could make a cat be alive and dead at the same time, we are just not quite there yet". That is not what Schrödinger was trying to say.

What he was trying to say was that this was ridiculous, you cannot have a cat that is alive and dead at the same time, there is something seriously wrong with his own equation. He was actually trying to argue against his own equation: he was saying that the Schrödinger Equation does not explain how the world operates.

Question: Could you expand on the idea that there might be a hiatus or gap in quantum mechanics, suggesting that something is missing for it to be considered a complete theory? In which ways do you envision this incompleteness, and how it might be related to our understanding of consciousness?

Roger Penrose: Yes, a huge gap. But it is sort of an independent gap, that is a physics argument. But I'm saying is that there is a gap there and maybe that is where the non-computability lies. When I wrote *The Emperor's New Mind*, that was the pitch I was trying to get across. The weakness of the book was that I had no idea how neurons could do this.

I thought that by the end of the book I would have learnt enough about neurophysiology to see where there could be a gap of this kind, and I did not. I sort of gave

up on the book and I had to finish it somehow and I was not very happy with all that I said, but, nevertheless, that was the end.

But then, Stuart Hammeroff, who read my book, wrote to me and told me about certain structures that I did not know about, which are these microtubules, and he had a theory, which has now gained a lot of support from other angles, that general anesthetics can raise a different angle on consciousness.

Hammeroff's angle was the following: what is it that turns consciousness off, particularly, what turns it off in a reversible way, a very specific way – you can turn it off and turn it on back again (meaning: you can be in an induced coma with no consciousness, and then awake and have your consciousness back).

Being an anesthesiologist, Hammeroff is not only trying to put patients into unconsciousness and then wake them up again, but he is also trying to understand what is actually happening with the substances that 'turn off' consciousness. This is very interesting because the conclusion is that it is not a chemical process: what links these different substances together is related to a physical process, rather than a chemical one. And the question is, what is going on?

What is interesting is that, although we got together in the 1990s, this was regarded as a far-out point of view, but it is now becoming one of the views that people take

seriously outside the three or four seriously considered viewpoints about consciousness, this is now one of them. I found that is actually quite reassuring, that people are taking it somewhat seriously and there are very interesting experiments on the general action of anesthetics and what structures they affect.

It does look like microtubules could as well be an important ingredient to this whole story of consciousness: things are moving in a direction which may open it up in this way to see if this is a correct point of view or not.

Question: Let us delve a bit deeper into your theoretical model of consciousness. Essentially, you criticize the so-called “Copenhagen interpretation” of quantum mechanics, as you argue that what actually happens is contrary to what this interpretation describes: we have the collapse of the wave function, but we do not need any conscious subject for this to happen, we don't need an observer. Following this, how can consciousness arise from this process of the collapse of the wave function through the influence of gravitational forces, which is what you propose?

Roger Penrose: Yes, as you said, it is the other way around. There are several ingredients: one that I did not mention is that the collapse of the wave function, in my view, is do with the combination of gravity and quantum

mechanics. This is not yet experimentally confirmed or disproved, but there are experiments aimed at resolving this issue. There are theoretical reasons arguing for this point of view. I have an argument, which I have put forward a few decades ago, that tries to show that there is an incompatibility between the fundamental principles of quantum mechanics, which involves the principle of superposition that I mentioned, and the principle of general relativity, which is the Galileo Principle, that says you can cancel gravity by free-fall.

Galileo imagined that if a big rock and a little rock falling from the leaning tower of Pisa, with no atmosphere, they would fall together and if there was an insect sitting on one of the rocks looking at the other one, it would think that there was no gravity. So, you can cancel gravity by free-falling and that is the principle of equivalence, which is the foundation stone of Einstein's General Relativity, but that foundation stone is inconsistent with the foundation stones that are in quantum mechanics, which is the superposition principle.

My argument is that there is a conflict there, which will need a new theory. I do not know what the theory is, all I can say is an estimate of when this theory comes in, in what sort of scale you would start to see the collapse of the wave function being a physical process that you can actually measure. As you said, it is the opposite of what

many people used to think in the early days of quantum mechanics, namely that the collapse of the wave function, which is when you make a measurement, depends on the observer, after all, making a measurement is making an observation, depends of the observer, a conscious being.

People like John von Neumann and Eugene Wigner argued for this kind of view. I actually talked to Wigner about this on some occasion and I found out that he was not so dogmatic about this as some people would think: he was certainly open to other alternatives. He seemed to think that this was a point of view that should be taken seriously, but only up to a point because you can see it does not really hang together and that you cannot rely on conscious beings to collapse the wave function. I think it has to be a physical process, but as you said, the other way around is the brain making use of that physical process in the production of consciousness.

Question: How close do you believe we are to establishing a genuine connection between states of phenomenal consciousness, such as emotions or feelings, and the physiological reduction of quantum superposition states in the brain, particularly within microtubules?

Roger Penrose: There surely is a long way to go. I do not see that directly. I would say, although, I'm quite impressed by how far things have gone, I was not expecting even the progress that has been made. I was at a conference recently, in Canada, where there were many interesting talks about these ideas in serious ways and with experiments. I was impressed with how far it has moved, but, nevertheless, it has not moved very far in the directions of actual consciousness.

The experiments on microtubules or on tubulin proteins or on nerve propagation and things like this, one is beginning to understand things which were not known previously, and I think that there are very interesting and unexpected things to learn from these experiments.

Most particularly, I know this is not part of your question, but there are some puzzles about how people can react so quickly and when you have a particular game like ping-pong or tennis, but specially ping-pong where you have to react so quickly – I used to play that game myself – you can see that conscious decisions could influence what you did much more quickly than they should, because when you work out where these nerve signals were and how long they take and what part of the brain had to be involved and so on, it looks as though there is no way that this could act that quickly, so it raises lots of very curious problems.

Many people say that those things that you think you know what you are doing are all done unconsciously, so it is not really telling you anything about consciousness. I do not believe in this, since when I used to play ping-pong, I thought I was really deciding whether to flick this down one way or the other way, and I was making that decision very rapidly.

The reduction of the quantum state is a very peculiar phenomenon when you try to fit it together with other principles of general relativity and quantum mechanics, and when you try to fit these things together, you come up with very strange relationships in relation to the way time progresses. They are not inconsistent, but they are very peculiar, and it tells you that the time we think perceiving things are happening is not quite what we think.

There were some early experiments, which I did actually describe in the *Emperor's New Mind*, done by Benjamin Libet, where he had a patient in a brain operation (I won't go into details of it), but it was very hard to explain the temporal line, what happened before and what happened after: it did not make any sense.

The argument is that there is something very peculiar going on in relation to the temporal nature of experience, when you think you experience something, and when you actually experience something, and to make sense of it involves some very curious things,

which, I think, will be very intriguing in any theory which we come to later, but we are a long way from it.

Question: To conclude, when considering the phenomenon of spirituality, numerous books explore themes like quantum synchronicity, chakras, chi, among others dubious concepts. If you could provide a brief comment, what are your thoughts on these approaches and the popularity of such best-selling books? Do you see any elements that could lend some validity to these alternative phenomena?

Roger Penrose: I am not a follower of any of these things: I have to confess not to having read much in this direction, my bias is not to think there is much in it. We may learn some or other thing about consciousness by looking at these ideas, but I have only concentrated on a very limited aspect of consciousness, which is this quality of understanding, specifically, when it comes to mathematical ideas.

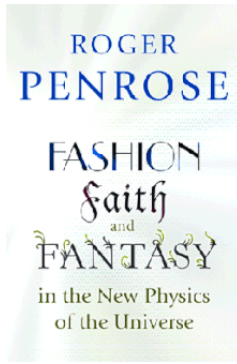
There is a lot more to understanding it, a lot more to conscious experience than the small area on which I have concentrated.

I have stayed away from these other things, mainly because I do not see any way to talk in a nice, precise way about it. I think this is the problem, I'm a mathematician: I like to be able to put things together into a theory where I can see consequences coming out,

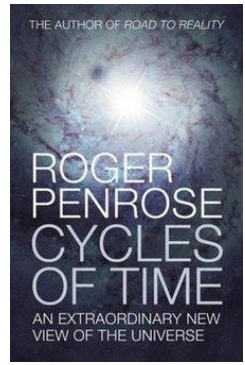
which are precisely determined in clear ways from the theory.

These things that you talked about, I do not see them in this form, so I have not studied them. I do not see a benefit to me, although there can be benefits to other people.

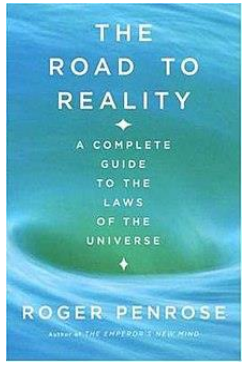
Books by Sir Roger Penrose



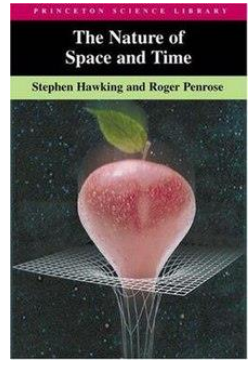
2016



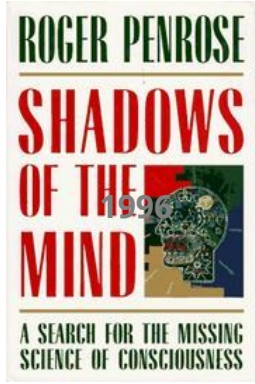
2010



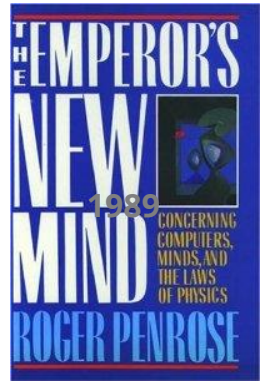
2004



1996



1996



1989

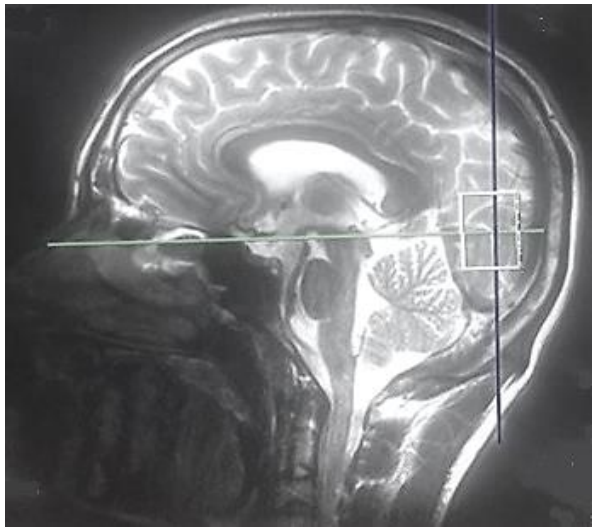
PART II

Brain

PART II

Brain

I. INTRODUCTION: BRAIN



1. fMRI image of a brain.

The philosophical and scientific nature of the brain, as depicted in the image above, has been a subject of debate for hundreds of years across various disciplines. We now have significantly more knowledge about the brain than our grandparents did. For instance, we understand that, on average, it weighs 1.5 kg.

Additionally, we know that it consists of various types of neurons that exchange electrochemical information among them. Furthermore, these neurons can form

distinct neural networks that play a role in different mental functions, such as long-term memory, emotions, or creativity.

The human brain comprises approximately 86 billion neurons interconnected through trillions of synapses. It is organized into various distinct regions. The most primitive nucleus, reminiscent of ancient evolutionary structures, governs basic survival instincts. Next, there is the limbic system, responsible for emotions and memory formation. Finally, the brain's outer layer, known as the cerebral cortex, is divided into several lobes, overseeing complex cognitive processes.

While our understanding of the brain has significantly advanced, we still have much to learn, as explored in this second part of the book. Throughout history, as humanity delved into philosophical contemplation of its nature as a being composed of flesh and bone, the brain has been a subject of controversy. Notably, two prominent thinkers offered insights on the connection between the brain and the human mind.

In a fervent debate, Aristotle of Stagira argued, in book III of his *On the Parts of Animals*,[50] that the heart was the seat and source of sensations, consciousness, intellect and everything that was relevant in the human being, believing that the brain had a single function: to cool down the temperature of the blood. Against this

[50] Online version: penelope.uchicago.edu/aristotle/parts3.html

perspective, Hippocrates of Kos claimed in his book *On the Sacred Disease*[51] that the Aristotelian perspective was mistaken: the brain should be considered as the source of pleasure, joy, sadness or pain, instead of the heart.

Although Ancient Greece is often considered a pivotal historical era, there is speculation that the earliest contemplations about the human brain did not commence with the ancient Greeks, as commonly believed, but even earlier in history.

At the end of the Stone Age, a period that began about 50,000 years ago in Europe and progressed into the Neolithic period – an era that spans from 3,000 to 2,000 BC – humanity developed some agricultural practices and domesticated certain animals.

During this period, the invention of polished stone tools, the creation of pottery and the formation of small settlements also emerged, which allowed something absolutely fascinating to happen.

From the same period, fossil records were found of the first perforations of the human skull carried out by surgical procedures with specialized stone tools, which indicates that Neolithic man used various types of sharp instruments to carry out an operation known as "trepanation" (from the Greek *trypanon* which means

[51] Online version: classics.mit.edu/Hippocrates/sacred.html

"to pierce"), which is the first known surgical practice to be performed by humanity.[52]

The first skull recognized as perforated was discovered accidentally in 1864 because of... bird feces. What would be the connection between skulls and feces? In 1861, the American Civil War broke out and it became essential to guarantee fertilization for agricultural production, becoming one of Abraham Lincoln's priorities.

Interestingly, the best fertilizers in the world, at the time, came from South America, with the extraction of a substance called "guano" that comes from the feces of various animals, being rich in phosphorus and nitrogen. Thus, Ephraim G. Squier, an archaeologist and journalist, is sent by Lincoln to South America – specifically, to Peru – to guarantee the production and shipment of fertilizer to the United States.

After completing his task, Squier decided to explore the country, having arrived at the Inca cemetery in the Yucay Valley, where he found a perforated skull that had a 15x17mm rectangular hole in its structure, something that left him astonished, given that it is not It is common to find right angles in nature. With this discovery, Squier decided to return home, and

[52] Based on: Weber, J. e Wahl, J. (2006) "Neurosurgical aspects of trepanations from Neolithic times", *International Journal of Osteoarchaeology*, 16: 536–545.

presented the artifact for the first time to the New York Academy of Medicine in 1865.

Regrettably, the discovery was not taken seriously, as the scientists in the organization doubted the possibility of indigenous peoples possessing advanced surgical knowledge, reflecting a clear racist bias.

Undeterred, Squier continued his analysis of the skull and decided to send it to Paris. There, it would be examined by the eminent founder of the first Anthropological Society in France, the renowned doctor and professor, Paul Broca, considered the greatest scientific authority on the study of the brain at that time.

Broca was completely surprised when he analyzed the skull in detail and realized that the perforation contained therein had clearly been produced by a specialized cutting tool and involved some sort of protosurgery, something that was considered completely unfeasible until then.

This was the first of many perforated skulls that were discovered in the following decades, most of which revealed an even more astonishing detail: evidence of bone growth around the perforations. Now, what does that mean? It means only one thing: that individuals subjected to these primitive surgeries may have survived several months or even years, which is extraordinary.

Another historical source can be found in the Nile region of Ancient Egypt. The Great Pyramids and the various funerary procedures were indicators that the afterlife was something of great relevance to the ancient Egyptians.

Now, to guarantee a successful passage to the “other world”, the Egyptians believed that, to ensure that the soul reached the right place, the body would also have to be preserved. It is this belief that explains the entire mummification process that became famous in several literary and cinematic works.

Interestingly, the most important part of the body was the heart, as it was believed to represent the person's self. The intestines, lungs, liver and stomach were also considered important – being embalmed and stored in canonical jars alongside the mummy.

On the opposite side of relevance was .... the brain, which was simply thrown away and extracted through the nostrils using an iron hook. As far as we know, the ancient Egyptians gave little importance to the brain in its relationship with the mind or consciousness.

Through various Egyptian writings – particularly those that adorn tomb walls – we know, however, that the heart was not only seen as the repository of the soul's earthly actions, but also possessed cognitive and conative capabilities (emotions and feelings).

This primacy of the heart was maintained, remarkably, until biblical times. In fact, when the Old Testament was translated into Greek, it was accepted that Man's intellect and emotions resided in the heart and not in the brain, and there is not a single reference to the brain in the Bible.[53]

However, despite this discredit for the brain, we can possibly find in the ancient Egyptians the oldest written reference to this particular organ: a medical papyrus known as the "Edwin Smith papyrus" that describes 48 types of injuries to the head and neck, along with advice on treatment and surgical intervention, and which contains references to texts written up to 3,000 BC.

The authors of the papyrus seem to have had a certain understanding of the brain's function. They described specific head injuries and correlated them with various symptoms, such as paralysis or loss of speech. However, they also made gross errors, such as stating that a lesion in the right hemisphere would cause injuries to the right side of the body, when, in fact, the left side should be affected.

Another fascinating historical reference about the brain can be found in the adventures of Achilles and Ulysses, namely in the Homeric texts of the *Iliad* and the *Odyssey*,

[53] Based on: York, G.K. e Steinberg, D.A. (2010) "Neurology in Ancient Egypt" In Finger, S., Boller, F. e Tyler, L. (eds) *Handbook of Clinical Neurology*, vol. 95, Elsevier: Amsterdam.

which inspire the title of this book. According to the Greeks, the self would be composed of several forces, one of them being the *psyche*, the vital force that would keep the person alive. Homer would end up identifying the mental in several other “forms” of the soul that resided in the chest.

The most important of these forms would be the *thymos*, mentioned more than 450 times in the *Iliad*, which would be located in the diaphragm, being the source of emotions that drive someone to act. In terms of intellectual capabilities, they would be part of *noos*, situated in the chest. Similar to the Egyptians, Homeric texts do not attribute great importance to the brain.

There is, however, a very relevant linguistic influence from Homer, who introduced three fundamental concepts into the vocabulary: (1) *enkephalos*, the brain itself; (2) *muelos*, referring to the spinal column (from which we get the word 'marrow'); (3) *sinew*, which gave rise to the concept of “neuron”.[54]

Despite this disregard for the brain, we can find in Ancient Greece some thinkers who, contrary to this belief shared by the Egyptians, argued that this particular organ should have greater relevance than it had been given until then.

[54] Based on: Singer, C. (1957) *A Short History of Anatomy and Physiology from the Greeks to Harvey*, Dover: New York.

One of these thinkers was Alcmaeon of Croton, who considered the brain as the organ of sensations, although all his texts have been lost and no direct references have survived to this day. Despite this, several indirect references have come to us that point to the proto-anatomical brilliance of this philosopher.

For example, Theophrastus of Eressos stated that Alcmaeon would have been the first to study the anatomy of various animals through dissection techniques, having discovered that there were two channels that would physically connect the back of the eye to the brain, what we currently recognize as the optic nerves, leading to the argument that the senses were all connected to the brain.

Another indirect reference can be found in Aetius, a contemporary of Alcmaeon, who argued that the latter was known for claiming that intelligence, like sensations, would be related to the human brain. Due to these and other discoveries, the impact of Alcmaeon's investigations should be comparable to the discoveries of Copernicus and Darwin.[55]

As we saw previously, the other great Greek reference on the relevance of the brain is the founder and father of Medicine, Hippocrates. In the *Corpus Hippocraticum*, there are multiple references to the brain, with the

[55] Based on: Mithen, S. (1996) The *Prehistory of the Mind*, Phoenix Books: Guernsey.

most relevant text being *On the Sacred Disease*, which largely focuses on an attempt to understand the nature of epilepsy.

Hippocrates is the first that tries to determine that this disease could not be explained by the current medical theory of his time, the so-called "Theory of Demonic Possession", where it was argued that epilepsy was linked to a form of possession by a particular demon as divine punishment for some sin committed by that person.

