

Digital Sentience

Recognizing sentience in machines

Manu Herrán



International e Symposium on Research, Innovation and Entrepreneurship 2020
D Y Patil College of Engineering (Formerly University of Pune) Akurdi, India
www.dypcoeakurdi.ac.in

DYP

D. Y. PATIL COLLEGE
OF ENGINEERING,
AKURDI



[linkedin.com/in/mherran](https://www.linkedin.com/in/mherran)



Manu Herrán

Lead Researcher en Sentience Research

Madrid y alrededores, España · Más de 500 contactos

[Únete para conectar](#)

-  [Sentience Research](#)
-  [University of Deusto](#)
-  manuherran.com

Acerca de

Independent thinker and visionary of new technologies, science, and society. Working with tools, systems, and methods that allow individuals to reduce suffering and enjoy greater control over their lives.



Lead Researcher

Sentience Research

sept. de 2019 – Actualidad · 9 meses

Sentience-Research is an association of researchers on sentience with the objective of develop and promoting professional and scientific research on sentience aimed to reducing involuntary and useless suffering. Sentience is a concept very close to the idea of consciousness, but sentience is, in our opinion, focused on what is more morally relevant. Sentience is concerned with studying experiences that can be positive or negative, such as pleasure and pain, suffering and enjoyment.



Associate

Organisation for the Prevention of Intense Suffering (OPIS)

mar. de 2016 – Actualidad · 4 años 3 meses

Madrid y alrededores, España

The Organisation for the Prevention of Intense Suffering (OPIS) is an international think-and-do tank developing new ways of effectively addressing the most urgent issue in our world: the intense suffering of sentient beings. Our activities include research and thinking about ethical principles and strategies for effectiveness, advocacy, and the development of projects and creative campaigns to spread compassion and prevent the suffering of human and non-human animals.

Founder (Science Magazine)

REDcientífica

sept. de 1997 – Actualidad · 22 años 9 meses

Founder of "REDcientífica" (SCInetwork), a digital magazine of Science, Technology and Thinking, with more than 500 authors, 5,000 documents and 50,000 monthly visitors, receiving several mentions and awards (Google, Society of Information)

Contact Center, IT & Business Consulting (Banking, Insurance, Telecom, Energy, Public Sector)

M2C Consulting

mar. de 2008 – feb. de 2014 · 6 años

Spain, UK, Poland

IT presales, IT consultancy and IT projects management with specialization in Contact Center, online businesses, customer relationships management, automation and process improvement in insurance, health, bank, telecom, legal, info, energy and public sectors.

17 years of experience in Spain, UK and Poland, plus several professional experiences in United Arab Emirates and Latin America.

Main achievements:

- EUR 1M of annual savings through automation and process optimization
- Open and develop a consulting business line with annual revenues of EUR 1M
- Improved customer relationship processes in various companies (telco, insurance, energy, health, services, and public sector)

Ph.D. in Customer Intelligence. Design of specific methodology for creating

I MIGHT BE WRONG

- ✓ I might be wrong and that's ok because a fundamental aspect of science compared to other ways of obtaining knowledge is the recognition of ignorance, in a permanent methodological skepticism that always takes into account the possibility of being wrong (both others and ourselves).
- ✓ Maybe machines can't feel
- ✓ If they feel, maybe there is nothing we can do about it
- ✓ In that case, I'm losing my time (and your time!)

EVIDENCE

- ✓ There are several ways to obtain evidence. Some better than others; some more scientific than others. But whatever method we use, we must always recognize the possibility of being wrong. Skepticism is a fundamental aspect of the scientific attitude, along with fairness (Impartiality) and honesty.

EVIDENCE

Very scientific

Repeated observations

Replicated experiments (induction)

Logical deductions (deduction, rationalism)

Observations (empiricism)

Interpolation, extrapolation

Authority (prestigious sources)

Consensus

References (sources, what some say)

Popularity

Intuition (reasoning difficult to explain)

Superstition

Revelation

Myth

Faith

Dream

Very unscientific

Scientific method



What are the foundations of the scientific method?

HONESTY

IMPARTIALITY

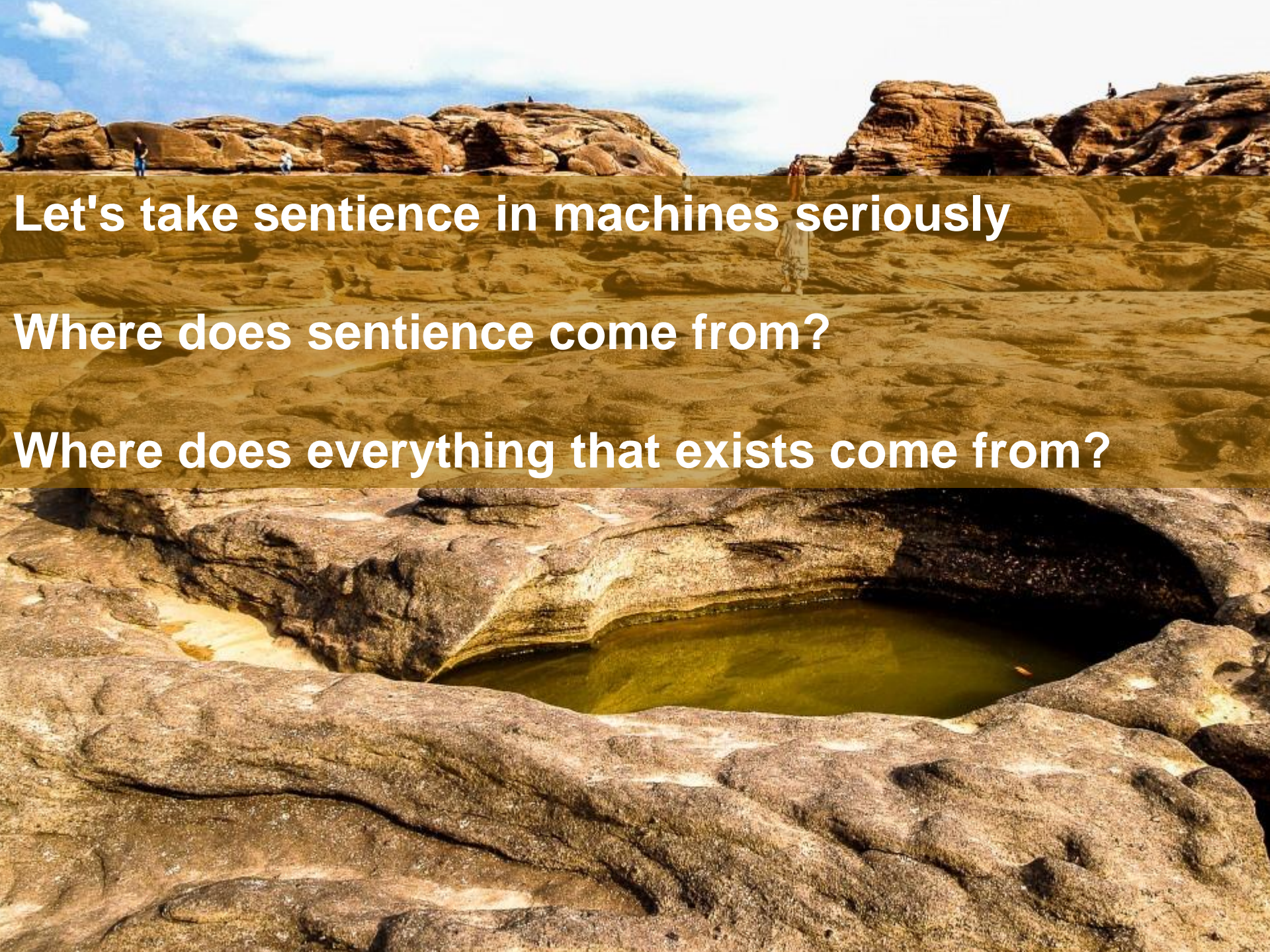
SKEPTICISM

I MIGHT BE WRONG

- ✓ Maybe machines can't feel
- ✓ Or if they feel, maybe there is nothing we can do about it

WHY IS RELEVANT

- ✓ I might be wrong, but the question of whether a hypothesis is probable or not is not the only factor to consider.
- ✓ We must take into account not only the probability of the hypothesis, but also the consequences that would result from it if it were true.
- ✓ Even if the probability of sentience in machines were extremely small, while there is a higher than zero probability, and considering that is not very clear where sentience comes from, we might think twice before disregarding this idea, because if true, its implications would be immense.

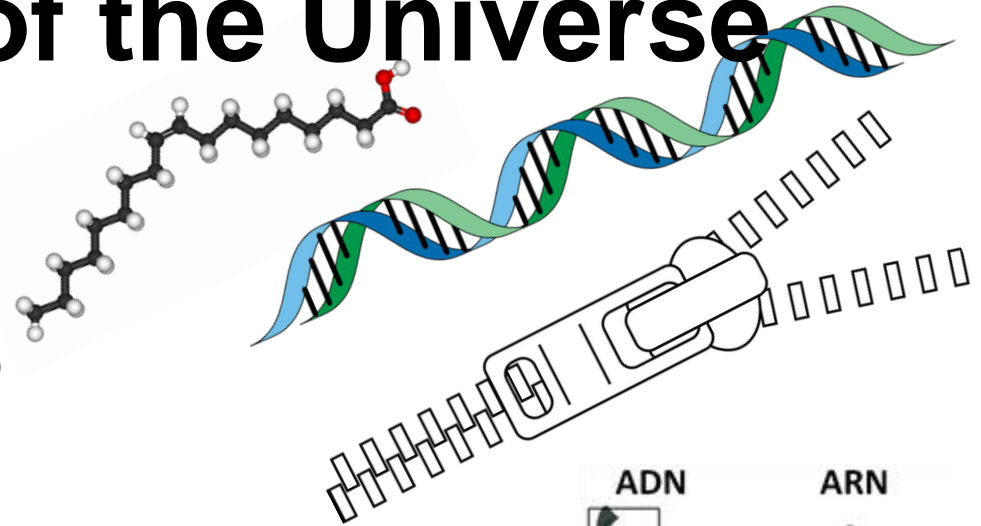


Let's take sentience in machines seriously

Where does sentience come from?

Where does everything that exists come from?

Small story of the Universe



Nothing

Something (Big-Bang) 13.800 M.y. ago

Expanding matter (Earth 4.470 M.a.)