The founder of medicine tried to counter this theory, arguing that it was a notion defended by charlatans and healers who did not really want to know about a treatment for this disease. He asserted that epilepsy was a brain disorder caused by an excess of phlegm, obstructing the flow of air in the blood vessels. According to his argument, only an epileptic seizure could rectify this blockage.

In addition to this naturalistic explanation (which does not appeal to obscure demons!), Hippocrates would end up arguing that the brain is responsible for all our mental activity. It is worth paying attention to the words of the philosopher and doctor:

> "It ought to be generally known that the source of our pleasure, merriment, laughter and amusement, as of our grief, pain, anxiety and tears, is none other than the brain. It is specially the organ that enables us to think, see and hear,

> and to distinguish the ugly and the beautiful, the bad and the good, pleasant and unpleasant. [...] It is the brain too which is the seat of madness and delirium, of the fears and frights which assail us, often by night, but sometimes even by day; it is there where lies the cause of insomnia and sleep-walking, of thoughts that will not come, forgotten duties and eccentricities..."[56]

This remarkable account from 400 years BC could have been taken from any neuroscience book from this century, indicating the brain's involvement in sensory perception, judgment and emotion, as well as its association with mental disorders.

In the aforementioned book, Hippocrates also reveals his incredible anatomical skills, where he provided several anatomical descriptions of the human brain, arguing that it is similar to many animals, and that it is divided by a structure that splits it into two halves, a clear reference to the body callosum and both hemispheres.

Plato also sought to understand the role of the brain in humans, associating it with the concept of the soul. In Platonic philosophy, "soul" is distinguished from its usual religious meaning, and was composed by a tripartite structure: the *epithymetikon*, the *thymos* and

[56] Original publication: Chadwick, J. e Mann, N. (eds.) (1983) *Hippocratic Writings*, Penguin: London.

the *logistikon*. The first was associated with the liver and intestine, where the individual's basic vegetative needs would be located, while the *thymos* was located in the heart, which instigated emotions such as anger, fear, pride and courage.

Plato believed that these two distinct parts of the soul were also present in other animals and would cease to exist at the moment of death. However, the *logistikon* differed from these two initial forms: it was considered a unique spiritual force in humans, granting individuals thought and intelligence, being immortal, and capable of reincarnation.

Now, the most interesting part of this tripartite distinction is that Plato believed this last part resided in the brain. For example, in the book *Phaedo*, Plato states that: "the brain is the originating power of the perceptions of hearing, sight and smell, and memory and opinion can come from it". In the book *Timaeus*, Plato writes: "the head... is the most divine part and dominates the rest of the body."[57]

Against this Platonic perspective, we can find Aristotle, as mentioned briefly earlier. Aristotelian thought did not attribute much importance to the brain. Interestingly, this conclusion was drawn from (very

[57] Based on: Crivellato, E. e Ribatti, D. (2007) "Soul, mind, brain: Greek philosophy and the birth of neuroscience", *Brain Research Bulletin*, 71: 327–336.

rudimentary) empirical data and observations, but it was nonetheless misinterpreted by the Stagirite.

Aristotle was fascinated by the functioning of the human body and was among the first to dissect a multitude of animals at various stages of development, including fish, reptiles, mammals, and even elephants.

He made several anatomical descriptions, observing two membranes that covered the brain, which we now know to be the meninges (dura mater and pia mater). He also observed a structure located in the "back" of the brain that he called *parencephalis*, which had a very different nature, both in texture and appearance: we now know this to be a reference to the cerebellum.

Despite these precise descriptions, Aristotle did not attribute any relevant role to the brain: he was not enchanted by its uniform and cold structure, having relegated greater importance to the heart, taking into account that it was hot and irregular. It's important to note that one of the fundamental beliefs of that time was linked to the thesis that heat was essential for life, as living bodies were considered hot, while corpses were cold.

Aristotle supposedly examined various embryonic stages of the hen's egg and concluded that the first organ to emerge was the heart. The brain was then considered a kind of biological cooling tool due to another empirical observation: the fact that heat rises.

This led the author of *Metaphysics* to deduce that the numerous blood vessels covering the brain served to cool the blood in the heart. This also provided an explanation regarding the size of the human brain: since human beings were warmer than other mammals and animals in general, they needed a larger "device" to lower blood temperature.[58]

A few decades after these rudimentary experiments by Aristotle, another philosopher named Herophilus emerged in the city of Alexandria, making admirable contributions to brain anatomy. He recognized, for example, very specific connections between different parts of the body and the brain through the spinal cord, which we now call cranial nerves. He described seven pairs and formulated their origins: the facial, auditory, optic, hypoglossal, trigeminal, and oculomotor.

Thus, Herophilus contributed to the thesis that it was the brain – and not the heart! – responsible for mental functions in general. For his truly remarkable contributions, he became known as the first great anatomist, having founded the Alexandria School of Medicine, where he conducted much of his research on cadavers.

After sailing through Ancient Egypt and Ancient Greece, we now arrive at the next stop on this historical journey:

[58] Based on: Gross, C.G. (1995) "Aristotle on the Brain", *The Neuroscientist*, 1 (4): 245–250.

the time of the Colosseum in Rome and the Roman Empire, where we encounter Galen of Pergamum. He is considered the founder of physiology, and unlike Herophilus, he was prohibited from dissecting human corpses. Instead, he used several animals to draw his conclusions.

Galen achieved the remarkable achievement of differentiating a dual constitution of the nervous system. On one hand, it consisted of nerve pathways entering and leaving the brain through the base of the skull. On the other hand, there was a set of nerve pathways connected through the spinal cord. Admirably, in his work *On Anatomical Procedures*, Galen describes 10 of the 12 pairs of cranial nerves present in each human being, although he may have confused some of them with each other.

These discoveries – and other discoveries throughout the medieval period – paved the way for Modernity. With Descartes, a new perspective emerged, dividing the world into two fundamental substances: the material part, representing the majority of existing entities, and the immaterial part, composed of the human soul or mind. This conceptual framework is termed "substance dualism" as it delineates a specific duality within the world.

In Descartes' conception, the human being is dual in the sense that the body and the brain are composed of something material and physical, while the soul or mind

is immaterial and non-physical. This dualistic framework suggests a separation between the physical and non-physical aspects of human existence. What follows from this dualistic scheme?

This dualistic conception of the human being have a problem: the challenge of explaining the interaction between two substances of different natures. If the mind is immaterial, how could it influence something material like the body or the brain?

Descartes faced this problem and, following the footsteps of the early anatomists, produced a very rudimentary study of the human brain. He observed that, anatomically, most brain structures existed in pairs (e.g., right and left hemisphere). However, one particular structure seemed to be singular: the pineal gland.

Based on this discovery, Descartes contended that the interaction would occur in this specific organ: the pineal gland ensures the causal interaction between the mental (immaterial) and the physical (material).

The author of *Principles of Philosophy* was, in addition to being a philosopher and mathematician, tutor to Princess Elizabeth of Bohemia, to whom Descartes would end up writing the dedication of his book. Remarkably, Elizabeth posed one of the most formidable objections to this dualistic perspective: how can something immaterial (like the soul) be causally

influenced by something material (the body and the brain)?

How could a mental belief, for example, "I want to stop reading this book," result in the physical, bodily act of closing the book? This objection was so powerful that it prompted numerous philosophers to propose alternative solutions to address the issue.

One way to solve this dilemma is to acknowledge the objection and posit that, in reality, there is no interaction between the physical and the mental. French philosopher Nicholas Malebranche argued in this manner, suggesting that the mental and the physical are connected through an occasional relationship — this stance is termed "occasionalism." Gottfried W. Leibniz also argued along similar lines but with a distinction from Malebranche: he proposed that the mind-body relationship occurs in a state of "parallelism."

The issue with these two responses to Elisabeth's objection is that they both assume the existence of a God to substantiate the argument. Occasionalism posits that God is consistently orchestrating the mental and physical aspects in each moment: when the reader forms the intention to close the book, God intervenes and ensures that the body carries out that action.

On the other hand, parallelism argues that God does not act at every moment, but that he created, at the

beginning of the universe, a perfect synchronization between the mental and the physical so that it appears that there is a causality between them, as if they were two perfectly synchronized clocks in a pre-established harmony.

Thus, whenever the reader forms the belief "I'm really going to close this book!", your body closes the book, and this happens without causality, but by mere synchronization: imagine that the mental and the physical lines are two parallel lines that never touch, but are perfectly synchronized with each other.[59]

Now, these problems with dualism have raised suspicions, leading philosophers and scientists to consider another position on the mental and the physical: the monist thesis. Monism argues that physical and the mental are constituted by one and the same substance – the material.

This new position gives rise to the possibility of studying and investigating the nature of the brain and mind from a scientific point of view, based on rigorous and informed models and observations, which led to the founding of a new discipline: neuroscience.

The first steps of this discipline began in 1810 with the Phrenology project by Franz-Joseph Gall and J. G. Spurzheim. Phrenology attempted to locate various

[59] Based on: Gouveia, S. (2018) *Philosophical Reflections: Art, Mind and Justice*, Braga: Editora Húmus.

functions, from language to perception or consciousness, through the shapes and irregularities of the human skull.

However, this "localizationist" project was criticized shortly afterward by the physiologist Jean-Pierre Flourens, who, from France, rejected the fundamental idea that certain functions were limited to specific regions of the brain. Hughlings Jackson, a neurologist in England, argued that, based on his work with epileptic patients, different functions were responsible for different regions of the human brain.

In 1861, the famous case reported by Paul-Pierre Broca seemed to corroborate the thesis that a lesion in a specific area of the brain – in the left frontal lobe, which would later become known as “Broca's Area” – was related to a particular injury called aphasia, linked to the inability to understand and express language.

In Germany, Karl Wernicke specialized in this topic, receiving his doctorate in 1876 with a thesis focused on a victim of a stroke that had affected another region of the brain relevant to language, which would come to be known as the “Wernicke’s area.”

Soon after, in Italy, Camillo Golgi made a significant contribution to the field of neuroscience with his development of the Golgi stain, a groundbreaking technique that revolutionized the study of neurons.

Golgi's method involved infusing brain tissue with silver chromate, allowing for the visualization of individual neurons by staining a limited number of cells in their entirety. This breakthrough provided a new perspective on the intricate structure of neurons, including their dendrites, axons, and cell bodies.

Using the same method, Santiago Ramón y Cajal, in Spain, contributed with the idea that neurons were unitary structures that transmitted information only from dendrites to axons, contrary to what was thought until then. These – among many others – discoveries have led philosophers and scientists to formulate an increasingly precise link between mental aspects and specific parts of the brain.

The study of neurological diseases has also played a crucial role in the advancement of neuroscience. Erwin Strauss, Kurt Goldstein and Hartmut Kuhlenbeck dedicated themselves to analyzing the "abnormal" brains of several soldiers from the First World War. Initially, they focused on the specific location of the injuries and examined the impact on behavior and functions associated with these injuries.

Notably, in a unique approach, they interviewed patients to understand how their injuries influenced their bodily experiences and relationships with the world. This method established an indirect connection between phenomenal descriptions of consciousness

and the neuronal characteristics of the brain for the first time.

Advances in the development of specific technologies to explore the brain more directly were also fundamental for gaining several fundamental insights into the nervous system. For instance, the investigations of the German scientist Hans Berger made it possible to record, for the first time, the electrical activity in the skulls of human beings.

This technique would be improved over several years until, in the 1930s, Berger developed the technique of electroencephalography (better known as EEG), which allowed the identification of several particular brain rhythms.

For example, less frequently, we find the delta wave, which is in the range of 0.5-4 Hz, and is associated with stages of deep sleep. More frequently, the gamma wave, which is in the range of 30-100 Hz. Hz, associated with higher processes, such as memory, perception or learning.

Finally, other technologies relevant to the study of the brain have been developed in recent decades, including positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). These technologies do not directly measure electrical activity, but instead assess biochemical or metabolic activity in neurons.

For example, using fMRI, we can quantify with some precision the amount of oxygen molecules consumed by neurons, and based on this data, infer the level of their electrical activity. On the other hand, if we employ PET, we use radioactive substances to trace specific receptors or different biochemical balances in neurons.

The fundamental distinction between EEG and fMRI/PET technologies lies in the kinds of resolutions that can be achieved: in the case of EEG, we are able to achieve high-resolution time measurements on the order of milliseconds, but with low spatial resolution.

With fMRI/PET, we achieve high spatial resolution but with a low temporal resolution in the order of seconds. Due to this distinction, many current studies in neuroscience aim to combine both EEG and fMRI/PET to obtain a more comprehensive and accurate understanding of the data.[60]

After this brief historical odyssey about the brain and philosophical reflections on its nature, composition, and relevance, we will, in this second part, introduce some stimulating approaches to the nature of the brain. We'll delve into discussions with four internationally renowned neuroscientists – Anil Seth, Karl Friston, Christof Koch, and Joseph LeDoux – with whom I had the privilege of discussing these issues.

---

[60] Based on: Gouveia, S. (2022) *Philosophy and Neuroscience: a Methodological Analysis*, New York: Palgrave Macmillan.

Let us therefore begin this second part – dedicated to the brain – by introducing a more contemporary view of this important organ that argues that, contrary to what we thought until then, the brain can be seen as a "prediction machine" whose function is to create hypotheses about the world.

II. THE PREDICTIVE BRAIN

The first dialogue of this second part, dedicated to the brain, features the neuroscientist who popularized the idea that views the brain as a prediction machine. He has one of the most viewed TEDx Talks in the world, with 10 million views and the suggestive title 'Your Brain Hallucinates Its Conscious Reality.[61]

We are referring to Anil Seth, a Professor of Cognitive and Computational Neuroscience at the University of Sussex in England. Additionally, he serves as the co-director of the Sackler Center for Consciousness Science. Seth advocates for a theory known as 'Predictive Processing' (PP).

This approach responds to another theory that assigns to the brain a purely passive role – that of receiving stimuli from the external environment (reality) and, based on these stimuli, constructing a map or a mental model of that reality. In this perspective, the active role of the brain is deemed irrelevant.

Hence, this new theory criticizes the passive stance and posits that the brain functions as a highly active prediction mechanism. Its role is to anticipate stimuli

[61] YouTube: https://www.youtube.com/watch?v=lyu7v7nWzfo.

from the world, combining them to formulate the optimal 'guess' regarding the possible causes of these stimuli. Consider your visual perception. The previous approach sees perception as a mechanism in which the reader receives visual stimuli from the words in this book. It was only after receiving these signals that your brain constructed a mental representation of the book.

Now, Predictive Processing argues that this interpretation is incorrect: perception possesses a constructive nature engaged in a continuous "dance" with signals received from the world.

According to this theory, perception can be seen as a kind of "controlled hallucination" aiming to predict a model of the visual signals the reader will encounter.

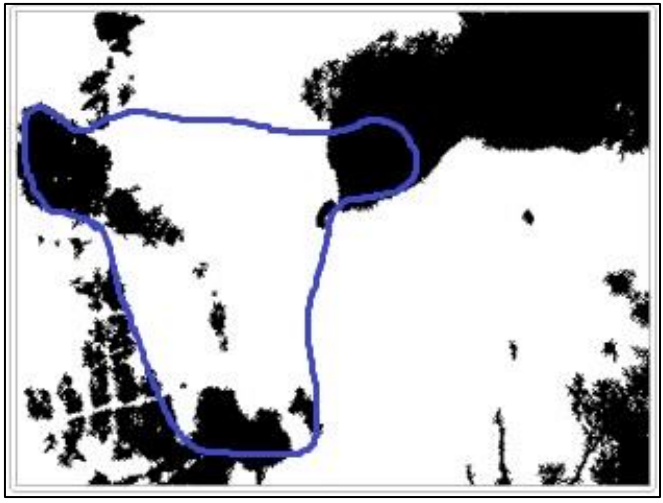
Only after this prediction does the brain confirm whether these signals are accurate or if they need to be updated by additional signals.

If this explanation seems a bit confusing, I encourage you to direct your attention to the following image:

2. Illustration of Predictive Processing (visual perception) I

If you never saw this image, you might perceive it as making little sense being, at best, a piece of dubious aesthetic taste that could have been produced by a Pollockian baby: there is nothing that exists in the world that we can identify with those shapes and shadows.

However, even though your brain may struggle to identify something specific, the approach proposed by Anil Seth becomes more plausible when you now observe the second image, which has a slight difference:



3. Illustration of Predictive Processing (visual perception) II

In this second image, by adding a line to outline part of the shapes and shadows that previously held no meaning, your brain is now able to understand what is present: the second image facilitates the brain in updating the prior model and forming a new model where it can discern the rudimentary shape of the face of a ... cow.

The intriguing aspect of this visual "experience" is that, upon revisiting the first image, the reader may now perceive the cow's face without the aid of the line – a detail your brain was unable to discern initially.

But how does Predictive Processing explain this phenomenon? Essentially, it posits that, after your brain processed the second image, it updated the model that

initially lacked meaning, incorporating this new information into the existing model. This process enables something that previously appeared meaningless to now hold meaning.

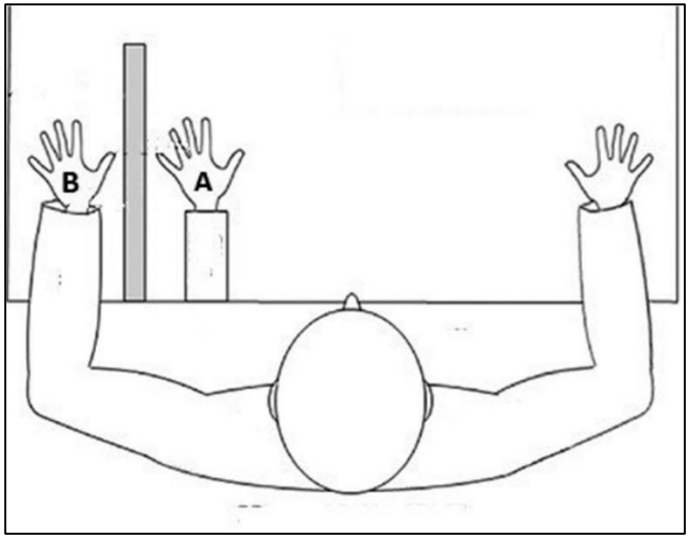
Notice that the sensory information processed by your brain, both the first time you glance at the initial image and when you see the same image a second time (after seeing the second image), remains exactly the same – the visual stimulus has not changed but remained consistent. What has changed is the most plausible "guess" that your brain forms, influenced by prior experiences.

In this sense, this theory contends that what we perceive is a result of both the impact of the world on the brain and, perhaps even more significantly, the influence of the brain in predicting that world.

It is important to note that Predictive Processing can be applied not only to perception and the senses but also to various other aspects of the mind. Essentially, this theoretical approach seeks to elucidate the nature of all mental phenomena – from consciousness and memory to psychiatric illnesses – based on the same fundamental and unifying principle.

See another interesting example to illustrate the potential of this framework when applied to the self. The rubber hand illusion is a phenomenon studied in psychology that delves into the way your brain predicts,

much like in perception, your self – what your brain deems to be part of your body. This experience has the following structure:



4. Illustration of Predictive Processing on Self; Hand A = rubber/fake; Hand B = true.

Now, imagine that you are the participant in this experiment. The scenario unfolds as follows: your left hand is resting on the table, but you cannot visually see it due to an obstruction (in the image, positioned at point B). Farther to the right of your left hand, a rubber hand is placed within your visual field, positioned where your (actual) left hand would be if it were not covered (in the image, at point A).

The subsequent step in the experiment involves having a person in front of you who will simultaneously touch your real hand (which you cannot see due to the obstruction) and the rubber hand (which is visible in front of you).

Finally, after a few minutes of stimulation, something magical seems to occur: if, for instance, Michael Myers were to suddenly appear and thrust a knife into the rubber hand that you can see, you would instinctively and immediately lift your left (real) hand off the table, reacting as if it were your actual hand that had been stabbed, rather than the rubber hand.

What happened here? The idea is to argue that the consolidation of sensory information from touch combined with visual information is adequate for your brain to construct a model of your body, wherein the rubber hand is presumed to be a part of it.