Replicants (order) 4.000 M.a. ago

Cells 3.800 M.a ago

Multicellular 1.700 M.a ago

Engines, Sensors (Flavors / smells / eyes ...) and Brains

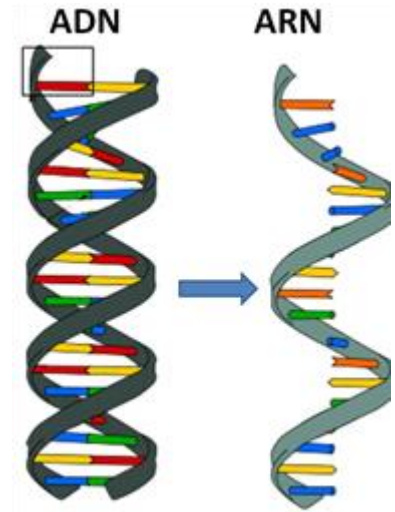
Water → Land → Air

Dominant species

Machines (Industrial and information revolution)

3D self-replicating printers, strong AI

Colonization of the galaxy



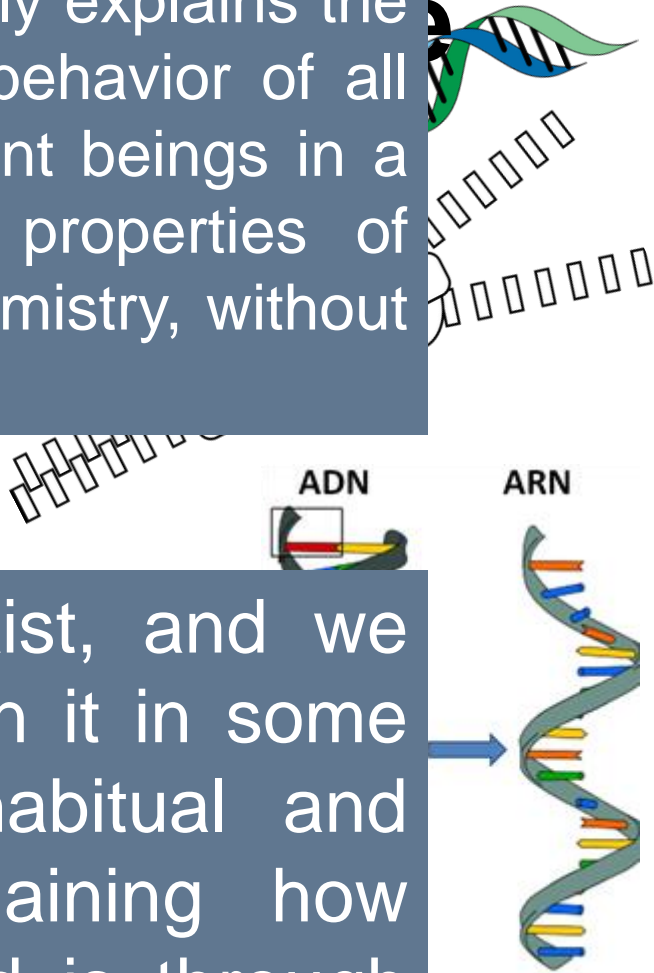
Where is the capacity to feel?

Evolution is a theory that perfectly explains the whole appearance and all the behavior of all living beings including all sentient beings in a reductionist way, through the properties of matter, through physics and chemistry, without including sentience.

Nothing
Something
Expanding matter (Earth 4.470 M.a.)
Replicants (order) 4.000 M.a. ago
Cells 3.800 M.a.
Multicellular
Engines,
Water →
Dominant
Machines
3D self-re
Colonizat

However, sentience do exist, and we have the impulse to explain it in some way. One of the most habitual and prestigious ways of explaining how sentience can be produced is through the evolutionary emergentist paradigm.

Where



It's very intuitive to think that we animals are sentient because our biological wet brains: complex systems that are result of evolution, which allow the emergence of sentience, that is selected because is useful.

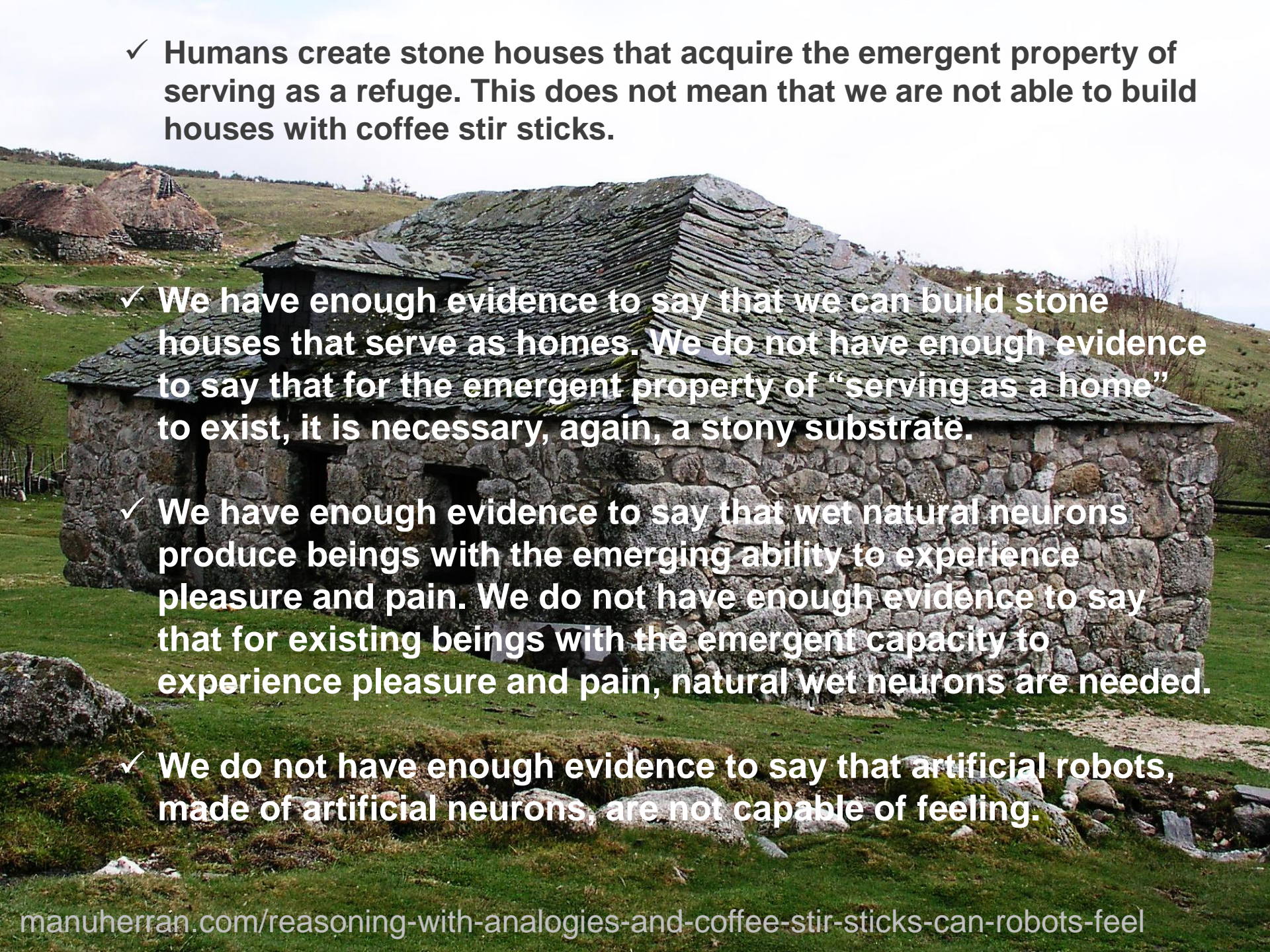


A photograph of a rocky coastline. The foreground and middle ground are dominated by large, layered, reddish-brown rock formations. Several people are scattered across the rocks, some standing and some walking. The sky is bright blue with scattered white clouds. The overall scene is a natural, rugged landscape.

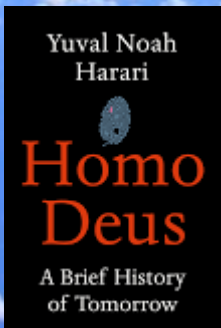
Simplest approach to argue about the idea than machines can feel:

- **Nature can naturally create systems in which the property of being “recipient” emerges.**
- **This does not imply that the only one way to create recipients is by the forces of nature.**
- **Humans can create systems in which properties emerges.**
- **If something emerges naturally, possibly it can also emerge artificially.**

- ✓ **Humans create stone houses that acquire the emergent property of serving as a refuge. This does not mean that we are not able to build houses with coffee stir sticks.**

- 
- A photograph of a traditional stone house with a slate roof, situated in a rural landscape with rolling green hills and other stone buildings in the background. The house is built from rough-hewn stones and has a simple, rectangular structure with a gabled roof. The surrounding area is grassy and appears to be a rural or agricultural setting.
- ✓ **We have enough evidence to say that we can build stone houses that serve as homes. We do not have enough evidence to say that for the emergent property of “serving as a home” to exist, it is necessary, again, a stony substrate.**
 - ✓ **We have enough evidence to say that wet natural neurons produce beings with the emerging ability to experience pleasure and pain. We do not have enough evidence to say that for existing beings with the emergent capacity to experience pleasure and pain, natural wet neurons are needed.**
 - ✓ **We do not have enough evidence to say that artificial robots, made of artificial neurons, are not capable of feeling.**

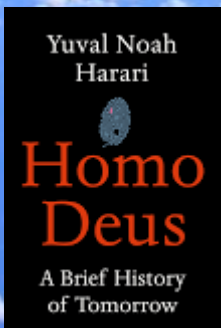
Is sentience useful for something?



Utility vs. Inevitability

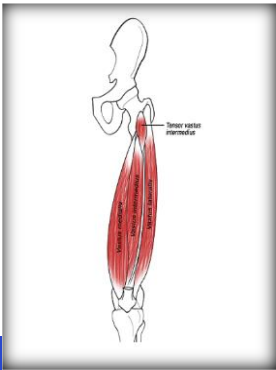
Is sentience useful for something?

If we believe that “sentience emerges from Central Nervous Systems” (matter → experiences). There are two options: sentience is useful by itself vs. sentience is inevitable. Harari: *“sentience could be like the roar of the engines of the plane: not useful to fly but inevitable”*.



Utility vs. Inevitability

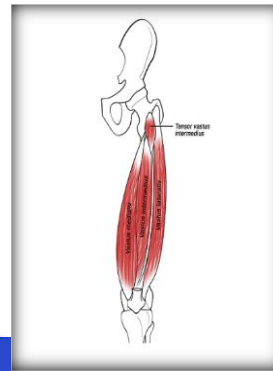
The central nervous system acts on the muscles.



MATERIAL THINGS

The central nervous system acts on the muscles.

The rest of the body also sends information to the central nervous system.