From the standpoint of this theory, even the perception of what constitutes our body is a "controlled hallucination" by the brain – a guess or a prediction.

Now, just as we can be deceived with visual illusions, the way we experience our body can also lead to mistaken models: this thesis can provide an important contribution to the comprehension of psychiatric illnesses that deal with disturbances of the self, such as bipolar disorder or schizophrenia.

But how does this continuous prediction really happen in our brain according to this theory, both in relation to the external signs of reality and the internal signals of our body?

A fundamental concept to understand this approach is that of 'Bayesian Inference,' which enables us to explain how, on the one hand, we form our predictive models, and on the other, how we can constantly update them and adjust our beliefs to new information.[62]

In practice, Bayesian inference involves combining two sources of knowledge: the predictions generated internally by the brain, known as 'prior,' and the sensory information coming from reality, referred to as 'verisimilitude.'

These two sources are assimilated to form a new estimate, termed 'posterior,' representing the updated model of the world. Essentially, our brain aims to strike a balance between its internal expectations and external evidence, allowing for constant adaptation to reality.

Let's examine a simple example that illustrates the utility of this method. We'll start with the ‘prior’: before entering the house, the reader assumes there is a

[62] Based on: Gouveia, S. & Curado, M. (2020) (eds.) *The Philosophy and Science of Predictive Processing,* New York: Bloomsbury.

minimal probability (1%) of finding a cat, given the knowledge that you don't own any cat.

However, upon opening the door and hearing a meow (an observation falling under 'verisimilitude'), and considering that cats meow, the belief is updated, suggesting a high probability (90%) of a cat being in the house. By combining the initial belief with the new evidence, the 'posterior' conclusion is reached: there is a high probability of a cat being in the house.

Other fundamental and essential concepts crucial to understanding the foundations of the Predictive Processing theory as a unifying theory of the human mind, such as 'Active Inference,' 'Prediction Error,' and the 'Free Energy Principle,' will be introduced in the next section. Before delving into these concepts, let us explore the usefulness of this approach in explaining – at least in part – the nature of various psychiatric illnesses

Following what was said previously, the brain is a prediction machine that combines internal models with stimuli from the world. We can argue that, in schizophrenia, this ability to make predictions more effectively is compromised for a several reasons. What are those reasons?

For instance, individuals may struggle to distinguish between internal and external sources of information,

causing them to confuse certain internal thoughts with external sensory stimuli.

It is well known that one of the most common positive symptoms of this disease is auditory hallucinations, where the subject believes that the internally generated 'voices' have an external origin in the world – that is, that someone is actually telling them something.

If we shift our focus to individuals with autism, it can be argued that they may encounter challenges in integrating external sensory information. This difficulty might lead them to become excessively focused on their own internal models.

As these individuals encounter more challenges in incorporating new sensory stimuli, their ability to update their models and, consequently, to adapt more effectively to the world is significantly compromised.

It is recognized that these individuals tend to prefer maintaining specific routines and patterns, facing challenges in social relationships precisely due to their diminished ability to predict the behavior of others.

Imagine now an individual whose brain has a negatively biased tendency to anticipate models of the world: this would lead that person to consistently harbor pessimistic expectations, distorting their ability to predict positive outcomes.

This could be rooted in a "rigidity" within the self's relationship with the world, causing the relational model between them to persistently remain negatively inflexible. In other words, it does not allow the entry of new positive information, perpetuating the negative bias.

In this scenario, we would be dealing with a patient experiencing ... depression, an illness that impacts millions of people worldwide. It is known, for example, that individuals with depression tend to view themselves negatively, selectively processing information that reinforces low self-esteem.

And what if a brain assigns excessive predictive importance to stimuli, constantly generating predictions of future events deemed threatening, leading to exaggerated anticipation that results in excessive worries about danger? The psychiatric problem such an individual would have is, indeed, anxiety (in this case, chronic!).

Due to a distorted interpretation of various threats in the environment, an anxious individual's brain will encounter challenges both in updating internal models of "safety" in their surroundings and in magnifying everyday stimuli as potential dangers. This amplification of stimuli contributes to the symptoms of anxiety.

As I tried to show you with these brief examples from psychiatry, the Predictive Processing theory exhibits significant explanatory potential and could indeed contribute to the elucidation of specific mechanisms in various mental illnesses.

Moreover, it can also shed light on how our perception, self, memory, dreams, and even consciousness operate.

Next, and in continuation with this approach, we will introduce some ideas that constitute the basis and the foundation of this perspective, as developed by one of the most influential neuroscientists globally, Professor Karl Friston.

III. THE FREE BRAIN

The second dialogue features the participation of Professor Karl Friston, one of the foremost authorities in brain science and a founder of the perspective described in the previous section. Following a brief introduction to Predictive Processing, I will now present some principles that constitute the basis of this approach to the human brain. This section will directly engage with the previous one.

Let's begin by introducing the "Free Energy Principle." This principle has a long history and was developed by several generations of intellectuals from various disciplines. It was formalized more recently in neuroscience by Karl Friston, who is currently the most cited living scientist in the world. In essence, it is suggested that the human brain has an intrinsic propensity to minimize energy consumption when engaging in various cognitive and mental tasks.

The idea is that, to conserve energy, the brain strives to organize itself in the most efficient manner, optimizing the predictions about the world as described earlier. According to this principle, the brain seeks to create internal and accurate models of the world with as little "effort" as possible.

In practical terms, the Free Energy Principle implies thinking the brain as continuously striving for a balance between (i) keeping internal models stabilized and (ii) being flexible enough to be able to adapt to new information from the world. This approach allows to formulate how organisms deal with uncertainty in the most economical and efficient way possible.

In a nutshell, we can claim that the Free Energy Principle emphasizes the 'saving' of energy as a guiding principle in the organization and functioning of the human brain, aiming to ensure the organism's subsistence and survival over time.

We should note that this principle has a complex mathematical form that we cannot explore in an introductory book of this nature. However, in essence, it is a principle that can be considered "simple" to understand. Imagine that you are driving on a road that is familiar to you (for example, from work to home, a routine undertaken every day).

In this scenario, your brain has constructed, through numerous previous trips, an efficient internal model of the route you must take. This model incorporates the positions of curves, traffic lights, traffic signs, and other reference points. The purpose is to conserve energy by minimizing the effort required to predict the characteristics of the road each time you drive through it.

This internal model utilizes information gathered in the past and employs the same data to predict potential outcomes. For instance, if you are aware of a traffic light at the end of the street, your brain will anticipate this fact, preventing surprise if the light turns red. This process facilitates the formation of a "stable" road model, which is updated only when necessary, resulting in significant energy savings.

Nevertheless, if, for some reason, a change occurs (e.g., a new traffic light is installed), your brain will strive to adapt and integrate this new information into the existing model. Instead of reformulating the entire internal model of the road, it updates the previous one with this new data, consistently aiming to minimize elements of surprise.

Despite this simple example, it is crucial to grasp the potency of this principle: fundamentally, it elucidates how an organism maintains its survival. This principle is applicable to a broad spectrum of organisms, ranging from simple cells like bacteria to highly complex beings such as humans and other animals.

Both simple and sophisticated organisms are regarded as dynamic systems in nature, distinct in their existence from the environment where this existence unfolds. This setup implies an indirect interaction and a continual exchange of information: the goal of this interaction is to fulfill a second fundamental concept, namely the concept of "Prediction Error Minimization."

The aim of what has been discussed so far is to ensure that the prediction models created by the brain minimize the possibility of making mistakes: an accurate model will be closer to reality and, therefore, will not require such an expensive energy expenditure, when compared to a model less accurate that requires frequent updates.

Thus, all organisms try to minimize errors in predictions made about the environment through the Free Energy Principle. But what does “free energy” mean exactly? In this context, "free energy" can be considered a metric linked to surprise or uncertainty in predictions: the more an organism can minimize its free energy, the better it can ensure its survival.

It is crucial to note that this principle operates not only in predicting the external states of the environment, as highlighted earlier but also in predicting the internal states of the body. The organism, in addition to predicting models of the environment at a given moment, simultaneously predicts the conditions that internal organs must produce to ensure survival in this environment.

This could imply the following: if there are errors in predicting certain external (environment) or internal (bodily) conditions, we will encounter a scenario leading to high free energy. This, in turn, will manifest in states of surprise or high unpredictability.

If, however, the prediction is accurate, the outcome will be low free energy. Consequently, all organisms, whether complex or simple, mammals or birds, plants or bacteria, strive to minimize this uncertainty or surprise.

It is pertinent to highlight another characteristic of this principle. While most biological theories aim to explain why a biological phenomenon works in a certain way, this approach attempts to reverse the question. Instead of focusing on describing what an organism must do to exist, this approach seeks to understand, assuming the organism already exists, what it needs to do to continue prevailing.

This is relevant to understand that the Free Energy Principle, in itself, is not a falsifiable theory about how organisms behave: only its postulates can be falsifiable or not, depending on their applicability.

Finally, to conclude this brief introduction, we need to address another fundamental concept, that of “Active Inference”. This concept allows the organism, in cases where the brain identifies a problem in the models generated in a given context, to seek new information relevant to that particular situation in order to adjust the incorrect or imprecise predictive model.

Hence, Active Inference can be regarded as the brain's capacity to actively influence the information we gather from the environment. This entails selecting specific

stimuli and directing attention to factors (which, in this case, can be external or internal once again) that may be relevant in reducing the identified ambiguity or uncertainty.

Consider an example related to predicting internal states of the body, such as physiological states. Imagine that, suddenly, you are transported from Copacabana beach to the incredible Swiss Alps. Your body, through interoception, identifies a change in temperature and produces a sensation of cold. This interoceptive signal indicates that your body temperature will decrease if you remain in that environment.

In this scenario, what Active Inference allows is to seek information about that environment and also to explore other internal body signals that may have been activated. For instance, your brain may have focused on the tremors your muscles would begin to feel, which would be an indicator of your body's actual temperature.

Considering all these elements, your brain reformulates the temperature model it had in Copacabana (where, for example, you would be sweating to fight the heat) to a model where other elements, such as tremors (which serve to increase internal heat), become integrated.

Therefore, in this example, it is through Active Inference that the brain actively optimizes – focusing its energy and attention on both the external signals of the

environment and the internal signals of the body – the previous predictions to maintain adequate body temperature regulation. If this adjustment hadn't occurred, it could have led to some rather ... unpleasant consequences.

Some neuroscientists and philosophers have criticized this approach[63] arguing that not everything happening in the body seems to happen with the sole aim of minimizing energy consumption, given that there are several mental functions, such as creativity or imagination, that seem to demand a much greater expenditure than what is considered normal.

Interindividual neuronal variation (between individuals) itself seems to differ, and this diversity seems to be difficult to explain by appealing to the action of the same principle. But perhaps the most significant challenge is to actually measure and quantify the entire predictive mechanism that takes place in every part of the organism.

Next, we will address another perspective of looking at the brain and its role in the development of conscious processes, presenting a theory that aims to provide precisely a way of quantifying those conscious states, developed by the neuroscientist and President of the

[63] Article by Professor Karl Friston: Friston, K. et al. (2023) "The free energy principle made simpler but not too simple", *Physics Reports*, 1024: 1-29.

Allen Institute for the Brain in Seattle, Professor Christof Koch.

IV. THE INTEGRATED BRAIN

In this third dialogue, we will introduce one of the trendiest neuroscientific theories about the brain and its role in the development of the conscious mind, developed by Professor Christof Koch, author of the book *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed*.

This theory, titled Integrated Information Theory (IIT),[64] has been revolutionizing the way consciousness is considered within neuroscience. Given its conceptual and theoretical nature, it has influenced several adjacent areas, such as philosophy, physics or even artificial intelligence. Because of its significant scope, this theory has also prompted interesting ethical considerations.

Firstly, it is important to highlight that this approach aims to offer a scientific and objective answer to the problem posed by the existence of consciousness in the physical world, seeking to identify the essential properties – referred to as "axioms" – and infer the

[64] The first author to develop this theory was Giulio Tononi in 2004, with whom Christof Koch, after having worked with Nobel Prize winner Francis Crick, joined forces to further develop this theory. Original publication: Tononi, G. (2004) "An Information Integration Theory of Consciousness", *BMC Neuroscience*, 5 (1): 42.

necessary and sufficient conditions that any substrate must satisfy to be conscious – termed “postulates”. All of this is expressed through a specific mathematical notation that, due to its complexity, we will not address in this book.

The IIT strategy aims to reverse a typical approach found in neuroscience's scientific research of consciousness. Rather than attempting to identify the neuronal processes in the brain responsible for conscious states, Koch prefers to start with phenomenology (consciousness itself!) and then inquire about the kind of physical mechanisms that could explain the phenomenology of experience.

For this approach, the definition of conscious experience closely aligns with that offered by Descartes, who considered consciousness as a fundamental property of the world. Building on this assumption, IIT advances with the following 5 axioms:

- Axiom of Existence (AExis): this axiom is, in essence, a reformulation of the Cartesian cogito, replacing the act of “thinking” with the act of “experiencing”: “I experience, therefore I exist”; this means that we have an indubitable and immediate certainty of our own first-person perspective on the world; consciousness is, following this axiom, intrinsically real;
- Axiom of Information (AInfo): this axiom indicates that an experience must always specify

something, always being distinct from other possible experiences; thus, we can claim that every experience informs us in a particular way through a contrast with other particular experiences;

- Axiom of Integration (AInteg): this axiom claims that consciousness is unified, and that the content of experience is irreducible to independent parts, being rather integrated into a whole; we do not have, therefore, isolated experiences from each other that only later form a mental set; rather, when the reader experiences reading these words, you experience the font and color in a unified experience, and not separated from each other;
- Axiom of Composition (AComp): this axiom advocates that all experiences have a structure, made up of various aspects and various combinations of each other; again, if you are experiencing these words, it contains different phenomenological aspects such as the colors and shapes of the letters;
- Axiom of Exclusion (AExcl): this last axiom indicates that conscious experience always excludes other experiences, given that, when it specifies an experience (cf. AInfo), it will necessarily exclude other experiences; in other words, conscious experiences have boundaries; furthermore, they have a temporal "grain": that

is, the contents of the experience have a specific duration.

Let's see how these axioms can be applied to an everyday example, such as reading a book, in order of the axioms presented.

Imagine that you are at home, on the couch, in your precious free time, enthusiastically leafing through this book, which indicates the presence of intrinsic conscious experiences (AExs). During this reading, the information in the book is processed in a unique way in your consciousness: the various patterns and knowledge contained in the words demonstrate the informative nature of this experience (AInfo).

Moreover, this information present in the book forms a unified and cohesively integrated conscious experience between the words, the images, and the emotions that you may feel when reading the book (AInteg). Of course, this conscious reading experience is made up of particular elements, such as the various discussions and the author's introductions: all these elements contribute to a unique composition of your reading experience.

Finally, during the time you are reading this book, many other information may be happening simultaneously – like a football match of your favorite club being on television. However, if you find the content of this book interesting, your conscious experience may even

exclude all other simultaneous experiences, demonstrating the unique nature of consciousness at a given moment (AExcl).

In addition to these axioms – which are taken as self-evident truths – IIT proposes, based on each of these axioms, the kind of properties a physical system must possess to be conscious. Let us then examine the postulates of IIT:

- Postulate of Existence: this postulate states that the existence of consciousness implies a system of mechanisms with the power of cause and effect; for a physical substrate to exist, a necessary condition is to possess causal power;
- Postulate of Information: this postulate states that, if consciousness is indeed informative, it must have the ability to specify or differentiate certain experiences from others; this implies that any mechanism within a physical system must possess cause-and-effect powers; all these "repertories" form the cause-effect structure that allows the system to specify a certain state;
- Postulate of Integration: this postulate indicates that, for the integration of consciousness to occur into a unified whole, the physical system must necessarily be irreducible: the parts of that system must be interdependent; all mechanistic elements of the system must be capable of causing something as a whole, as well as being

affected by it; if a given physical system can be divided into parts without affecting its cause-effect structure, then that system loses its integration and cannot be conscious;

- Postulate of Composition: this postulate implies assuming that the elements (mechanisms) of a physical system must have the capacity to be combined with each other, guaranteeing that these combinations have cause-effect power; that is, if an experience has a determined structure, the causal process that generates that experience must necessarily be structured;
- Postulate of Exclusion: finally, this postulate involves the idea that a conscious state of a physical system must be finite: different mechanisms of a system can have varying cause-effect powers; now, for there to be consciousness in a system, only one of these mechanisms can have the most irreducible cause-effect structure, which represents the highest level of integration in terms of information.

Let's consider another example to understand the relevance of what has now been described. Imagine that the reader suddenly wakes up from your beauty sleep and see, in front of you, the blue wall of your room. Furthermore, you also see where you bed is, and you realize that, at that moment, you exist consciously of yourself and what is around you.

Now, what IIT states is that this experience exists for oneself, and not for anything else; it is a specific experience (and not a generic one); it is a unitary experience: your right eye is not experiencing anything different from your left eye (both experience the same experience); furthermore, that experience is defined specifically by what you are observing in front of you and not what is behind you or to your side; finally, that experience is structured by several elements that relate to each other (e.g. your body, in your bed, which is situated so that you can see the blue wall in front of you) that allows you to have precisely that experience and no other.

Did you notice what we just did? That's right: through this example, we applied all the IIT postulates.

It is important to emphasize that, for IIT theorists, these axioms (and consequently, the postulates) are part of a complete and finite list: there is no other property of conscious experience that should be considered essential in its existence. If a physical system – whether biological, artificial, or even alien – wants to be conscious, it will have to reproduce, through postulates, all the axioms now presented.

Another attractive aspect of this theory is that it provides an attempt to quantify, to a particular extent, the ability of a system to integrate information. This quantification is presented through the measurement Φ – the Greek letter *Phi* – which represents the ability

(between zero and maximum) of a system to integrate the information processed in a particular system.

According to IIT, consciousness “happens” when a system reaches a high level of integrated information: the more intrinsic and irreducible this information, higher its Φ and, therefore, higher the degree of consciousness. The calculation takes place through the amount of information that is globally integrated in this system compared to the sum of information in each of its parts.

Thus, whenever a system reaches a maximum Φ, that system will be considered conscious: below this maximum value, it is argued that the system will not have enough integrated information to reach a conscious level.

This quantifiable approach – together with the axioms and postulates of IIT – provides an incredibly elegant theory and, above all, with the capacity to test its assumptions. Let's explore some interesting predictions, based on real empirical data, showcasing the explanatory power of this theory.

The cerebellum is an important region of the nervous system responsible for motor function located at the back of our head. Although this structure weighs only 150 grams and makes up around 10% of the brain's total volume, it has 4 times more neurons than the rest

of the brain combined, being highly connected with other brain regions.[65]

Now, despite this, if this organ is removed from the reader's brain, you won't have to worry too much: your consciousness will remain completely intact! How can this be the case? According to IIT, as the cerebellum is made up of modules that are independent of each other, the information processed there is not as integrated as in other areas of the brain and, therefore, its Φ is lower than the minimum necessary for consciousness.

Do you remember some ideas we presented in the first part of the book about sleep? IIT also predicts something interesting about what happens to our brain and consciousness when we go to sleep. Assuming, then, that consciousness can be measured by Φ, it is intuitive to postulate that when we sleep without any conscious experience – without entering into lucid dreams, for example – our brain should experience a marked decrease in neuronal connectivity and, consequently, be less integrated.

What is interesting is that some empirical studies on human brains during sleep show that this prediction is

[65] Original publication: Lemon, R. & Edgley, S. (2010) "Life without a cerebellum", *Brain*, 133: 652–654.

indeed accurate: the brain decreases its cortical connectivity.[66]

Another intriguing prediction of this theory is related to a kind of neurosurgical procedure known as "split-brain" callosotomy. This procedure does something curious: it divides the brain into two hemispheres disconnected from each other through a precise incision in the corpus callosum.

In a famous scientific experiment on patients with epilepsy,[67] after removing the corpus callosum – to reduce symptoms – something predicted by TII happens.

If one of these patients observe an image with only the right eye, processing visual information through the left hemisphere of the brain, and we ask him what he saw at that moment, he will tell us that he is unable to remember seeing something.