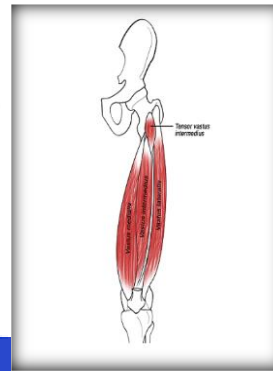


MATERIAL THINGS

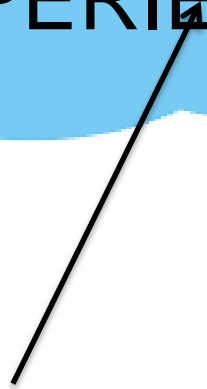
EXPERIENTIAL THINGS

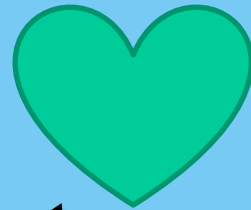


According to the emergentist paradigm, the central nervous system produces the emergence of sentience



MATERIAL THINGS

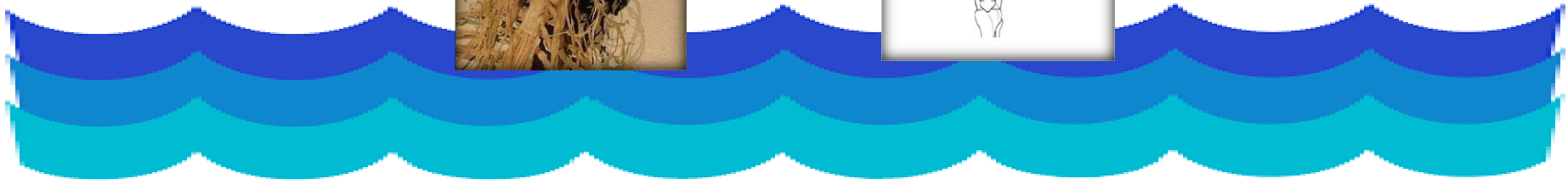
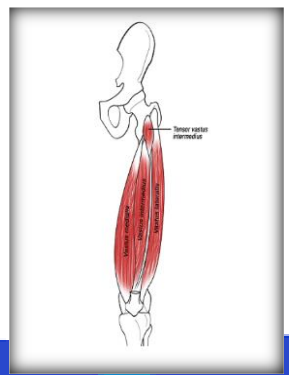
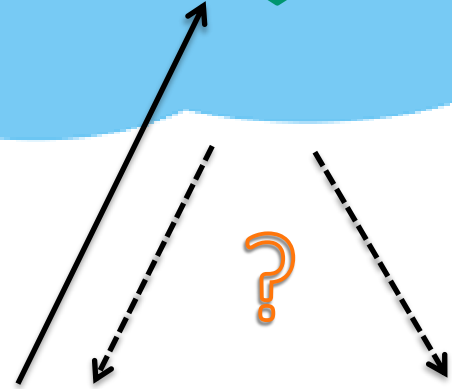




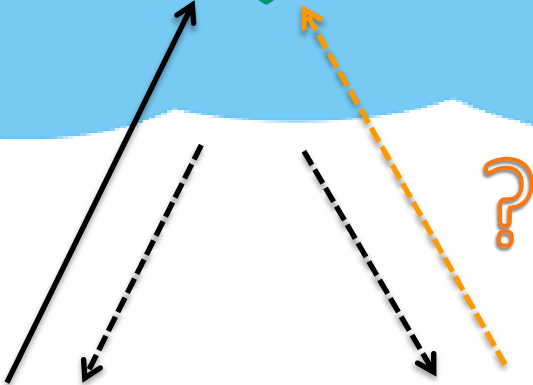
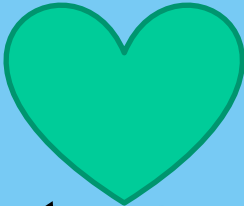
Is sentience useful by itself or is it just an inevitable sub-product?

If it were true that sentience is intrinsically useful, then we would be saying that matter does not comply with the laws of physics, since at least under certain conditions, matter would be affected by something that is not material

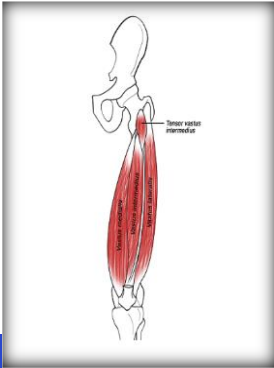
To be useful by itself, it should have an effect on the matter



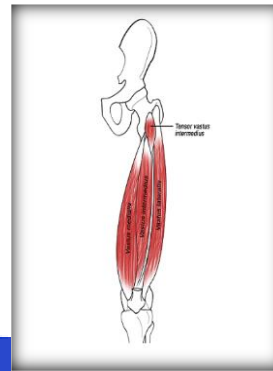
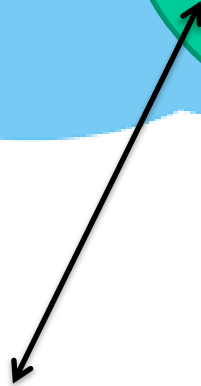
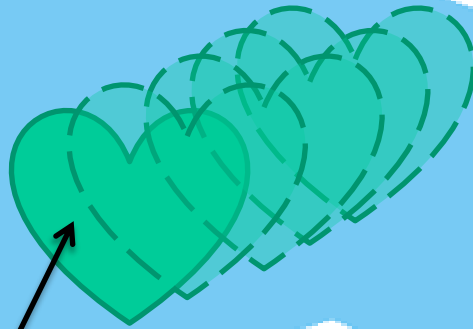
Other questions / options



Another type of matter other than the central nervous system could have or provoke experiences?