However, if we ask the same patient to draw what he saw, he will be able to accurately depict the observed image. How can this happen? Are we dealing with some demonic magic again? Certainly not: the reason for this

[66] Original publication: Massimini, M., ..., Tononi, G. (2005) "Breakdown of cortical effective connectivity during sleep", *Science*, 309: 2228–2232.

[67] Original publication: Gazzaniga, M. (1967) "The split brain in man", *Scientific American*, 217 (2): 24-29.

occurrence is due to the fact that different functions are located in different parts of the brain.

While language is a cognitive function mainly present in the left hemisphere (which did not process the information being assimilated in the right hemisphere, as the connection between the two was disconnected), the function related to drawing is present in the right hemisphere.

Now, according to IIT, when the corpus callosum was disconnected, the brain ceased to be, in its entirety, a complete system, creating two independent systems in each hemisphere that could maximize its Φ and, thus, creating two streams of consciousness independent of each other.

Another interesting prediction linked to brain malfunction is focused on brain injuries: this theory predicts that any kind of injury that affects the brain's ability to integrate information will prevent the formation of conscious states.

Recent studies confirm that this prediction is indeed precise: through transcranial magnetic stimulation, several patients with brain injuries exhibited substantial changes in their brains compared to healthy individuals.[68]

[68] Original publication: Casali, A. et al. (2013) "A theoretically based index of consciousness independent of sensory processing and behavior", *Science Translational Medicine*, 5: 198ra105–198ra105.

Could this theory, in addition to providing an explanatory narrative of how the biological brain is conscious, also shed some light in relation to the development of conscious artificial systems? The answer is mixed: on the one hand yes, and on the other, no.

On the positive side, in theory, IIT does not specify that a system must be biological to be conscious: it only indicates that it must be a system that can integrate information following the axioms advanced previously.

However, on the negative side, it is clear that current Artificial Intelligence, based on feed-forward structures in which information is only processed in one direction (from input to an output), will never be conscious.

This is because a system of this kind will not have any Φ, even if it replicates conscious behavior perfectly; it would only be a simulation. For consciousness, the system would need to have a re-entrant structure in which outputs can be used as inputs and vice versa.

To close these introductory notes on IIT, it remains to address a criticism – which can be seen as an advantage – raised against this theory. For some authors, IIT may have counterintuitive panpsychist consequences. And what is panpsychism? Panpsychism is a theory developed by various intellectuals throughout history –

from Spinoza to Leibiniz or Bertrand Russell[69] and, more recently, Philip Goff[70] – which contends that everything that exists in the physical world has fundamental mental properties.

Now, this means stating that consciousness is a fundamental property, which goes against the idea advanced by IIT that some complex systems, when they reach high levels of integrated information, become conscious.

Furthermore, both panpsychism and ITT agree that consciousness is not an all-or-nothing phenomenon: on the contrary, consciousness must be seen as a spectrum, between the absolutely unconscious and the completely conscious.

Despite these similarities, it seems somewhat radical to identify IIT as a panpsychist theory. Although both have, in fact, Cartesian commitments by accepting subjective consciousness as fundamental in its nature, panpsychism advocates that consciousness is fundamental to all existing entities, while IIT only suggests the necessary conditions for consciousness to emerge in certain complex systems.[71]

---

[69] Original publication: Russell, B. (1927) *The Analysis of Matter*, London: Kegan Paul, Trench, Trubner & Co.

[70] Original publication: Goff, P. (2019) *Galileo's Error: Foundations for a New Science of Consciousness*, London: Pantheon Book.

[71] Section based on: Tononi, G. (2012) "Integrated information theory of consciousness: An updated account", *Archives italiennes*

To conclude these introductory notes, we will address another perspective on the brain through emotions and their role in the development of our mind, presenting some of the central theses developed by the neuroscientist and Director of the Emotional Brain Institute at New York University, Professor Joseph LeDoux.

de biologie, 150: 56-90; Tononi, G. & Koch C. (2015) "Consciousness: Here, There and Everywhere?", *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370 (1668): 20140167.

V. THE EMOTIONAL BRAIN

In this fourth and final dialogue of this odyssey into the world of the conscious brain, we will present some of the key concepts developed by neuroscientist and author of the book *The Emotional Brain*, Professor Joseph LeDoux. The first step towards understanding what this emotional brain entails is understanding the anatomical structures that compose this vital system for the survival of each of us.

The part of the brain that processes various emotions is called the 'limbic system' in neuroscience. Comprising several parts and located on the medial surface of the mammalian brain, it plays a fundamental role in the functioning of our emotions and, consequently, in the way we relate to ourselves and others.

Via the autonomic nervous system, it positively or negatively influences visceral functioning and metabolic regulation in the organism, controlling behaviors considered essential for the survival of all mammals.

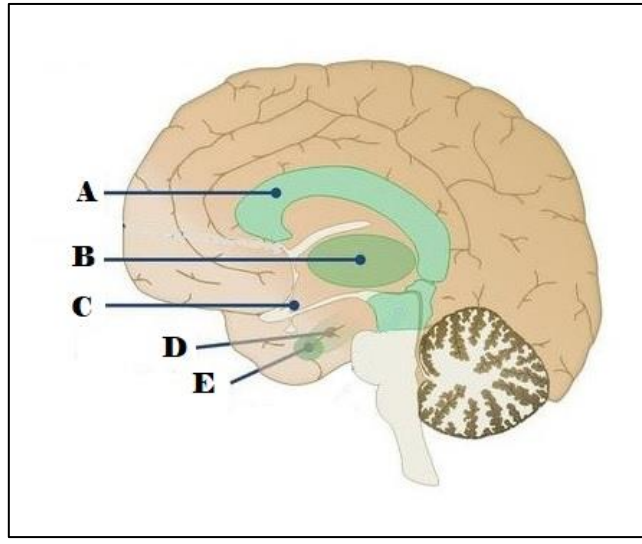
The appropriate emotional response is then generated by integrating sensory information into a specific mental state, attributing affective content to the registered stimuli and associating them with pre-existing memories.

Let's look at some of the main anatomical structures that belong to the limbic system:

- Amygdala: located in the anterior temporal lobe, it is connected to the hypothalamus and is regarded as the “center” that guides us in dangerous situations inducing states of alertness or sensations of fear; its removal in mammals causes a sexually indiscriminate posture, indifference to risky situations and loss of affective sense of external stimuli; its electrical stimulation generates aggressive and violent behavior;
- Hypothalamus: located in the center of the brain, it is located just below the thalamus and above the pituitary gland, being considered the most fundamental part of the limbic system since it is responsible for the vegetative functions of the brain (maintenance of internal balance, body temperature, hunger, thirst, hormone production, etc.); furthermore, it is directly related to behavior, playing an essential role in emotions (the middle part is linked to aversion and the lateral parts to pleasure and anger);
- Thalamus: located in the diencephalon, just below the corpus callosum, it is in this part of the limbic system that sensory information captured from the environment is selected, and is also responsible for the motor part (muscle activation); it is also related to states of alertness

and is considered fundamental in the perception of pain, being important in the regulation of sleep and wakefulness states;

- Hippocampus: located in the inner part of the medial temporal lobe, it is responsible for short and long-term memory, allowing the organism to compare current situations with past experiences, thus increasing its survival capacity; It is also responsible for spatial orientation and navigation.
-



5. Limbic system: A = Corpus callosum; B = Thalamus; C = Hipothalamus; D = Hipocampus; E = Amygdala.

Starting from this succinct neuroanatomical description, we can now delve into the hypotheses advocated by Joseph LeDoux. This neuroscientist supports a novel conceptual framework that redefines crucial concepts in the examination of emotions, particularly the functions and circuits pertinent to the survival of organisms.

In this novel approach, we aim to shift the focus from the conventional method of examining emotions. Rather than asking whether the emotions consciously experienced by humans are also present in other animals — an approach prone to anthropomorphism — we inquire about the extent to which the functions and circuits crucial for the survival and preservation of animals are either present or absent in us, humans.

It is crucial to grasp that, according to LeDoux, these circuits and functions are not causally related to emotions, even though they do play an indirect role in contributing to them. So, what precisely do these survival circuits entail? Essentially, this wider concept aims to encompass all types of mechanisms that are pertinent to the organism's preservation and success in its daily existence.

And why introduce this new concept instead of discussing 'emotions' directly? LeDoux puts forth a compelling argument to justify this shift: he contends that the concept of 'emotion' is philosophically inconsistent, lacking consensus in the scientific

community regarding its precise meaning. Therefore, he suggests this methodological inversion to avoid being 'trapped' by the uncertainty inherent in this conceptual ambiguity.

Following this interesting philosophical innovation, LeDoux aims to concentrate on phenomena that are relevant to the study of emotions, avoiding, however, the use of particular emotional concepts. Instead, the emphasis is placed on the specific circuits that instantiate certain enabling the organism to survive, such as sensing and responding to challenges and opportunities. And it will be these particular instantiations that can be labeled with some emotion afterward (but not vice versa).

These circuits involve developing capabilities for identifying danger, potential mating partners, the presence of food or water, defense and energy maintenance, thermoregulation, among many others. Essentially, the scientist's challenge lies in describing all these processes without resorting to the use of confusing and ambiguous language associated with emotional concepts.

Note that these survival circuits are conserved across all mammalian species, and perhaps in many other animals and organisms. While there may be some differences, essential components of these functions are shared by all animals.

We thus observe the utility of the LeDouxian inversion: as these survival functions are instantiated in circuits conserved by evolutionary history, we can avoid anthropomorphizing the question by asking, 'what human emotions are present in other animals?' and instead, inquire, 'which circuits present in other animals are also present in humans?'.

The idea behind this inversion is to consider emotions and related concepts (motivation, reinforcement, inhibition, excitement, etc.) as components of a broader process, not confined to specific feelings.

Therefore, what follows from this approach is not an attempt to explain or define emotions: rather, the aim is to provide a framework for thinking about some phenomena associated with emotions – phenomena related to survival – in a way that is not confused with the search for the too broad meaning of “emotion”.

This allows to focus on key aspects that avoid endless debates about a correct definition of “sadness”, “fear”, “happiness”, among all the other concepts we use to describe our feelings.

This emphasis on the connection between survival functions preserved by evolution and emotions is not entirely novel; rather, it follows a tradition dating back to at least Charles Darwin (1872). Hence, it won't come as a surprising innovation for neuroscientists that these circuits are somehow linked to the brain.

LeDoux's strategy – and its novelty – lies in focusing precisely on these circuits to substantiate various emotions. This marks a departure from the conventional approach in neuroscience, where the usual strategy involves starting from *a priori* definitions of emotions and subsequently attempting to identify the circuits related to them. Simply put, the goal is to first identify the circuit and only then associate it with the emotion, as opposed to beginning with the emotion to pinpoint the circuit.

A second interesting issue in the investigation of emotions – where LeDoux's insights are relevant – involves understanding whether there are innate emotional circuits in the brain or if they are creations of the human mind influenced by social and cultural aspects. The complexity of this debate revolves around the question of whether the so-called 'basic emotions' are natural or not.

Basic emotions are those believed to be expressed by all individuals and are also present in the closest animals, evolutionarily speaking, to human beings. Ekman's list (1972, 1992)[72] provides a canonical example of basic emotions, including happiness, fear,

[72] Original publication: Ekman, P. (1972) "Universals and cultural differences in facial expression of emotion" In J. Cole (Ed.), *Nebraska Symposium on Motivation*, Lincoln, Nebraska: University of Nebraska Press: pp. 207–283. An interesting update can be found at: Ekman, Paul (1992). "An Argument for Basic Emotions", *Cognition and Emotion*, 6 (3/4): 169–200.

sadness, anger, surprise, and disgust. Many current neuroscience investigations aim to identify the neuronal basis of these emotions.

Despite numerous neuroscientific investigations, the concept of 'basic emotions' has posed several challenges. Various theories either include additional emotions or exclude some from Ekman's list. This is likely due to the difficulty of aligning an emotional concept with a diversity of biological states that may or may not coincide with that concept.

Furthermore, there are two more central difficulties in this approach to “basic emotions”:

> (i) some argue that emotions are psychological and social constructions arising from the interaction between the physical or social environment; therefore, there is no inherent biological determination from the outset;
>
> (ii) the vast majority of theories associated with basic emotions are grounded in research on the brains of animals that do not align with Ekman's list or other proposed basic emotions (Panksepp, 1998, 2005).[73]

[73] Original publication: Panksepp, J. (1998) *Affective Neuroscience*, New York: Oxford U. Press. A relevant update can be found at: Panksepp J. (2005) “Affective consciousness: Core emotional feelings in animals and humans”, *Consciousness and Cognition*, 14: 30–80.

For example, Louise Barrett (2006)[74] considers to be meaningless to argue that basic emotions were conserved through evolution through neuronal circuits for three particular reasons:

- some imaging studies show that areas activated by the same stimulus were associated with different basic emotions;
- much of the evidence in favor of basic emotions has been collected in animal brains through retrograde techniques (e.g. electrical stimulation of the brain) with a marked lack of precision;
- the basic emotions identified in studies on humans do not coincide with the basic emotions identified in animals.

Now, this controversy can be clarified, once again, through LeDoux's approach: instead of discussing 'basic emotions', scientists and philosophers should favor the term "survival circuits." This shift would enable them to address and resolve conceptual confusions – which have direct implications for empirical research – caused by the use of the language of emotions in the first instance.

These survival circuits encompass systems associated with nutritional maintenance, fluid balance, defense, reproduction, and thermoregulation. These circuits are

[74] Original publication: Barrett, L. (2006) "Are Emotions Natural Kinds?", *Perspectives on Psychological Science*, 1:28–58.

believed to have been present in the earliest forms of life on Earth. For instance, bacteria demonstrate the ability to escape toxic environments and seek favorable conditions for their development by employing some of these circuits.

With the development of multicellular eukaryotic organisms, these survival circuits became increasingly complex, especially with the development of specialized sensory receptors, and a central nervous system capable of coordinating all bodily functions and diverse interactions with the environment.[75]

It must be clear, however, that LeDoux's goal, when describing the various survival circuits, is not to link them directly to 'basic emotions': rather, the aim is to free neuroscientific language from concepts based on subjective feelings of the human being and focus, neutrally, on such circuits conserved throughout evolutionary history.

To conclude, let's consider a brief example related to understand the relevance of this approach. We know that aggression is an emotion that does not have a single fixed "representation" in the brain: rather, there are different kinds of aggression depending on the context in which it occurs.

[75] Original publication: Shepherd, G. (1988) *Neurobiology II*, New York: Oxford.

For instance, aggression can be triggered by the defense circuit (in protecting the organism); it can be triggered by the food circuit (in the hunt for food); or it can be triggered by the reproductive circuit (in competition for mates).

Note, therefore, and against much research in neuroscience, that a survival circuit is not exclusively tied to a particular emotion, much less to a specific neuronal activation in the brain. On the contrary, different survival circuits can be related to one and the same emotion, and vice versa.

This could mean that these circuits do not have a direct relationship or causal role with emotions. Of course, there must certainly be an indirect influence: but the focus of these circuits is only one, to guarantee adaptive purposes in order to assure the preservation of the organism in the different complexities required by the interaction with the environment.

Following this example, it is now clear how the inversion proposed by LeDoux is useful: by looking at the shared survival circuits of animals and humans, and only then making considerations about the so-called "emotions," we will be closer to assuring relevant explanations and understanding the nature and role of these emotions in humans.[76]

[76] Section based on: LeDoux, J. (1996) *The Emotional Brain*, New York: Simon and Schuster; LeDoux, J. (2003) "The Emotional Brain,

Having concluded these second introductory notes, the reader will be able to enjoy the next dialogues with four more internationally renowned intellectuals on the nature of the brain through multiple perspectives. This will certainly enrich your knowledge about this fascinating odyssey that takes place inside your head.

Fear and the Amygdala", *Cellular and Molecular Neurobiology*, 23 (4/5): 727-738; LeDoux, J. (2012) "Rethinking the emotional brain", *Neuron*, 73 (4): 653–676.

DIALOGUES II

Brain

DIALOGUES II

Brain

VI. Dialogue with Anil Seth

Anil Seth is Professor of Cognitive and Computational Neuroscience at the University of Sussex, where he is also Co-Director of the Sackler Center for Consciousness Science. He is also co-director of the Canadian Institute for Advanced Research (CIFAR) Program on Brain, Mind and Consciousness and the Leverhulme Doctoral Fellowship Program: "From Sensation and Perception to Consciousness".

He is the Editor-in-Chief of Neuroscience of Consciousness (Oxford University Press) and was the Conference Chair for the 16th Meeting of the Association for the Scientific Study of Consciousness (ASSC16, 2012).

His research was funded by the EPSRC (Leadership Fellowship), the European Research Council (ERC, Advanced Investigator Grant), the Wellcome Trust and the Canadian Institute for Advanced Research (CIFAR).

He has published several books, the most recent being *Being You: a New Science of Consciousness* (Faber & Faber), which was considered one of the top 7 science books by The Guardian in 2021.

More information: https://www.anilseth.com/

Question: You are one of the leading neuroscientists advocating for the Predictive Processing framework, which has a long history and interesting assumptions about the human mind and the brain. When did you first encounter the idea that the brain is a prediction machine?

Anil Seth: It was not remotely my idea, it just seemed to me a really attractive way to think about not only how brains work in general but also in the topics I was particularly interested at the time, and still am: how do we think about the neural correlates of consciousness; how do we think about relating what happens in the brain to what happens in our consciousness experiences; and also, how do we understand emotions and the self.

Thinking about the ideas of predictive coding, which I first heard about probably about twenty years ago but then started working on them properly about twelve or thirteen years ago, just seemed the right kind of language to use and apply to the brain and its functions.

The other thing that was really appealing to me was that predictive coding is something that you can think about from different levels: a very mathematical level, a computational level (you can write code that does it), but you can also think about it at a conceptual level and at a philosophical level.

In that sense, it is a really nice way to join these disciplines and I think that that is very important in current neuroscience.

Question: Delving deeper into the origins of the Predictive Processing framework, could you share a more comprehensive account of how this perspective unfolded? Who were the pioneering figures and influential advocates who initially proposed and championed the notion that the brain operates as an active agent in its interactions with the world?

Anil Seth: Plato (laughing)! It depends how far we go back. We know already back in Greek Philosophy we had the idea that what we take to be real, what we perceive in the contents of our experience, is not necessarily what is there.

It is a sort of reflection Plato has in his 'Allegory of the Cave' where he talks about the prisoners chained to the wall of the cave who take the shadows on the cave to be the real world. They do not know and cannot know any difference between the reality and the illusion of the cave. But of course, there was no idea of predictive coding as a principle of brain function at all back then.

There's also the Hindu concept of Maya, which emphasizes the process by which what we perceive has the character of seeming to be real – this, too, is

intimately related to modern concepts of predictive processing.

Then we have in the 15th century this Arabian Scholar, Ibn al-Haytham, who was probably the first person to think about perception as inference, this idea that what we perceive is a judgment about what is going out in the world, something that the organism estimates.

The next landmark would be Herman Von Helmholtz, the German scientist in the 19th century who approached this perspective in much more detail, coming up with the first formulation of the idea that perception is a process of brain-based inference, where the process of inference is itself unconscious. Von Helmholtz's formulation naturally aligned with the mathematics of inference coming from Bayesian reasoning.

But predictive coding itself, as a specific idea and algorithm, came out of engineering and signal processing, where people were trying to figure out how to compress long signals without losing information. The idea was: if you can predict something, then you do not have to explicitly encode it. This algorithm of predictive coding was developed in the 1950s.

Then, in about 1990 there was a first paper published, a classic reference, about the visual system: Rajesh Roh and Dana Balor talked about the visual system might actually be doing this, and they used it to explain

properties of the visual system. And since then, it has been taken on as a much more general theory of how the brain works.

Question: You adopt a critical stance towards the well-known dichotomy proposed by David Chalmers, distinguishing between the 'easy' and 'hard' problems of consciousness. Fundamentally, you reject this division, asserting that a more compelling approach involves addressing the 'real' problem of consciousness. Can you elaborate further on this strategy of conceptual change and how it facilitates a scientific and objective exploration of consciousness?