Other questions / options



Are experiences generated or invoked? Are they dependent on matter or do they have some kind of independent Platonic pre-existence?

Some theories, approaches and paradigms related to consciousness, sentience and identity

Conventional

GOD

PARTICLE

There are different types of frameworks from which we can give an answer about if machines are able to feel, why, how and how much.

Creative

Empirical

MATRIX

EMERGENCY

Bold

MATTER - HARDWARE

IDEAS - SOFTWARE

The first group are what I call theories or worldviews of the GOD type, which refer to beings or realities superior to ours, and which in some way determine it, such as religions.

Conventional

The second group is called PARTICLE and it is about those theories or hypotheses that consider necessary, for the existence of the capacity to feel, some component in particular (usually, material), such as, for example, biological, wet components, based on the carbon.

The third group of theories are EMERGENCIAS, the most popular among modern scientists, who consider that, based on a material basis, sentience emerges if certain conditions are met.

Creative

The fourth group I call MATRIX, because according to these theories nothing is what appears, and they put in doubt our intuitions about sentience and reality in general.

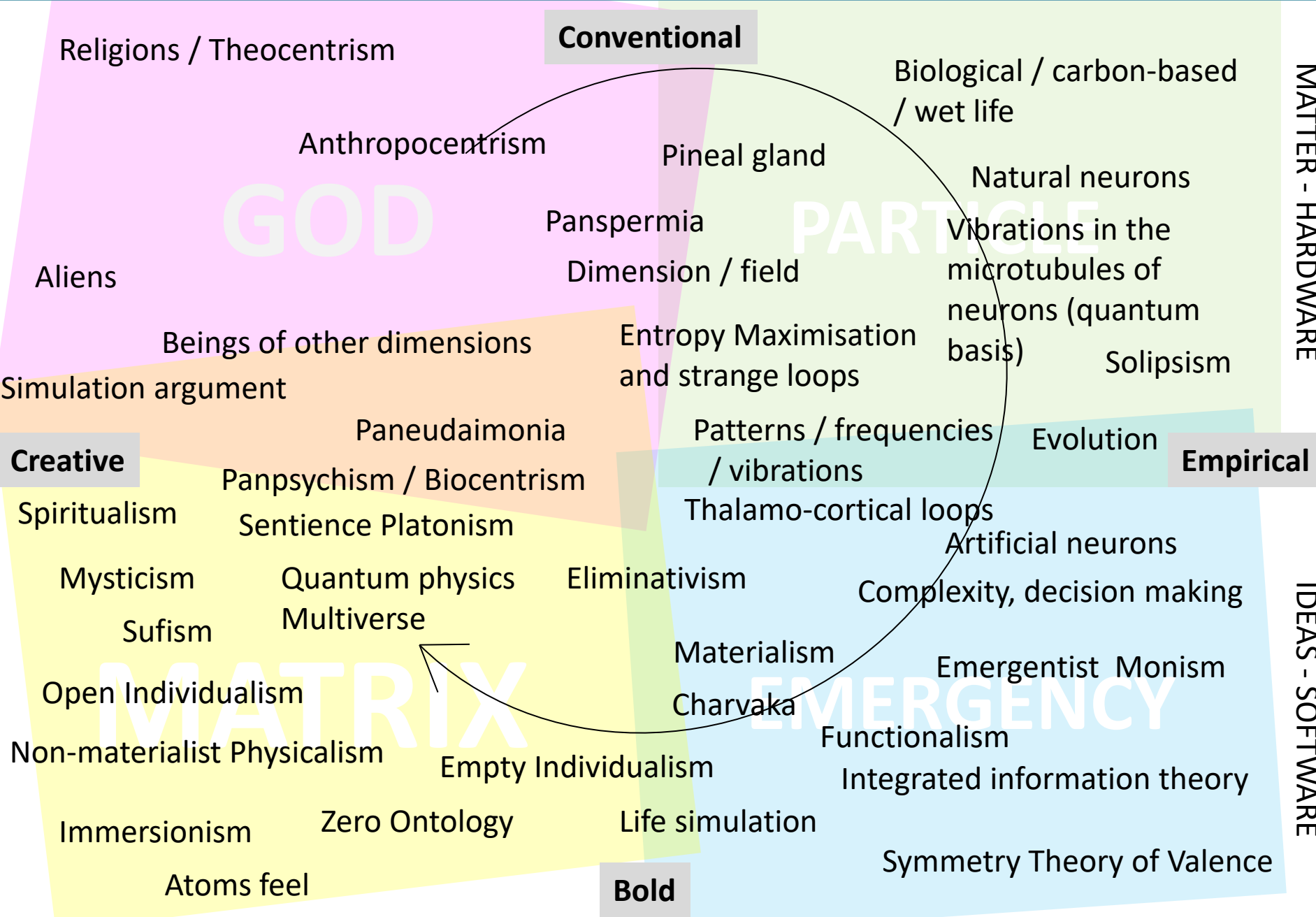
Empirical

Both historically and on a personal level, it is common to observe an evolution of beliefs in the indicated order, which I have illustrated with an arrow: GOD, PARTICLE, EMERGENCY and MATRIX. In some way, this intellectual journey returns to the starting point.

If you asked me about the probability that I assign to each of the four types of theories, I would say something like: 1%, 25%, 75%, and 99%. The sum of the probabilities does not necessarily have to be 100%, given that several hypotheses can be true at the same time.