Anil Seth: I called it "the real problem of consciousness" to partly to wind up David Chalmers – in a friendly way - since he has contributed so much to the philosophy of mind, and I think he articulates this hard and easy problem distinction very well.

However, I do think it is a little bit of an obstruction too, since it really embeds the intuition that no explanation of brain or any physical system could ever explain why parts of the universe are consciousness and other parts are not. It's true that consciousness does not seem to be the sort of thing that could ever be explained in physical terms, this is a really deep intuition. And, from one perspective, this intuition is what the hard problem formalizes.

But then, there is a side effect that you end up treating consciousness as one single big mystery and you try to find an enchanted solution, a special source – perhaps some kind of quantum wizardry – that you hope can magic experience from mere mechanism. And maybe that is what it would take, maybe we need indeed a new physics or some new eureka moment, but I think there is another approach – which is the real problem approach – that can be more useful.

Again, this is not a new idea, it is just a sort of new emphasis that locates it within the language that Chalmers has made so popular. The real problem is just the idea of saying: Ok, so we do may not know how physical systems give rise to consciousness or are identical to it, but we do know that consciousness exists, and that it is intimately related, in lawful ways, to brains.

Consciousness is a fact of the universe that we live in, that is in need of explanation. The real problem approach argues that we shouldn't address the problem 'head on' – as solutions to the hard problem attempt to do, but we can divide the problem a bit, and try to understand which are the properties of consciousness, and how do we account for those properties.

The analogy I use on my book *Being You* (2022) is this idea that people – not so long ago – thought that life could not be explained in terms of mechanism, so they

spent ages arguing about whether there was a spark of life, or an Elan Vital or not. But in the end, it was neither the case that people found that the spark of life existed, nor that they agreed that life did not exist. They just realized that life is not one thing, but that is composed of many different properties, and that you can explain each of them in terms of chemistry and physics.

Following this strategy, the “hard problem of life” was dissolved: it was not solved. This is the approach I think is more useful for consciousness, even if does not eventually succeed in completely dissolving the hard problem. But we will only know if it works or not if we try, and this “real” approach does that.

Question: Therefore, this approach or conceptual change can, in the end, even give rise to new problems that might not have emerged initially. By dissolving the somewhat problematic distinction between the easy and hard problems of consciousness, we open up a new avenue of investigation that may lead to significant progress. Before this shift, if we assert that consciousness is hard in Chalmers' sense, there is limited potential for empirical investigations into the nature of consciousness.

Anil Seth: Exactly. I think that is a really nice point. I think it is a very underappreciated aspect of the interaction between science and philosophy. It’s not

that we have this fixed “menu” of questions and we just solve them, or fail to solve them, one by one – even though this is how it is often presented to the public.

There are so many articles that you can find, in newspapers and even in science journals, with titles like “10 biggest unsolved mysteries”. This totally neglects the dialectic between theories, data and the questions: you can track progress just as much by how the questions we ask changed, as how our answers develop throughout time.

Question: And we might even discover that the question or the primary problem we were initially attempting to solve does not make any sense at all, right?

Anil Seth: Right. Back to that analogy, if you try to apply for a grant to look for the Elan Vital presently, you would not get that grant. It is a question that does not make much sense, and we have seen this in physics with the old concept of “ether” as well. What does that imply?

I think it implies that the conceptual apparatus that we inherit regarding a particular question or mystery should always be interrogated, since it is always up for grabs, and should not be taken on board unquestioningly.

Question: Returning to the theory of predictive processing, do you believe that this predictive approach to the brain and mind can illuminate the development of conscious artificial intelligence? There is a lot of talk about "artificial consciousness" these days, but it does not seem sensible to aim for the development of "artificial" consciousness without first understanding how "natural" consciousness exists.

Anil Seth: To start simply, I think there is actually already a large contribution of predictive processing, as it is considered in cognitive science, to Artificial Intelligence. As I mentioned before, engineering was where predictive processing first came about as a method of signal compression.

Then, you have things in AI like 'Helmholtz machines', named after the same Helmholtz that originated the concept of perception as unconsciousness inference, and 'auto encoders', and algorithms that actually implement aspects of what we would call the theory of predictive processing in cognitive sciences – the key computations that constitute that theory.

These kinds of systems already depart the deep learning networks that are so prevalent in research and AI applications currently.

The core ingredients for predictive processing and AI are: we need generative models - something that is capable of generating predictions about its inputs. Then

you compare the inputs to those predictions, and update both the predictions and, over a longtime scale, the model. So, you can do perception and you can do learning with the same algorithm.

In fact, one of my colleagues – a really fantastically gifted former PhD student called Beren Millidge – showed that if you implement predictive coding, you can actually also get back-propagation, which is the most widely used learning algorithm in AI.

People had always criticized back propagation because they said it is not neurobiologically plausible: it is not the kind of thing that brains can do, because it requires global error signals that get propagated throughout the network. But it turns out that predictive coding gives you back propagation for free, in a way that is potentially biologically plausible, so it is another strong clue that this might be what brains actually do.

There is actually a lot of research in AI that is following this route and building not just deeper and deeper networks, but networks that have generative models: we see them all over the place now, such is an adversarial networks, some of which are useful for neuroscience, others of which may be less so, but I think there is real fertile territory there in this intersection.

Question: You published your new book, *Being You: A New Science of Consciousness*, which was very well

received, having been considered book of the year in the science category by *The Financial Times* and *The Guardian*. What main idea of the book would you like to highlight?

Anil Seth: I am going to say three main messages that I tried to argue for in my book. First, that consciousness can be addressed scientifically and philosophically without addressing the hard problem head-on. That is, we adopt a real problem approach, related to a neurophenomenological approach, and here I should give credits to Francisco Varela, who first proposed this approach to study consciousness.

The second idea is that we can use predictive coding to try explain not just what organisms do, how they take sensory input and use it to guide behavior, but we can map the contents of perceptual predictions onto what perceptual experiences are like. In this view, different kinds of visual experience can be explained because they involve different kinds of perceptual prediction.

The third and final thing is that all this predictive machinery also applies to the self, and this is really where the book goes. The self is not the 'thing' that does the perceiving, it is not that you have this little "mini me" inside your head that is perceiving all the perceptions and deciding what to do. What we experience as the self is also a sort of perceptual inference. In this case, the sensory signals come largely from within the body, from

your heart rate or your stomach, what we call interoception.

This is the big thesis of the book: everything that appears in consciousness is a perceptual prediction of some kind, and all of the perceptual experiences that arise are ultimately grounded in the body and in our nature as living machines: there is a very intimate connection between the living body, the experienced self, and consciousness in general.

Question: Can this cutting-edge approach carve a path through the intricate landscape of psychiatric disorders? Can we claim that many of these disorders might be based from a delicate imbalance between the anticipated input and the actual signals that ought to be processed by the brain's inferred model?

Anil Seth: I really hope that is the case, and this is certainly something we have tried to do in my research center in Sussex over the last ten years. I think that with all these things, there is a worry about overpromising. But I do think there is a lot to be done, since in mental health treatment and psychiatric medicine, many current approaches still primarily rely on treating the symptoms, rather than addressing the underlying mechanistic causes.

To expand on this, most current approaches address the symptoms of mental health disorders in the same

way you might take paracetamol to relieve the symptoms of a fever that is caused by an underlying illness. Fortunately for many illnesses we also have treatments that address these underlying causes – antibiotics, for example. In psychiatry, in many cases, we do not really have the equivalent: we don't have 'psychiatric antibiotics' that target the mechanisms that give rise to the symptoms.

I think that understanding psychiatric disorders through the lens of predictive processing can offer us a better route to identify plausible mechanisms, but even so there will still be a long way to go before we have in hand effective treatments.

There are promising signs out there, but there is no panacea that will solve everything related with mental disorders. Not even psychedelics …

Question: To conclude, I would like to hear your perspective on the ongoing debate about free will. Some scientists, such as Robert Sapolsky, assert that free will does not exist, advocating for radical determinism. On the other hand, philosophers like Dan Dennett argue that we possess degrees of freedom. Where do you stand in this debate?

Anil Seth: I have a chapter in the book about it, and it was a really challenging chapter to write since it is the one thing I had not written about before.

The concept of free will started to make the most sense to me when I started to think about it in the same way that I had been thinking about every other kind of experience: as a kind of perceptual prediction.

Now, what does that means in terms of how the debate is usually phrased? You have people who – like Sapolsky – who come out very strongly with the view that free will does not exist – and they usually say this because of some suspicion that the universe is deterministic somehow, that its course is fully set. This is certainly the case for Sapolsky. But arguments like this are often guilty of strawmanning: arguing against a version of free will that isn't worth hanging onto anyway.

This undesirable version of free will is what philosophers would call it "libertarian free will". I am very unkind to it and call it "spooky free will" – which is basically idea that consciousness can swoop in and alter the course of physical events, a kind of strategically-savvy uncaused cause. This sort of free will makes no sense, but is not the sort of free will we need or should want.

The main problem with discussions on free will, as I see it, is that people tend to make assumptions that they do not realize are being made or presupposed in the first place. For me, the question you should ask is: is the brain a sufficiently complex system that it can react to situations in very flexible ways, that are not immediately determined by the environment? And the

answer to that question, for the organisms like us is: yes! We can indeed perform actions that have their causes that come more from within the body – and which trace back in time deep into our histories - then from the world, from the immediate situation out there in the environment.

We have many degrees of freedom, in a very strict engineering sense again. I can do many things, but I do one thing. And our highly complex brain controls these degrees of freedom, integrating many prior causes in very sophisticated and subtle ways into a single behavioural 'final common path'.

Alongside this, it is a fact that there are some occasions – when I do some actions – in which I feel an experience of volition and agency. Sometimes I experience an action as being freely willed, and other times I do not. Essentially, what I do is take the same perspective as for other kinds of perceptual experience and ask "Why does that experience feel the way that it does?" and "What is the point of that experience?".

Following this line of thought, one can understand experiences of free will as being like experiences of color. They do not reflect reality as it really is, but they are still very useful for the organism. The color red does not exist objectively, as a property of the world, but is a very useful thing for us to experience since it tracks a useful property of how objects in the world behave.

For me, experiences of free will track a very useful property of how us, complex organisms, behave. These experiences label actions that have a certain 'freedom from immediacy' to use a term from Mike Shadlen. And this is useful – not because experiences of free will actually cause actions – but because they allow the organism to learn so that next time, if things didn't work out, a different action might happen instead.

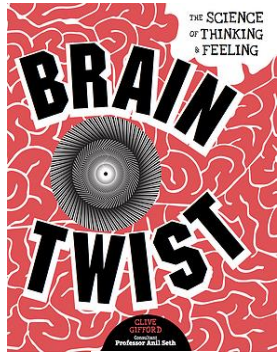
The mistake is to think that the experience of free will causes an action: they do not do that. Again, the experience of free will is useful because it labels particular actions as having their causes more from within the organism that outside. If you see it like that, the whole debate about determinism and stochasticity becomes largely irrelevant.

You can think of free will as a natural biological phenomenon, but it is real in the same sense and that experiencing colors is real, and it is useful in the same sense as an experience of color is useful.

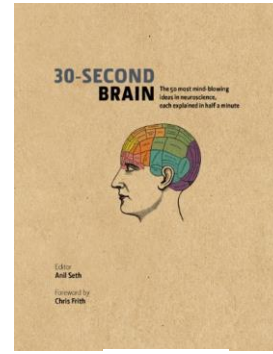
Thinking about free will in this more nuanced way, as something that comes in degrees, means that we can talk about degrees to which it varies, a discussion which becomes complicated when we talk about moral responsibility and so on, but that complexity goes with the territory.

Books by Anil Seth

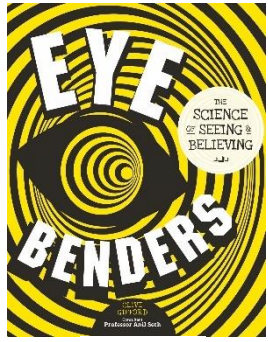
2021



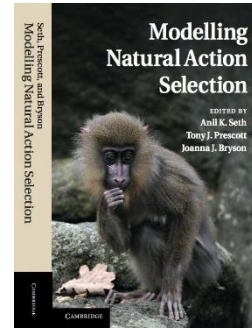
2015



2014



2013



2011

VII. Dialogue with Karl Friston

Karl J. Friston is Professor at the Institute of Neurology at University College London, a neuroscientist and an authority on brain imaging, having invented statistical parametric mapping, voxel-based morphometrics and dynamic causal modeling (DCM).

He was awarded the “Minerva Golden Brain Award” and was elected “Fellow of the Royal Society”. He received a “Medal College de France” (2008) and the “Weldon Memorial” prize (2013) for contributions to mathematical biology.

In 2016, he received the “Charles Branch” award for unparalleled discoveries in Brain Research and the “Glass Brain” award. Friston holds honorary appointments from the University of Zurich and Radboud University.

He has published several books, such as *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (2006), *Principles of Brain Dynamics: Global State Interactions* (2012) and more recently *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior* (2022).

More information: https://www.fil.ion.ucl.ac.uk/~karl/

Question: You are regarded as the world's foremost authority on the Free Energy Principle (FEP) and its various theoretical and empirical consequences. My first question is, in fact, a meta-question: how did the idea of the Free Energy Principle originate in your research?

Karl Friston: We can argue that the idea has been around since days of Plato, probably present in Kantian thinking, very nicely articulated by Helmholtz, with the notion of unconscious inferences.

But we can also find it in the ideas of Richard Gregory, like perception as hypothesis testing and formalizations of this, or by people like Peter Diane and Jeffrey Hinton, with the Helmholtz Machine, where they borrowed the mathematical notion of free energy from Richard Feynman.

Feynman was trying to use it to create an insoluble inference integration problem into an attractable optimization problem – technically, that is quite important, since it means that you can view the brain as an optimization machine, where one is trying to optimize its beliefs and then you can interpret the brain as an inference machine, a statistical organ.

Those ideas have been developing for ages literally, but certainly formalized by the work of people like Richard Feynman, and in terms of artificial intelligence, and machine learning, by people like Jeffrey Hinton and

Peter Diane. The generalization of these ideas to everything has a few steps though.

The first generalization would be to apply exactly the same ideas not just to sense-making an inference and measurement and observation, but also to say that our behavior and our actions can also be understood as trying to minimize free energy.

If we use, for example, machine learning as a kind of free energy – which is negative free energy – if I were talking to physicists, they would be talking about minimizing free energy or minimizing prediction error, minimizing surprise.

Following this we can think now about everything that we do is also trying to minimize surprise in the sense that we predict that we are going to move and then we do things with an ideomotor theory, we realize those predictions, and that can be understood as minimizing exactly the same free energy or prediction error or surprise that we use to actually make sense.

So, it is a long-winded answer to the question that I did not have the idea in the first place: I have leveraged the idea which I have been slowly formalizing, crystalizing over a long period.

Question: Following the Free Energy Principle, living organisms strive to minimize the entropy of their free

energy. In light of this proposition, do you believe there might be an issue when considering altered states of consciousness? These states often involve abnormally high or low entropy. Do you think such cases could pose problems from your perspective? I am considering, for instance, drug-induced states or other similar induced states.

Karl Friston: I do not think that is at all problematic. I think they are very important, very informative windows onto this view of the brain as trying to make sense of this world and to do so actively in the sense of active sensing. Both psychopathology and different states of mind that you have when taking certain drugs, – say psychedelic drugs – both reveal very important mechanistic aspects of this perspective.

You mentioned a sort of entropy. Entropy is just a description of a probability distribution, so it is very important to work out what probability distribution you are referring to. The free energy principle, when applied to self-organization of any sort – ranging from small particles through human beings –, in these cases, the entropy that people are talking about that is being minimizing is the entropy of the outcomes, it is the things that I actually experience. That would be the homeostasis perspective that you are just trying to keep your exposure to the environment in your exchange with the lived world within viable bound.

You are trying to minimize that entropy but, on the other hand, to do that, you actually have to maximize the entropy of your beliefs in accordance with things like Occam's principle: you want to find these very simple explanations without committing too much to a particular explanation. You can Immediately see that there is a dialectic between trying to do the best kind of inference, which would require maximizing the entropy, and the consequence of that, which is minimizing the physical entropy of my outcomes.

The same argument also applies to the entropy of your neuronal activity that is encoding your beliefs about the causes of your sensations: at the same time, you are maximizing the entropy of your beliefs, but you are also going to be largely minimizing the entropy of the neuronal dynamics, so they are minimally complex – you try to make it as efficient as possible.

I think one should be careful when talking about entropy, since there are different kind of entropies, such as the entropy of the general activity, the entropy of our beliefs, the entropy of the observations. But all of those different kinds of entropies are very useful ways, as you point out, of summarizing different states of consciousness.

The history in relation to psychiatric conditions, mind-altering drugs and indeed different altered states of consciousness that are physiological, like sleep states, for example, or dreaming, they all speak to one very

interesting aspect of this process of free energy optimization. Sometimes, with predictive coding, if you make some simplifying assumptions, you can read the "physicist" free energy as prediction error more precisely, a precision weighted prediction error.

In this case, you are not trying to minimize all prediction errors, just the ones that you think are very precise, informative and reliable. Once you think about the brain as in the game of being driven by the imperative to minimize precise weighted prediction errors, you realize that you have to not just make predictions about the content of the observations, but also the precision.

You also have to quantify your uncertainty, and what that means, from the point of view of a physiologist, is that you have to estimate the excitability of various neuronal structures in reporting prediction error to higher levels that are responsible for accumulating the evidence of doing the belief updating, and, very simply put, what that means is that the modulation of the excitability of the neuronal structures reporting prediction errors, reporting free energy, has itself to be predicted. And if you get that wrong, you will get some very odd beliefs and odd inferences.

That is important for two reasons: first of all, it links it to the psychological aspects of decision weighting, which would be attention. Attending to something is just basically allowing it to privilege those predicting errors, to have privileged access to belief updating, but

not those: “I am going to ignore that, that is not new, whether has low precision; it is false news. This is interesting. I am estimating this is important information or precise information. I am going to increase the attentional gain of these prediction errors and do my updating.”

From the pharmacologist point of view, the drugs that have these effects are exactly the same drugs that affect the excitability of that precision weighting. From all this, now you have got a nice link between the role of psychedelics like Psilocybin and or LSD and the visual attention accounts of why you get abnormal perceptual inference under those psychedelics. But you do not have to take psychedelics to do that.

You can look at exactly the same failures of neuromodulation as phenotypes of certain psychiatric conditions that will then lend themselves to false inference. And by this inference I just mean sort of classical type one and type two errors: inferring something that is there when it is not, which is just like a hallucination, or infer something that is not there when it is, which would be a dissociative syndrome or neglect syndrome. Both of which are features of many neurological and psychiatric conditions.

Question: My next question pertains to how we can apply this approach to mental disorders like

schizophrenia or autism. I believe the strength of your ideas lies in the ability to explain both the 'normal' or healthy brain and the ‘abnormal’ brain, altered states of consciousness, and even lucid dreams. Do you view this array of explanations as a sign that you are on the right path?

Karl Friston: Yes, lucid dream is an interesting one.

Question: I believe you have a paper published that addresses lucid dreams from your perspective of active inference and free energy, dated 2018, if I'm not mistaken.

Karl Friston: I just mentioned my friend and colleague, Alan Hobson, who I studied with a couple of years ago and he loved lucid dreaming, but he was also very anxious to point out that if you wanted evidence that the brain is a constructive organ in the sense that it generates fantasies, hypotheses for how the world works, then you don’t need to look any further than dreaming.

In dreams we have this neurochemical / pharmacological / neuro-modulatory shutdown of the precision of all our sensations, with the exception of the eyes, because the eyes are part of the central nervous system. And yet we still perceive, we still dream.

So, I think that that is a wonderful example of the brain as a sort of creative constructive active organ that just contextualizes itself by getting this precision weighting and this sort of gain control in the right balance.