Bold

Some theories, approaches and paradigms related to consciousness, sentience and identity



Some theories, approaches and paradigms related to consciousness, sentience and identity

Conventional

Religions / Theocentrism

Biological / carbon-based

MATTER - HARDWARE

Entropy Maximisation and strange loops: [Roshawn Terrell & Nell Watson](#).

Thalamo-cortical loops: [Francis Crick, Anil Seth](#)

Vibrations in the microtubules of neurons (quantum basis): [Roger Penrose](#)

Pineal gland: [René Descartes](#)

Simulation Argument: [Nick Bostrom](#)

[Open, Empty and Closed Individualism](#) ([Daniel Kolak](#))

Do atoms feel? [Brian Tomasik](#)

Panspermia: [Fred Hoyle](#)

Panendemia: [Manu Herrán](#)

Sentience Platonism: [Manu Herrán](#)

Immersionism: [Manu Herrán](#)

Symmetry Theory of Valence: [QRI](#) (Mike Johnson & Andrés Gómez Emilsson)

[Non-materialist Physicalism](#) and [Zero Ontology](#) (David Pearce)

Biocentrism: [Alberto Terrer](#)

s
m
sism

Empirical

IDEAS - SOFTWARE

aking
ism
neory

Aliens

Simulation argu

Creative

Spiritualism

Mysticism

Sufism

Open Individ

Non-materialis

Immersion

Atoms feel

Bold

Symmetry Theory of Valence

As for the moral models (prototypes) of each quadrant, I think that more or less could be as follows.

Quadrant 1 (GOD) The moral prototype of those who hold these beliefs is solidarity people, altruists, concerned about human rights, against torture and the death penalty and who collaborate with humanitarian organizations. They consider and value all human beings equally regardless of their intelligence, culture, country, age, sexual identity, sexual preferences, political preferences, race, skin color, abilities, etc. They are contrary to (involuntary and harmful) experimentation with human beings.

Quadrant 2 (PARTICLE) These people share the moral concerns of quadrant 1, but they also include all animals with a central nervous system. They are defenders of animal rights. They try to minimize the suffering of all beings that feel. They are contrary to experimentation with animals, and also with biological neural systems, since these could generate sentience and suffering.

Quadrant 3 (EMERGENCY) In addition to assuming the moral positions of quadrants 1 and 2, these people consider the possible emergence of sentience in machines and therefore robot rights, computer simulations and, in general, software, which has been constructed in a similar way or under similar conditions like those under which we -biological beings that feel-, have been built. In particular, they prevent the implicit risk in the construction of very complex physical or digital systems, capable of reasoning and / or capable of evolving.

Quadrant 4 (MATRIX) Those who consider these hypotheses, in addition to taking into account the three moral positions described above, take into account other possibilities related to the physics and philosophy of suffering that can be very unintuitive and could even be considered improbable, but whose implications in relation to prevention of suffering, if true, would be immense; and therefore consider it morally correct and necessary to devote at least a part of the resources available to investigate about these possibilities.

The answer to the question about the possible sentience in machines, according to each one of the quadrants, with nuances, seems to me to be the following:

Quadrant 1 (GOD)

“The question is absurd, machines can not feel, nonhuman animals can do it, but it is not very relevant, since the only relevant being is the human being, made in the image and likeness of God, the chosen people, anointed of divinity, what legitimizes we humans to use animals for our benefit and of course, also machines.”

Quadrant 2 (PARTICLE)

“Dry machines, made of metal and plastic, can not feel, whereas a biological machine, built using artificial biological cells, could do it.”

Quadrant 3 (EMERGENCY)

“We humans, as well as other animals and all living beings, are, in short, machines, therefore, what are known as robots, and in general machines built by humans and even artificial simulations can feel if met certain conditions of complexity and evolution in an appropriate environment, as has happened with us, animals.”

Quadrant 4 (MATRIX)

“Not only robots could feel. Atoms and even ideas could feel. We do not understand reality well and we do not know what is possible.”

HOW TO DEMONSTRATE SENTIENCE?

- METHODOLOGY: how to address the problem
- Three ways to solve the problem of recognizing sentience:
 - I – THEORIES OF SENTIENCE (DEDUCTION)
 - II – SIMILARITY WITH MYSELF (INDUCTION)
 - III - THE BEST POSSIBLE EXPLANATION (ABDUCTIVE REASONING)

HOW TO DEMONSTRATE SENTIENCE? METHODOLOGY

In the field of philosophical ideas we can not (easily) make predictions, but we can prove and demand philosophical hypotheses to have:

- In relation to the proposed idea:
 - Clarity
 - Internal coherence
 - Compatibility with the evidence (observations, experiences)
 - Explanatory capacity
 - Leave out accessory or arbitrary elements
- Regarding the author, the creation process and its context:
 - Honesty
 - Impartiality
 - Skepticism
 - Recognize the intention
 - Recognize the motivation