Question: I have a philosophical provocation for you. In some paper, it seems that you are supporting an embodied or anti-representationalist view. But in other papers you actually seem to argue for a representationalist view. What is your real position on what is known as the "representation war" in Cognitive Science and Philosophy of Mind? Are you for or against representations in general?

Karl Friston: My answer is the following: it depends on who am I talking to. I heard a nice saying on an Australian sitcom the other day about being a person pleaser. So, if I am talking to a skeptical – somebody who is an anti-realist – then you can use the free energy principle to say: "yes, that is absolutely true!"

Why? Because you will never ever actually be directly exposed to what is out there, if there is anything out there. All you have is your sensorium. You are a brain in a box and everything that you believe about the world is just a fantasy. So, this would be very consistent with the skeptical approach.

On the other hand, if you are a realist or an externalist, then the active inference reading or application of the

free energy principle will celebrate that, because this is all about embodied active sensing: it is about how I physically engage with my world in order to get the right kind of sensations that make my internal beliefs most consistent with that world. So, you can literally read self-evidencing as gathering evidence from the real world for an I-model of that world. And if I am that model, I am gathering evidence from my existence, and that is one way of just looking at the existential imperatives, they are entailed by this kind of self-evidence.

But you brought up an interesting notion which is representationalism. If you consider the free energy principle, I can say definitively, irrespective of who I am talking to, that is quintessentially representationalist. Absolutely. That is the whole point of separating the brain from the world: that the world can now have beliefs about the world. And that that is just a probabilistic representation. So, the free-energy principle and active inference would not admit radical enactivism. But if it is a softer kind of enactivism, I think the theory can happily accommodating them.

Question: What consequences does the Free Energy Principle have regarding what philosophers truly seek answers to – that is, the existence of subjective qualities of experience, the so-called *qualia*? Does your approach contribute any relevant knowledge to address this philosophical notion?

Karl Friston: That is a very interesting and challenging question, which has become a recent focus among a number of my colleagues getting into the phenomenology of subjectivity. My best summary of current thinking in terms of people who do philosophy of mind and try to frame that within the free energy principle and active inference would be that to have a qualitative experience requires that to be a covert action. There has to be an internal action.

So, if I was a psychologist that would be the same as saying that I have to be able to attend to something. If I cannot choose whether to attend or not, then I cannot have the qualitative experience of that. Very often this is linked to Metzinger's notions of phenomenal transparency and opacity: for something to be rendered normally opaque – and I am now reading that opacity as sort of isomorphic with a quality of experience of a percept – then you need to have the ability to change the precision of the sensory evidence that matters, which is just the prediction error.

What that would say, if you are a neuroanatomist, is that to have qualitative experiences (or at least to have those kinds of percepts that can be rendered opaque) means that I should be able to find projections, anatomical neuronal projections in the brain that have a neuromodulator effect on some parts of the brain in a given hierarchy.

And, if I was a psychologist again, I would expect us to see those projections support the ability to endogenously attend to different parts of the sensorium. So, what you have got is a story which links the quality of experience with attention, with the active selection of sources of evidence for my fantasies and my belief updating. For me, the active selection is not sort of overt behavior.

You can certainly see parallels with, for example, Rizzolatti's premotor theory of attention, in the sense that acting on the world to get the right kind of precise information can actually involve overt action as I visually palpate the world. But if I internalize that, and now think about that palpation of the sensorium going on the inside by selectively various sources of prediction errors or sensory evidence by dating them with the precision control, that is a kind of mental action, that is a kind of covert action. That was, for me, an acid test for the quality of experience, or reading qualia as qualitative experience.

If you subscribe to that, there are a couple of other conditions that have to be satisfied before you can engage in that kind of mental action: you have to have a generative model that can generate the predictions of the precision that you are using as the basis of your mental action, which immediately tells you that you have got to have a generative model of the future. That tells you that the kind of creatures, or systems, or

particles, that could possibly have qualitative experiences, have to be those quite sophisticated creatures that have generative models of the future, because you need to have a model of the consequences of your action, even if that action is covert.

I think that is quite important. There is a temporal thickness that is entailed by this sort of mechanistic account of qualitative experience. There are other arguments which colleagues, in particular people like Maxwell Ramstead, Mark Solms, and other people, have recently written up that speak more to a more abstract formulation of sentience, thinking about experience being projected onto in the screens as a Markov Blanket, from the point of view of classical information theory, or as a holographic screen from the point of view of quantum information theory.

The basic idea is that maybe one possibly physically distributed system in the brain that can be further subdivided and the Markov blanket that surrounds this system – or if you took the view of a quantum information theorist, the holographic screen that contains the classical information that the internal aspects of this structure write to, or read from – separates that internal structure from the rest of the brain. There are arguments that claim that, to be conscious and to have conscious experience, it is required the existence of this unique Markov Blanket or

holographic screen, interestingly taking you much back to a Cartesian Theater again.

However, there is no essential dualism implied by this, but it brings you back, in an interesting way, to the notion of an internal screen. But crucially, the internal states that are observing this screen can never see themselves, because you cannot put in further internal streams. That story can be unpacked in the context of this attention and internal action or mental action story, simply because the only way that the inside can change the projections on this internal screen is by this mental action. But you always come back to this notion of mental action at the end of the day.

Question: To conclude, I'm not sure if you are aware of the new book *The Model of the Mind* by Grace Lindsay, a computational neuroscientist at New York University, who argues that your approach, based on the Free Energy Principle, is unfalsifiable from the point of view of the philosophy of science. She argues that it is not possible to understand, from a scientific point of view, whether your principle is true or not. What do you think of this criticism?

Karl Friston: That is absolutely correct: FEP is unfalsifiable. Why? It is very simple. The free energy principle is a principle of least action, like Hamilton's principle. In physics and in mathematics or indeed in

computer science, these kinds of principles are just methods or tools: they are not theories or hypotheses; they are just tools. The free energy principle is just a tool: it is neither right or wrong, you either apply it or you do not apply it.

The purpose of the free energy principle is, if you like, to be applied, to simulate or to reproduce intelligence, like sense making or sentient behavior. But sometimes it is also to match the simulations, to observe behavior, to actually reproduce observed behavior. It is a principle or a method that allows you to simulate or reproduce, or to realize sentient behavior of a basic sort.

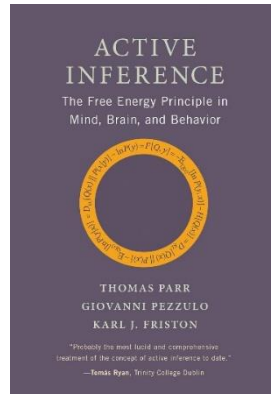
But in doing so, you now have to commit to applying it to a particular generative model. And at that point, if you are saying that this generative model is apt to explain this kind of behavior, this kind of creature or this particular patient, then that becomes a hypothesis and that is hardly falsifiable. What you do is that you apply the free energy principle to this hypothesis generative model, to that generative model, and to that other generative model; and then you actually use the energy principle to work out which is the best model for that particular context.

If you read the generative model that somebody applies the free energy principle to as a theory, then that is certainly falsifiable: there might be a better judging model that will explain this person's behavior or this

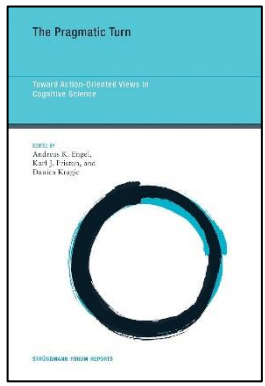
kind of creature's behavior. But the free energy principle itself is just a principle, it is like a T-Test, it cannot be falsified, you use it to do the falsification.

The free energy principle really provides you a theoretical formal framework in which you could understand the notion of falsification: it is just comparing evidence for the null hypothesis relative to the alternate hypothesis. You actually do it using the free energy principle. So that is why there is no problem and that is why the author is absolutely correct.

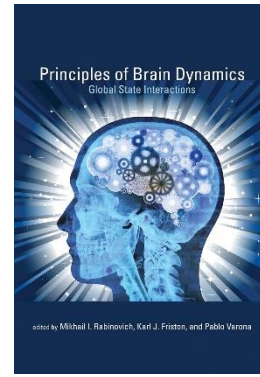
Books by Karl Friston



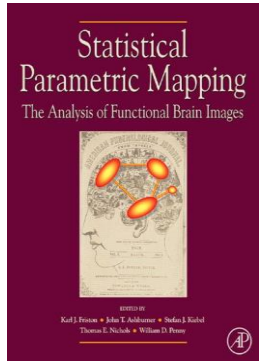
2022



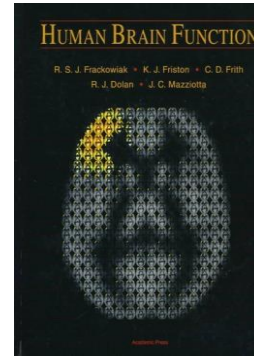
2016



2012



2011



1997

VIII. Dialogue with Christof Koch

Christof Koch is a neurophysiologist and computational neuroscientist best known for his work on the neural basis of consciousness.

He is the President and Chief Scientist of the Allen Institute for Brain Science in Seattle and is the Chief Scientist of The Tiny Blue Dot Foundation.

From 1986 to 2013, he was a Professor at the California Institute of Technology. Koch's main collaborator in the effort to locate the neural correlates of consciousness was Francis Crick, the Nobel Prize in Medicine.

More recently, Koch has worked closely with psychiatrist and neuroscientist Giulio Tononi and has been developing the Integrated Information Theory of consciousness.

Koch is the author of several books such as *The Quest for Consciousness: a Neurobiological Approach* (2004), *Consciousness: Confessions of a Romantic Reductionist* (2012) and *The Feeling of Life Itself - Why Consciousness is Widespread but Can't be Computed* (2019).

More information: https://christofkoch.com/

Question: Consciousness has proven to be a formidable challenge for contemporary science, prompting discussions on whether a redefinition of fundamental physical nature is necessary to provide a comprehensive explanation. In this context, what are your thoughts on the adequacy of current physics and physicalism in grappling with the intricacies of consciousness? Do you believe that our existing scientific frameworks need to evolve to better accommodate the nature of consciousness, or do you see potential within the current paradigms?

Christof Koch: It depends on what you mean by physicalism. I support a reformulation of IIT that claims that you cannot focus only on extrinsic causal power that is described by conventional physics, but you also need to study intrinsic causal power, because that is what consciousness is. It does not mean that we need a new theory of integrated quantum mechanics and gravity, unlike argued by Sir Roger Penrose.

We have to see where Physics evolves to, but right now, given the fact that the brain operates 300 degrees kelvin, I do not think that even with a reform – even once we have a complete single theory of physical laws –, that is not going to make a difference to the brain, since brains do not operate at that scale. I may be wrong, but that is my strong intuition.

Question: The Integrated Information Theory (IIT) is constituted by five axioms, with some authors criticizing the inclusion of the "Exclusion Principle", deeming it an arbitrary or *ad hoc* mechanism. Do you think that this exclusion principle is a weakness of IIT that requires refinement, or do you think that the fact that any conscious experience is defined by itself is enough to support the presence of the exclusion principle?

Christof Koch: I totally disagree with that criticism since I think that it is essential. Why? Because otherwise you get multiplicity of consciousness experiences, and also, by consequence, a multiplicity of causal powers. So, it is an essential part of the theory.

This exclusion principle was not added later, but was always part of the theory since ultimately, given all the combinatory possible, there is a very large number of possible mechanisms that exist for itself. IIT says that there is only one that exists, the one that is the maximum of all possible spatial, temporal and granularity.

That means that there is only one consciousness, there are not an infinite amount of consciousness in my head. Why are those particular spatial scale and temporal scale that are having that specific footprint over my brain? The answer for that question is the following: it is that specific one because that is the one that maximizes the intrinsic causal power.

If you had additional neurons, or if you had fewer neurons, it would have less causal power. If you look at a different time scale, it would have less causal power. That is the assertion of the theory: the theory might be wrong, but you cannot arbitrarily remove one axiom and still have the entire approach surviving.

Question: There has been a slight modification or update from the original version of IIT, now referred to as version 3.0 of IIT, where there is an attempt to redefine consciousness as the maximally irreducible cause and effect power of any network. Do you believe it would be more accurate to consider the theory as focusing on integrated causal powers rather than integrated information itself?

Christof Koch: I'm not sure, that seems a linguistic argument to me. Ultimately, it is about consciousness. Well, if I think a bit more, we can claim that it is both. The claim is that intrinsic causal power is what consciousness is.

As I said, it is an identity. At IIT, we labeled that integrated information as "phi", but the theory is more than just about phi, because any theory of consciousness not only has to explain how consciousness fits into the natural order of things but also why consciousness of pain is different from pleasure or why is different from seeing space.

You have to be able explain the phenomenology: you have to explain why particular states feel the way they do, and this theory does that in my opinion.

Question: Recently, there was a controversy in the scientific study of consciousness where several scientists and philosophers signed a public letter[77] arguing that IIT was a pseudoscientific theory, which was not falsifiable. What kind of experiments do you think could falsify your IIT theory?

Christof Koch: Many of them, actually. For instance, the theory makes very specific tests. For example, you can look at a particular network and question if it does maximize the intrinsic causal power. It is a very specific prediction: once you have the transition probability matrix of that particular neural network, in principle, you can compute it.

And you can ask: does it maximize the power? Does the phenomenology truly explain how? I briefly refer to this in the paper[78] on the phenomenology of the space: what about the phenomenology of time flowing, what about the phenomenology of color?

[77] Available here: https://osf.io/preprints/psyarxiv/zsr78.

[78] Paper published here: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0268577.

There is, currently, an active collaboration[79] for testing predictions between IIT and GNW: Global Neural Workspace famously says that the substrate of consciousness is the prefrontal parietal, and IIT says is all in the back of the brain.

So, this is being tested right now. IIT says you are conscious for as long as you experience: if you experience something for 10 seconds, there will be a physical substrate of that conscious experience for 10 seconds. Global Neural Workspace says "no, it is only there when you first send the broadcast, and then it disappears" – again there is a whole variety of ways that this theory can be tested, and is being tested right now.

For example, this perturbation complexity test method is a way to assess in a clinical context where the patient is conscious or not. That comes directly from the theory and it is being tested right now in clinical trials.

Question: That is a good sign, right? Since many theories of consciousness cannot be tested, then this seems to me to be an advantage of IIT over other theories of consciousness.

[79] For example: Melloni L, et al. (2023) "An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory", *PLOS ONE*, 18 (2): e0268577.

Christof Koch: Correct. It is essential, I would say.

Question: Regarding Artificial Intelligence, you argue that computers, at least those using a Von Neumann architecture, can never be conscious. However, would you agree that, if the intrinsic causal powers of the brain were reproduced on an artificial silicon substrate, for example, then we could assume that they would be conscious after all?

Christof Koch: That seems plausible: I usually say, there is nothing magical about brains. There is nothing supernatural about brains. In a typical transistor you have one transistor that is able to "talk" to two or three other gates. A typical neuron interacts to 50 000 other neurons.

Furthermore, there is huge overlap among these 50 000 neurons. So, two nearby neurons, one projects to 50 000 neurons, the other one projects also to 50 000 neurons, but they overlapped to a very large extent.

Again, this is all very different from the connectivity, which ultimately determines their causal power. If you build that into a hardware – whatever hardware it is – then you may get consciousness, in my view.

But we have to be very careful here: intelligence is different from consciousness. They are really two very different things. There is no question that we are going

to get “artificial” intelligence, that is, general artificial intelligence. Because, ultimately, that is about doing. Consciousness is not about doing, is about being.

Consciousness is a state of being. Intelligence is, ultimately, about doing. And there is no problem with getting human or superhuman intelligence on machines, but that does not mean that they will be conscious at all.

Question: You mentioned a relationship between Artificial Intelligence and consciousness, stating that Artificial Intelligence is inevitable, but we are not sure whether it will be conscious or not. My question is: is it really possible for intelligence to exist without consciousness? Doesn't intelligence come after being conscious? How can Artificial Intelligence exist without being conscious?

Christof Koch: That is a good question. The only two instances right now, at least in the present, of true flexible intelligence, is human intelligence, and we are indeed conscious. However, many of us believe that it is just one way to become intelligent: it is a way evolution has “chosen” and there are other ways, maybe through silicon and software – I see no evidence of that so far.

Let us put it differently: I can easily imagine how powerful computers, Turing machines, universal Turing Machines, that can be as powerful or more powerful

than I am, certainly much faster than I am, that are intelligent or even super intelligent. I do not see a priori why consciousness is necessary for that.

If you thing about deep machine learning, if you think about generative networks, transformers networks, etc., they seem to be doing extremely well, they seem to scale very well; the bigger the scale, the better the performance of them. I do not see why consciousness seems to matter for them.

Now, who knows whether they are going to reach stumbling blocks in a near future. My supposition for now is that we can get to artificial general intelligence – including super intelligence – without that involving necessarily developing consciousness.

That is not the way evolution has done it but now, as humans, we can do things differently from evolution, and we appear to be doing things differently. Whether this is good for us, is an entirely different question.

Having these super intelligences is actually good for Homo Sapiens-Sapiens? I am very skeptical about that, but that is the voyage we are on now, for better or worse.

Question: There are several methods for studying the conscious brain. We can focus on disorders of consciousness, or methodologies based on

psychedelics, but we can also focus on brain stimulation, among others. Which of these methods do you think is the most promising for studying the nature of consciousness?

Christof Koch: I would say all of them. There is none in specific, since that depends on the history of the researcher or the scientist, and the expertise that they posse. All the different methods have drawbacks and advantages. Some of them are more objective; with some, you can do first-person while with others it is more difficult to do first-person.

You can actually ask people relevant things when you study them but, on the other hand, you cannot intervene in their brains as you can do in the brains of animals. So, each one of these techniques comes with advantages and disadvantages.

It really depends on what are you interest in, what is your background, where do you want to study consciousness. Psychedelics, for instance, have this advantage that they can massively impact your consciousness. They also have drawbacks: most are illegal, and the big thing is that we do not know a lot about them.

There were many studies with psychedelics in mice and rats, but it is very difficult to really know what they are experiencing when you give them psilocybin or other

substances. So, it really depends on your background and your professional interests.

Question: In the context of predictive processing theories, there is an ongoing discussion about the role of 'top-down' influences on the brain. 'Top-down' processes involve higher-level cognitive functions influencing lower-level sensory processes. How do you perceive the relevance of these 'top-down' influences in understanding the nature of consciousness? Do you see them as crucial components in the formation of conscious experiences, and if so, how might they contribute to our overall comprehension of consciousness?

Christof Koch: In most people – leaving aside schizophrenics – what you perceive is fairly stable: you can stare as these visual illusions for many minutes, and you will keep seeing the same thing.

That tells me that the influences from top-down from my expectation are not that strong: it is definitely there, there is no question about it, particularly when I have very little time, when I have 100 milliseconds and I have to report what I see under those conditions, top-down is more important.

But, in general, I think people over emphasize the role of top-down influences. I can look at a wall and I see the wall. There is nothing there to predict, nothing changes,

I just see a blank empty wall. That tells me that top-down is not as important, at least under these conditions, as people think it is. And I do not think it has anything to do with predictive coding in this case.

Question: What ethical consequences do you anticipate arising from IIT? Particularly in its association with panpsychism, positing that all particles possess some level of consciousness. Consider the implications for Animal Ethics, where the theory suggests that many animals, especially those with developed brains, are capable of experiencing pain and pleasure. Do you believe that IIT, with its theory of consciousness, carries direct ethical implications?

Christof Koch: Yes, absolutely. Many years ago, exactly because of that, I turned into a vegetarian. And, in fact, I do not even kill bees: I try not to kill insects anymore because they too feel. I feel very strongly about this, that they also feel something. But they do not have a voice.

A bee, for instance, does not have a voice and a head like we do. Clearly, her brain is much simpler than our brain, but she too feels happiness when she is just drinking some golden nectar and flying in the warm sun.

And, just like we, she too is bookended between two eternities, at the beginning and at the end of their life,

so we are all thrown into this universe and we should minimize the suffering of all creatures including creatures like bees and similar. So, yes, I think IIT has definite ethical implications that we need to take seriously.

Question: With all the knowledge you have accumulated in brain science, could we say that you are in an advantageous position to correct, for example, a bad habit you have in your daily life? Does this incredible and complex knowledge give you a gateway to any bad habit you might want to correct, or does this knowledge give you no real advantage in your everyday life?

Christof Koch: No, nothing at all. I have to struggle the same struggles that everyone else have. I am trying to drink less and be more compassionate with others, and it is just wisdom tradition.

I happen to be a big fan of the stoic Marcus Aurelius, and what I can say is that it is discipline and constant mental exertion. IIT does not give you any superpower so you can say "Oh, now I get it! This is how the brain works; therefore, I can magically get rid of all my bad habits!". No, it still requires a lot of mental discipline and will power, causal power to actively avoid those things that I have determined.

In fact, this is where the true freedom lies in the theory. The theory also makes some implications about free will: it claims that free will exist in the original sense; in the sense that I can make difference to me, “I am the master of my faith, I am the captain of my soul”.

And the way I do this is that I think about it, I reflect upon it, I come to a particular conclusion that this is a bad habit or that is a good habit, so this I what I want to reinforce and that is what I want to avoid.

It takes consistent discipline, day in, day out, and that of course will change ultimately my brain, and therefore my life. Hence, I have the freedom to determine my own fate, given the background conditions about which, of course, I can do much less.

Question: What is your opinion on the current trend of incorporating "spiritual" aspects into science, often involving concepts like quantum synchronicity and quantum healing? Do you consider these approaches to be questionable or even immoral? Is there any scientific evidence supporting the relevance of these ideas, or do they primarily gain popularity as marketable concepts that attract individuals lacking the capacity to grasp the underlying knowledge?

Christof Koch: There is no evidence, right now, that microscopic quantum mechanics, in particular entanglement – which is key to quantum computing –

plays any role in the brain; which after all is, by physical standards, a hard system.

If you look at quantum computers right now, the operating temperature of a quantum computers is 25 or 35 milli-kelvin, that is roughly 50 000 times colder than the human brain is.

Therefore, it is very unlikely that these effects play any role for the biophysics or the biochemistry of the brain. Of course, many people sell you all sorts of things, like crystal healings, but that does not tell us anything, that is totally divorced from anything the brain does.

Question: I believe you engaged in an interesting debate with His Holiness the Dalai Lama regarding spirituality and the brain. Could we conclude with some key insights or agreements that emerged during the discussion? How did the intersection of scientific knowledge and spiritual perspectives contribute to a deeper understanding of the brain and consciousness, according to your perspective? Additionally, were there any specific aspects of the conversation that stood out to you or influenced your thinking in subsequent explorations of the mind?

Christof Koch: I went twice to India and met and debate with His Holiness, the Dalai Lama. We agreed in many things such as minimizing the suffering of all

consciousness creatures, unlike, for example, Catholicism (I grew up in a catholic faith).

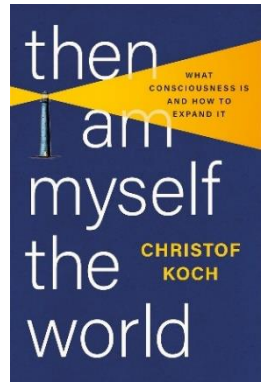
Buddhism, particularly Tibetan Buddhism, emphasize that consciousness is common to all creatures, not just humans who can speak about it. But we differed when we spoke about reincarnation.

The mantra I argued is constituted by four words: *No brain, Never mind*. Meaning that, once your brain dissolves, or dies, then the physical substrate of consciousness is gone and, in that sense, there is no more consciousness without having some kind of carrier.

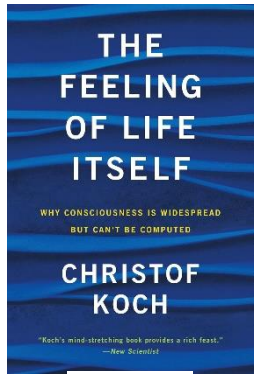
Without that carrier – even if it is exotic physics – there has to be some carrier of the brain, its memories and its trades. Typically, Buddhists will talk about the bardo, this liminal space between one life and their rebirth in the next life.

Again, if there is anything from my life that I carry in the next life, there has to be a place in space and in time that carries some substrate of my memories. Otherwise, I do not think it can exist. After hearing my argument, he just laughed with his deep belly laughs and just said “well, we will see”. We did not say anything further.

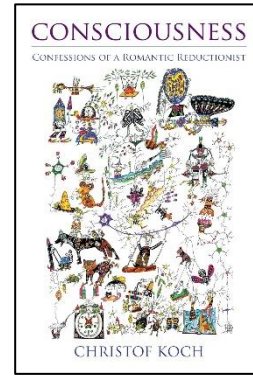
Books by Christof Koch



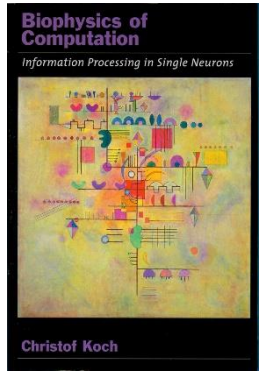
2024



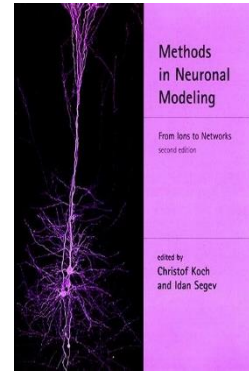
2020



2017



2004



2003

IX. Dialogue with Joseph LeDoux

Joseph LeDoux is the Professor of Neuroscience, Psychology, Psychiatry, and Child and Adolescent Psychiatry at New York University. He is the Director of the Emotional Brain Institute and the Nathan Kline Institute for Psychiatric Research.

He is vice director of the Center for Language, Music and Emotion at Max Planck-NYU and a member of the United States National Academy of Sciences.

LeDoux is the author of several important books about the brain: he is the author of *Deep History and Four Realm*s (2023), *Anxious* (2015), and *The Emotional Brain* (1996), among many others. Professor LeDoux is also a singer and songwriter for the folk-rock band *The Amygdaloids* and the acoustic duo *So We Are*.

He has won several awards such as the "Distinguished Scientific Contributions to Psychology" by the American Psychological Association (2010), the "Karl Spencer Lashley Award" by the American Philosophical Society (2011) and the "Gantt Medal" by the Pavlovian Society (2012).

More information: www.joseph-ledoux.com/.

Question: You are one of the foremost experts in the study of emotions and the mechanisms through which our brain, in conjunction with the body and the environment, generates specific behaviors. However, you express skepticism regarding the utility of vernacular or folk concepts related to emotions. Could you elaborate more specifically on how you approach and deal with this issue?

Joseph LeDoux: That is not exactly my position: I think folk language has an important role when we talk about mental states, since our mental lives are basically living in folk language. But where I draw the line is when we apply that language to behavior.

For example, the human brain can respond to a dangerous stimulus with a reflex or you might have a more complicated fixed action pattern, like with the freezing response. You might also – at a higher lever – learned a habit, so you can respond in that way, or you might even have a kind of goal directed cognitive model of the situation and respond that way, and all of these options happen unconsciously. But then, finally, you could also respond consciously.

We cannot take a simple behavior like freezing or fleeing and say that it is a pure indicator of fear, since there are many behaviors that fear can be expressed by and not all of them are in terms of conscious experience of fear.

Question: In your research, you have delved into a specific brain region known as the amygdala, responsible for detecting and responding to threats. However, you posit that labeling this anatomical organ with the emotion of "fear" is not appropriate. What prompted you to advocate for a conceptual shift from the conventional neuroscience term "fear circuit" to what you term the "defense survival circuit"?

Joseph LeDoux: Because I do not think fear is bubbling up out of the amygdala. Consider the defense survival circuit concept in general: every bilateral animal – that means every vertebrate that has ever lived – has to have had a circuit that can detect and respond to danger. It is a very primitive kind of thing and, layered on top of that, you can find the conscious experiences we have in which you know that it is you yourself that is having this experience.

I made a t-shirt out of it that says “no self, no fear”, meaning that, if you are not personally involved in a kind of autonoetic self-referential way, then there is no emotion. You have to be part of the experience in order for it to be an emotion: responding in a reflexive or instinctual way is not going account for an emotional experience.

If we see a dog hit by a car, lying on the side of the road, writhing in and in pain, we all project our emotions and feelings onto that dog. But what we are seeing is the

dog's reflexes when he is doing the twitching and growling and so forth. These are not indicators of pain.

I am not saying the dog is not in pain: instead, I am claiming that we have to draw a line and separate these automatic responses from responses that are associated with the conscious experiences of pain or fear, or whatever else you want to specify.

Question: This perspective appears critical of conventional approaches to scientific inquiry into emotions. For instance, in animal research, scientists often observe behavior and attempt to categorize that behavior under specific human emotions. Do you believe this method is flawed when investigating emotions in general?

Joseph LeDoux: This is an interesting point. I was writing about this a while back and I recently looked at a source that I used. It was a chapter of a book on anthropomorphism by Elizabeth Knoll in which she was writing about Darwin's perspective.

He lived in Victorian England, where anthropomorphism was in the culture, it was a kind of "way of life" at that point. Darwin was having a lot of trouble getting the theory of evolution/natural selection accepted by because of the religious implications and made an explicit decision talk about animal minds in human terms, rather than humans minds in animal

terms, because he did not think the latter would be well received.

Darwin's theories are really the starting point for the modern study of emotions, since his acolytes in the late 19th century viewed behavior as an ambassador of the mind. While that is certainly true to some extent, it is not a clear indication, because the behaviors we study in animals tend to be reflexive innate kinds of responses, and not responses that are necessarily products of the conscious experience of fear.

Free-wheeling attribution of conscious explanations without any evidence is what triggered the whole behaviorist revolution in psychology. You cannot say that "it looks conscious, therefore it is" unless you actually test that in some way. And it is very hard to do that in animals.

Sometimes, I am accused of denying animal's emotion but that is not true. I just think that it is methodologically very difficult to test that in an animal.

Question: You also argue that what is conserved by evolution is not behavior, as most scientists and philosophers claim, but rather the circuits linked to behaviors. Therefore, in your perspective, the same behavior can vary based on the particular circuit. Could you provide a suitable example of this distinction?

Joseph LeDoux: I think that it is not quite right. I am not saying that behavior is not involved. Behavior is a different level of analysis. The nervous system controls the behavior and the interaction between the animal's behavior and its environment is controlled by its nervous system. As species are evolving, changing and becoming other groups, other species, all of that is contributing in some complicated way to the behaviors.

But the point that you made about not being the behavior that was passed on; that was a point I made about tracing back to the beginning of life. The first cell that ever lived long enough to reproduce and give rise to other cells had to be able to detect danger and respond to danger in its environment.

Now, what that meant for a cell living 3.7 or 4 billion years ago is that it had to satisfy several key needs to stay alive: identify and turn away from dangerous (i.e. toxic) elements of the environment. But they also had to identify and incorporate nutrients and balance fluids and electrolytes. And for the spies to continue, replication was essential.

If a primitive bacterial cell encounters a high level of acidity, it uses its flagella to move away. Otherwise, it will not survive long enough to reproduce. These became fundamental physiological requirements of life, of all things that have ever lived. But each species solves the problem in its own way.

The point I am making is that what we have inherited through four billion years of life is not the behavior itself, but the requirement to respond to danger, identify nutrients, balance fluids and ions, and reproduce. These are survival requirements of any living organism, any living thing, whether it is a single cell or a whole gigantic organism, like us.

I was not trying to say that we have inherited our amygdala from bacteria, but what we have inherited from bacteria is the ability to survive, the necessity of being able to detect what is harmful and useful, and allow anything else that the animal or the organism has to do to survive.

I call these “survival needs” or “survival strategies” that have specific biological implementations in unique ways for each kind of group of animals and each kind of species of animal and, to some extent, each individual animal.

In short, all animals have these survival requirements, but only organisms with nervous systems have survival circuits.

Question: Contrary to many neuroscientists, you do not agree that emotions are biologically linked to the brain, but you argue that they arise from unconscious cognitive processing. Why do you think there is such a

divergence in views on the nature of emotions between you and many other neuroscientists?

Joseph LeDoux: It depends on who the neuroscientist is and what they are interested in. The cognitive theory of emotion is not new. William James had a version. But modern version goes back to Leon Festinger's theory of cognitive dissonance in the 1950s and Stanley Schachter and Jerome Singer's cognitive theory of emotion in the 1060s. I built on these in my research on split brain patient in the 1970s.

My mentor Mike Gazzaniga and I observed that when the right hemisphere of a split- brain patient would produce a behavior and we asked the left hemisphere "why did you do that", he would make it up, he would confabulate-- generate a narrative--to make sense of it. This is consistent with Leon Festinger's theory of cognitive dissonance which says that when you have discordant information, you have to resolve it some way.

Behaviors being generated from a non-conscious system – in this case the right hemisphere – would be a source of stress or anxiety since we all believe we have free will, whether we do or not, but we believe we do.

If our body is producing behaviors that we are not in charge of, it is very disturbing and we have to get some way to get around that. Michael Gazzaniga and came up with the hypothesis that maybe emotion systems are

ones of the systems in the brain that produces these behaviors unconsciously that might require some kind of cognitive interpretation in order to make them fit in with the mental assessment and our self-scheme about who we are.

That is why I have turned to studies of rats, because I wanted to understand unconscious behaviors that in humans might triggers us to have cognitive interpretations.

The Schachter and Singer theory of emotion came right out of Festinger's theory of cognitive dissonance. It has evolved quite a bit since the 1960s when Schachter and Singer proposed it, but it is a very viable approach to emotion. To me, it is a much more realistic way to think about our emotions.

Imagine you are on a mountaintop and you backpack has fallen off the cliff, the sun is going down, you do not have any food, you do not have any water or warm clothes: you are now in a state of fear and anxiety about what is going to happen to you. But the amygdala system, the so-called fear system, or what I call the threat system, evolved as a predatory defense system.

There is no predator on the mountaintop that is making you afraid. Fear is not something that is hardwired to a certain kind of stimulus: it is an interpretation. You monitor body signals from your stomach that indicate you are low on energy supplies, or you are starting to

get thermoregulatory signals that you are not warm enough. You then start to worry that you going to starve, freeze, or dehydrate to death. You can have fear for all kinds of reasons in life that have nothing to do with predators.

Question: This is interesting. How does your alternative view impact our understanding of anxiety disorders from a clinical standpoint, and what novel theoretical insights does it offer compared to more conventional perspectives?

Joseph LeDoux: We just published a paper in Molecular Psychiatry this week called "Putting the 'mental' back into 'mental' disorders".[80] The idea is that the entire approach of psychotherapy, the treatment of mental disorders that started in the 1950s, has been driven by a behaviorist agenda that marginalized the subjective conscious mind.

In the 1950s approaches like a behavior therapy and psychopharmacological therapy were become the standard. Behavior therapy obviously was straight out of Skinner, Watson and so forth, and paved the way for cognitive therapy, which started out with a mental angle (Arron Beck, the founder, was a psychoanalyst) but

[80] Original publication: https://www.nature.com/articles/s41380-021-01395-5.

slowly became more based on objective metrics and somewhat behavioristic.

But what about psychopharmacology? Who is working in these labs at psychopharmaceutical industries, testing animals to find out drugs that will help people? These researchers were trained by behaviorists in 1940s, 1950s, and 1960s.

The assumption was that, if you put a rat in a threatening or stressful situation of some kind, and you give him a drug that makes it less timid behaviorally, you are assuming that it is less timid because it is less fearful or anxious, and that when you give the drug to a human, the person will be less afraid or anxious.

But the pharmaceutical industry stared getting out of the anti-anxiety because it had failed to generate anything new. Most classes of medications were stumbled upon accidentally in the 1960s, like norepinephrine reuptake inhibitors or benzodiazepines.

The problem, in a nutshell, is that a medication that makes mice freeze or avoid less is not going to significantly alter human anguish- anxiety, worry and fear that a person experiences we need to help control the behavior physiology of the patient, but that is not enough, we also need to take seriously the subjective experience of the patient.

Question: Given the innovative nature of your work and its departure from traditional views on emotions, especially in attributing their origin to unconscious cognitive processing rather than biological links to the brain, do you see any parallels or influences from the theories proposed by Sigmund Freud? Freud, too, delved into the realm of the unconscious mind and its role in shaping human behavior and emotions. Do you acknowledge any resonance or departure from Freudian ideas in the development of your own theories?

Joseph LeDoux: What I would say is that the developments that were happening in the 1950s to create new forms of therapy were efforts to escape from Freud, because he was so subjective. However, in getting rid of the subjectivity that Freud brought into psychiatry, they threw the baby out with the bathwater.

Rather than saying "Okay, maybe the deep dark unconsciousness is not where everything is happening and it is not all sexual repression" and all of the other things that Freud was criticized for – but that does not mean that we should throw all of subjectivity out. You can't fix it all by changing behavior".

As I said above, even Cognitive Behavioral Therapy (CBT) has become a sort of behavior reporting approach. When Beck and Ellis started CBT in the 1950s and 1960s, there were more subjective element to it.

These were pushed out of the way later since the whole industry went towards insurance payments. Therapists had boxes to check, objective metrics, that you could use to identify or categorize what the patient's problem was. If you take the NIMH RDoc, for example.

It has all these metrics that you use to identify and classify things, and somewhere in the middle of the long list, you can find the verbal report—that is the closest thing to the subjective experience of the patient – but verbal report is just like any other symptom.

But why do people go to a therapist in the first place? Because they feel bad and they want to feel better. So, in my view, we have to fix the way they feel, and not just the way they behave.

Question: Consider the thought experiment in which you are tasked with programming a new type of human being from scratch, including designing their survival circuits. Do you believe it is possible to program these circuits in a way that directs the individual to seek happiness, goodness, or pleasure?

Joseph LeDoux: I definitely think that we could use a new form of human programming, that is a good idea because we are not in a good way as a species now. If you are going to plan that kind of project, you always have to start with the positive and the negative and try to build from there.

But we have to ask ourselves, in terms of the basic mechanisms, when a rat is being reinforced with a behavior, is it pleasure that is doing that? The whole idea of pleasure centers in the brain with Olds and Milner in the 1950s came about by accident.

Olds went on to write the paper in science called "Pleasure Centers of the Brain" and I asked Peter Milner about this once, a few years ago while he was still alive: "What were you guys really thinking about when you were studying these pleasure centers?"

And he said: "Well, first of all, we weren't studying pleasure, what we were trying to do is find mechanisms of reinforcement in the brain". If you read the article that Olds have published, you will see that the word "pleasure" never appears in the article, it is only present in the title. The journal editors perhaps added the title to the paper, Olds got seduced by it and begin to promoting the idea of "pleasure centers".

Rory Wise then started saying dopamine is the chemical of pleasure, but very soon both Wise and Olds rejected the idea that they were studying pleasure and that the dopamine was the chemical of pleasure. But once the cat is out of the bag with those things, it never goes back in.

So, now everyone talks about reinforcement in terms of reward and pleasure, but ultimately what we are talking

about is the effects of dopamine on cellular activity that causes a behavioral change.

That does not mean that an animal is not experiencing pleasure, or that the person is not experiencing pleasure, but that is not the same thing as the reinforcement of a behavior.

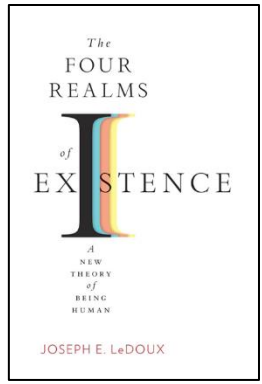
Question: In conclusion, I'm curious to hear your perspective on a question that has preoccupied philosophers for some time. Do you believe we are making progress in understanding how *qualia* are formed or sustained in the human brain, or do you think we are still far from providing a solid answer to the hard problem of consciousness?

Joseph LeDoux: One answer I would offer is that scientists need to be careful about which concepts they import from philosophy. Philosophy is a system of rules and reasoning which is great, but it does not mean that is the way the brain actually works.

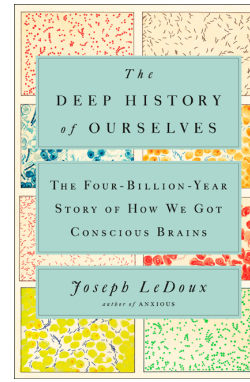
I'm not saying that *qualia* do not exist, but I think the so-called "hard problem" was created in a way that it can never be solved because it assumes mind-body dualism... That is okay philosophically, but to have a dualist perspective on the brain is a non-starter scientifically.

I think most neuroscientists are materialists and we want to find some mechanisms that makes qualia happening. I do have a new book out titled *The Four Realms of Existence: A New Theory of Being* where I propose a possible way to think about *qualia*, so maybe the reader can give it a try there.

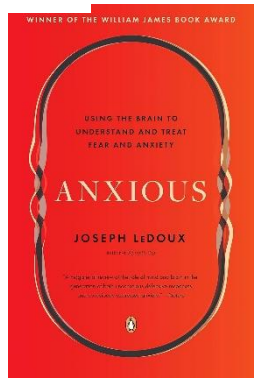
Books by Joseph LeDoux



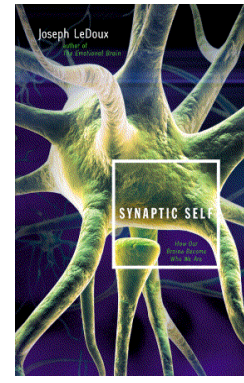
2023



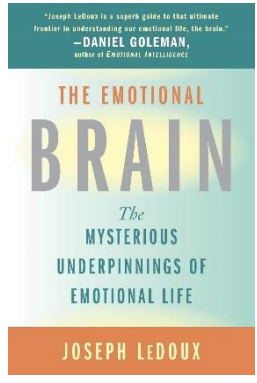
2019



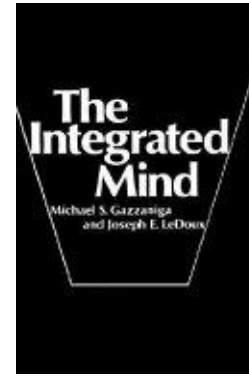
2015



2001



1996



1978

Conclusion

This book sought to introduce the reader to some of the most current and fascinating debates in philosophy and the science of the mind. Just like the fearless Portuguese navigators who dared to cross Cape of Storms, we faced the complexities of the mysteries of the mind and the countless navigational difficulties it poses.

The Cape of Storms is situated in Cape Town, South Africa, and stands as one of the southernmost points of land on the planet. Due to its unique location, it was deemed impassable due to the frequent, intense storms and challenging navigational conditions.

Until 1488, no human had successfully navigated around this cape. Everything changed with the Portuguese navigator Bartolomeu Dias who, for the first time, managed to cross the Cape of Storms, which would be renamed by King D. João II as the Cape of Good Hope.

This achievement marked the establishment of a direct sea route between the Atlantic Ocean and the Indian Ocean, connecting the West and the East, particularly

with India, an event that would forever change the economic, cultural and political history of our world.

Like Bartolomeu Dias, I hope that some of the progress presented in this book will continue to develop more robustly in the coming years, hoping that we will be able to transform the Storms of the Mind into a Cape of Good Hope.

Each page of this book is like another day on the high seas, in pursuit of a land that may never come. Perhaps the curiosity within each of us will be the wind propelling the forthcoming discoveries emerging from the shadows ahead.

Was it worth it? Let us trust the words offered by the Portuguese Poet:

"Was it worth it? Everything is worth it
If the soul is not small.
Whoever wants to pass beyond Bojador
Must pass beyond pain.
God gave danger and the abyss to the sea,
But it was in it that He reflected the sky."[81]

[81] Fernando Pessoa, Second Part: X. Portuguese Sea from the book *The Message*. Original version in Portuguese: "Valeu a pena? Tudo vale a pena / Se a alma não é pequena./ Quem quer passar além do Bojador / Tem que passar além da dor. / Deus ao mar o perigo e o abismo deu, / Mas foi nele que espelhou o céu.

May this *Odyssey of the Mind* offer the reader a safe point in the harbors of understanding and reflection. With the horizon in sight, I hope that each of you find in this book a reliable compass for your own explorations and that the wisdom acquired on this expedition illuminate paths to unknown futures, where the frontiers of the human mind can be further explored.

To close this conclusion and the book, a less poetic note (again). This book was inspired by several online courses that I organized as the Main Professor over the last 2 years. In this context, I had the privilege of reaching more than 600 students from around 35 different countries.

To all of them, I am deeply grateful for teaching me much more than I could have conveyed. Furthermore, the insightful curiosity of the students who, with their questions and interventions, inspired many of the debates presented here and made the courses much more stimulating and fascinating for everyone, especially for me.

May the *Odyssey of the Mind* continue!

Acknowledgements

Marcus Aurelius, Stoic philosopher and Roman Emperor, starts his famous book *Meditations* by thanking all the people who crossed paths in his life and from whom he learned something particular. Now, a work of this extent could not happen without the direct and indirect contribution of a hugely varied number of people, which is why I follow in the Emperor's footsteps by thanking all of them.

I begin, of course, by express gratitude to the 8 incredible scholars who gave me the honor and confidence of sharing the stage for this book: David Chalmers, Susan Blackmore (and Alison Seldon), Sir Roger Penrose (and Helen McGregor), Nicholas Humphrey, Joseph LeDoux, Anil Seth, Christof Koch and Karl Friston, a huge thank you.

In the execution of this book, Tássia Vianna was fundamental in transcribing the oral debates in English, which were then transformed into written dialogues that were approved by each of the contributors. Ana Monteiro was instrumental in the creative production of the cover of this book, whose faces were hand-drawn due to her pure talent and creativity. To both, a sincere appreciation.

To colleagues at the University of Porto, especially Professor Sofia Miguens, Coordinator of the Mind, Language and Action Group research group where I carry out my research, who has been the greatest influence in thinking philosophy in an open way to the world, and to Professor José Meirinhos, Director of the Institute of Philosophy, and the Dean of the Faculty of the University of Porto, Professor Paula Pinto Costa, for supporting the various projects without placing any obstacles. To Isabel Marques, science manager at the Institute, thank you for all your patience.

Next, I would like to thank to all those who have contributed to my ideas reaching the world, serving as a constant stimulus in my research.

In Brazil, Gabriel Mograbi and Paulo Taddei, from the Federal University of Rio de Janeiro; Nythamar de Oliveira, from the Pontifical Catholic University of Porto Alegre; and Maria Luiza llenaco, together with Osvaldo Pessoa Jr., from the University of São Paulo.

In Romania, Florin Piscociu and his team at Mindlifeline (Léa Chibany, Luana Aldea and the rest of the members) have been a constant support, whose work ethic inspires me every day.

In Malta, Ian Rizzo, Valdeli Pereira, Francois Zammit and the rest of my friends at the Philosophy Sharing Foundation have shown me the importance of a public philosophy, unconfined in the ivory tower and

accessible to all citizens. In Cyprus, Panayiotis Stavrou's enthusiasm and thirst for knowledge is a constant motivation.

In Chile, the friendship and admiration for the work done by Ricardo Ramirez, Camilo Garcia, Marcelo Martín, Nicole Nakousi, David Araya, and other colleagues, who gave me the privilege of a lifetime of receiving, at the age of 30, the distinction of Honorary Professor at the Faculty of Medicine Andrés Bello, a distinction that I hope to honor throughout my life.

Within the scope of my project on Ethics of Artificial Intelligence in Medicine, several colleagues should be nominated: Pekka Mäkelä, Raul Hakli and Pii Telakivi (Uni. Helsinki); Radu Uzskai (Uni. Bucharest); Simona Tiribelli (Uni. Macerata); Heidi Mertes (Uni. Ghent); Inês Dutra (Uni. Porto), Sabina Leonelli (Uni. Exeter) and, more recently, Antonio Chella, from the Robotics Lab, and Antonello Miranda, from the Advanced Studies Centre, both at the University of Palermo, where I had the privilege of being a Visiting Fellow during the summer of 2024.

To my friends at the Catholic University of Braga, with whom I had a short but intense collaboration, whose focus on multidisciplinary was an encouragement: a thank you to Augusto Soares da Silva, Carlos Morujão, Padre João Onofre, Paulo Dias, João Duque, António Melo, Bruno Nobre, Elton Marques, Ângela Leite, Tânia

Oliveira and Director José Lopes – thank you for welcoming me into your House.

To the philosopher Manuel Curado (Uni. Minho) and the neuroscientist Georg Northoff (Uni. Ottawa), supervisors of my PhD, a thank you for all the guidance in this still brief, but energetic academic career.

To the doctoral students that I have/had the privilege of supervising and trusted my guidance, I am grateful to Jaroslav Malík, Romeu Ivolela and Maria Luiza Ienacco for all their collaboration.

Randomly, to all colleagues spread across the various institutions who have valued sharing knowledge: Klaus Gärtner, Robert Clowes, Glorida Andrada, Rui Vieira da Cunha, Sâmara Costa, Sara Fernandes, Bárbara Sousa e Brito, James Grayott, João Cordovil, João de Fernandes Teixeira, Leonel Moura, Ralph Bannell, Roberto Pereira, Diogo Gurgel, Diana Tavares, Dina Mendonça, Joana Rita Sousa, Tomás Magalhães Carneiro, Ângelo Milhano, Björn Lundgren, Jorge Gonçalves, Luísa Neto, Peter Singer, Luiz Meirelles , Noam Chomsky, Yinchun Wang, among many others whose neurons in my hippocampus do not allow me to remember, but which always appear in my mind.

To conclude this already long note of thanks – I owe a lot to many people –, Marcus Aurelius attributes to the gods “the fact that my constitution has survived this kind of life for so long”: I can only thank my great friends

and family, who have helped me in this “kind” of life dedicated to knowledge: Mr. Joaquim, Maria José and family (Ana, Manuel, João and Maria Augusta); Rafael, Pedro, João Pedro and Carolina; brothers Kevin and Dylan; Carlos and Zaida (the true responsible for this “constitution”); grandmother Arminda, cousins Sandra, Nuno, António, Ana and uncles and aunts; Monica and Kiara; Isabel and Isaura; Roberta and Bruno; Ana Maria, Gilberto and Ana Mafalda; Luna and the cats Miró, Tareco, Mia, Didi, Manchinhas and Amarelinho (creative names, I know).

To All, a profound thank you.

nosce te ipsum

Author Biography

Steven S. Gouveia (Sion, Swiss, 30/09/1992) is a Research Fellow at the Mind, Language and Action Group of the Institute of Philosophy of the University of Porto (Portugal), funded by the Science and Technology Foundation (CEECIND.02527.2022), where he leads a 6-year project on the Ethics of Artificial Intelligence in Medicine, which has partnerships with the Universities of Exeter, Yale and Helsinki (cf. https://trustaimedicine.weebly.com/).

He graduated in Philosophy and at the age of 22 began his PhD in Philosophy of Mind at the University of Minho (Portugal), under the supervision of the philosopher Manuel Curado, having been a visiting researcher (2017 and 2019) at Minds, Brain Imaging and Neuroethics at the Royal Institute of Mental Health, University of Ottawa (Canada), under the supervision of neuroscientist and psychiatrist Georg Northoff.

After finishing his PhD (2021), he was a Researcher at the Center for Philosophical and Humanistic Studies at the Portuguese Catholic University, in Braga (Portugal).

Since June 2023, he has been Honorary Professor at the Faculty of Medicine of the Andrés Bello University, Viña Del Mar, Chile, a title awarded on the same occasion as Sir Roger Penrose Nobel Prize. He is also an Ethics

Consultant for an Effective Altruism company "Altruistic Careers".

He has published several academic books. In 2016, he edited the book *Philosophy and the Arts* in Portuguese, having invited some of the most relevant artists in national artistic practice, such as the artist Joana Vasconcelos, the rapper Valete, the musician Fernando Ribeiro (Moonspell), Leonel Moura (IA), the poet Ana Luísa Amaral, among others.

In 2017, he co-edited the edited book *Thinking Democracy* in Portuguese with the preface by Noam Chomsky and, with Ana Figueiredo Sol, he edited *Bioethics in the 21st* in Portuguese. That same year, he also edited *Philosophy of Mind: Contemporary Perspectives*, his first international book with Manuel Curado by Cambridge Scholars Publishing.

The following year, he published his first authored book in Portuguese with Editora Húmus titled *Philosophical Reflections: Art, Mind and Justice*, with a preface by the philosopher of mind João de Fernandes Teixeira.

In 2019, he published three edited books: *Perception, Cognition and Aesthetics* and *Film and Philosophy: Bridging Divides*, by the influential publisher Routledge and, for Vernon Press, he co-edited, with Manuel Curado, *Automata's Inner Movie: Science and Philosophy of Mind*.

He also published the edited book *The Age of Artificial Intelligence: an Exploration* by the same publisher, which features the participation of some of the most influential thinkers and transhumanists, such as Daniel Dennett, Ben Goertzel (creator of the famous robot Sofia), David Pearce, Natasha Vita-More, Roman V. Yampolskiy and Vernor Vinge, among others.

The following year, he published another edited book with Bloomsbury – the same publisher responsible for publishing Harry Potter – entitled *The Philosophy and Science of Predictive Processing*.

Furthermore, he published his most sought after authored book to date, in Portuguese: *Homo Ignarus: Rational Ethics for an Irrational World* with Editora Minerva, which has the preface by Peter Singer, a book focused on various ethical problems such as voting, humor, euthanasia, or artificial intelligence.

In 2022, he published his third authored book in English by the New York publisher Palgrave, part of the Springer Nature group, with the title *Philosophy and Neuroscience: a Methodological Analysis*.

Still in the same year, he published an edited book titled *Artificial Intelligence: Conversations about the New World* with the participation of several scholars such as Peter Singer, Wulf Loh or Sabina Leonelli. In 2023, he co-edit again a book with Manuel Curado entitled *Predictive*

Minds: Old Problems and New Challenges by Vernon Press.

In addition to this intense academic production, he is the host and producer of the international documentary "The Age of Artificial Intelligence: a Documentary", which features the participation of international researchers and is accessible for free on YouTube, with more than 55,000 views.

He was a speaker at dozens of peer-reviewed scientific conferences in Portugal and abroad. Furthermore, he has been invited as a speaker at universities around the world, in Brazil, Argentina, Chile, Malta, Italy, Czech Republic, South Korea, Cyprus, Romania, among others. Moreover, he was a speaker at the Science of Consciousness Conference for two occasions, the most relevant international conference on consciousness studies. He was also a speaker at TEDx NOVA, in September 2023.

He has been requested for various interviews, having participated in television and radio programs such as "A Minha Geração" (RTP3), "Sociedade Civil" (RTP2), "Linha da Frente" (RTP1), "Muito Barulho para Nada" (RTP2), "Filosofia na Rua" (Antena 2), in addition to several podcasts, such as "45 Graus", "Despolariza", "Pergunta Simples", "Desassossego", "Smart Meat" among many others.

He is the main professor of several online courses on topics such as democracy, conscience or ethics, with the participation of guest professor such as Sir Roger Penrose, Peter Singer, Noam Chomsky, Slavoj Zizek, Paul Bloom, among others.

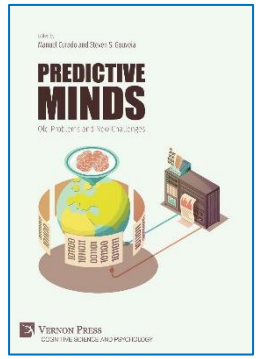
He was the founder and main editor of "Apeiron – Student Journal of Philosophy" (2012-2016), which included the participation of scholars such as Noam Chomsky, Peter Singer, Daniel Dennett or Noël Carroll.

Finally, he has been the main organizer of several international conferences, bringing to Portugal academics such as Peter Singer (Princeton), Luciano Floridi (Yale), William Child (Oxford), Shaun Gallagher (Memphis), Dan Zahavi (Copenhagen), Karl Friston (London), David Papineau (New York), Tim Crane (Cambridge), among others.

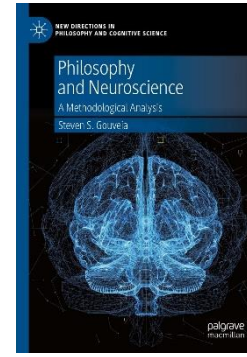
More information can be found at www.stevensgouveia.weebly.com.

Project funded by the FCT CEEC Individual Project 2022.02527.CEECIND by the Fundação da Ciência e Tecnologia, at the Mind, Language and Action Group, Institute of Philosophy (Instituto de Filosofia), University of Porto (Universidade do Porto), Address: Faculdade de Letras, Via Panorâmica s/n, P-4150-564 Porto, Portugal. The final revision of this English translation (from the original version in Portuguese) was also supported by a Visiting Fellowship at the Robotics Lab of the University of Palermo in 2024.

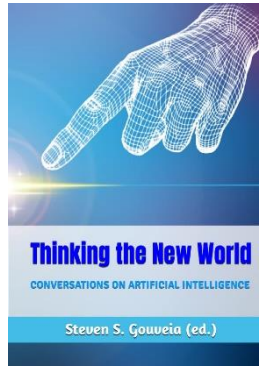
Books by the Author



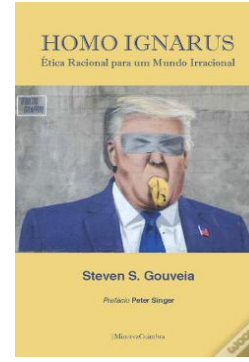
2023



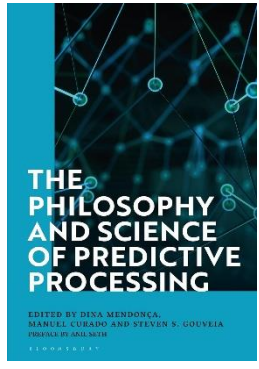
2022



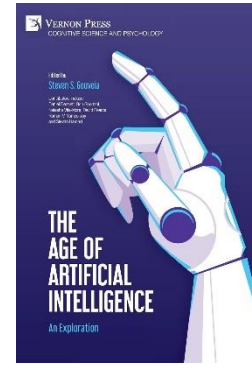
2022



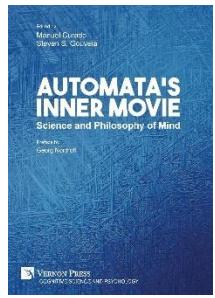
2020



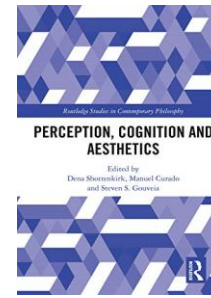
2020



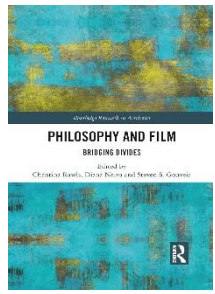
2020



2019

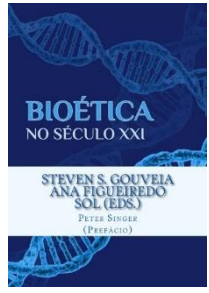


2019

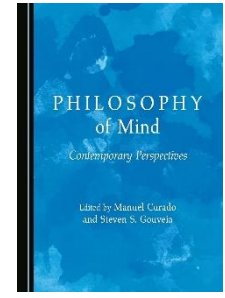


2019

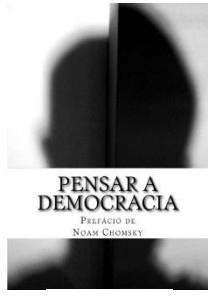
2018



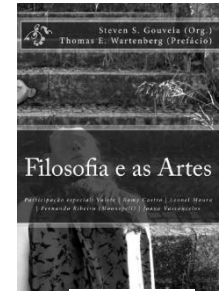
2018



2017



2017



2016