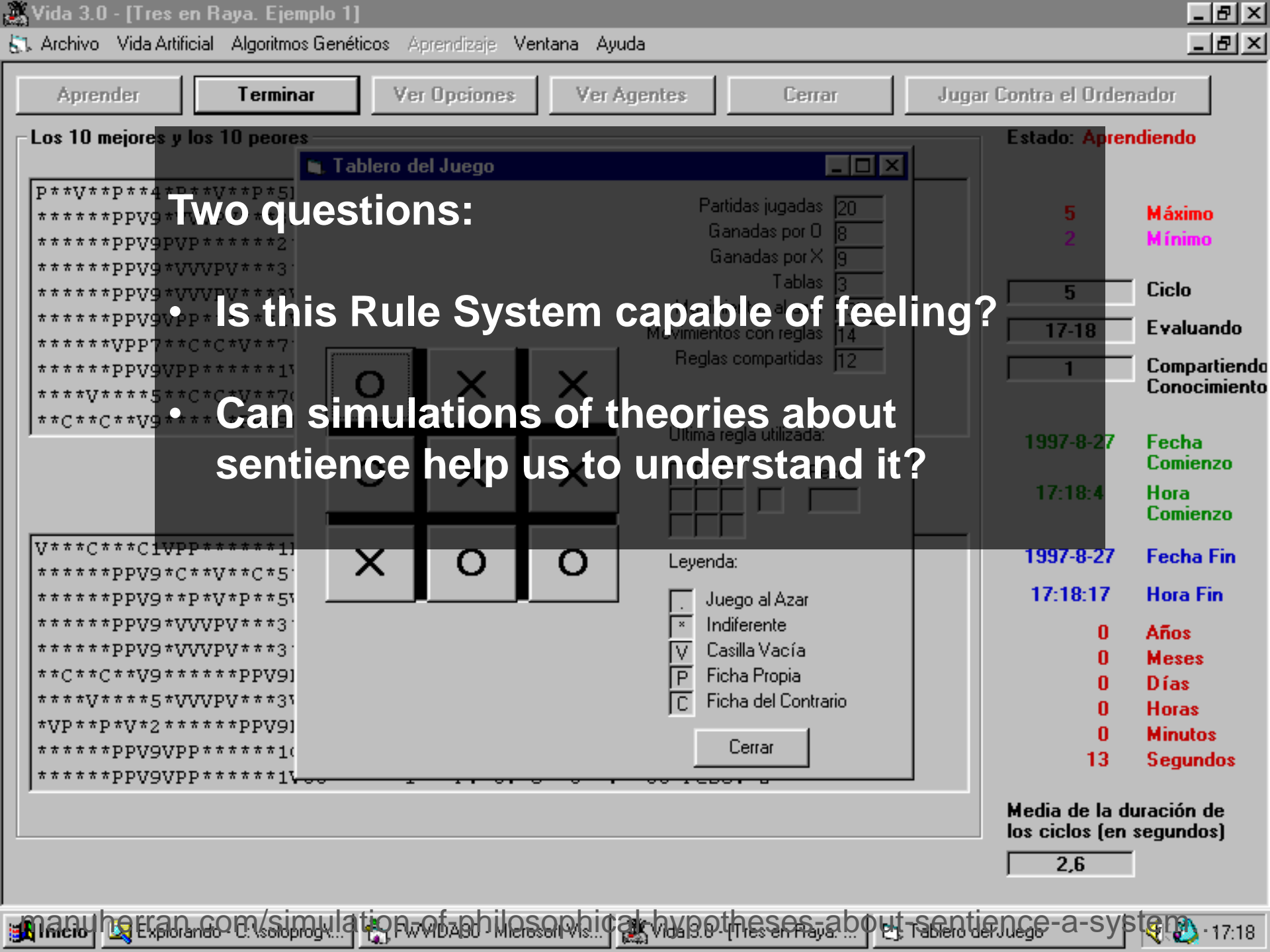
HOW TO DEMONSTRATE SENTIENCE?

- METHODOLOGY: how to address the problem
- Three ways to solve the problem of recognizing sentience:
 - I – THEORIES OF SENTIENCE (DEDUCTION)
 - II – SIMILARITY WITH MYSELF (INDUCTION)
 - III - THE BEST POSSIBLE EXPLANATION (ABDUCTIVE REASONING)

HOW TO DEMONSTRATE SENTIENCE?

I - THEORIES OF SENTIENCE (DEDUCTION)

- If we already have a theory of sentience in which we believe:
 - Use it to predict if that thing is sentient or not
- How can we know if a theory of sentience is correct or not?
 - Test their consistency in a simulation
 - Test their predictions in the real world
- How can we know we are not ignoring the right theory?
 - Use maps of theories of sentience and computer simulations of theories to try to cover them all



Two questions:

- Is this Rule System capable of feeling?
- Can simulations of theories about sentience help us to understand it?

Hormigas y Plantas 1.0 - [Hormigas y Plantas. Ejemplo 1] (Prg: 1, Ej:1, Aut:0, Iter:1) - [Hormigas y Plantas. 1]

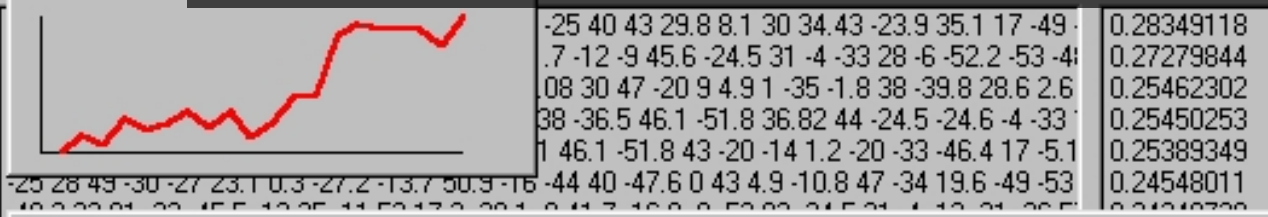
Fichero Edición Ejecutar Opciones Ver Ventana Ayuda

Another digital environment "Ants and Plants" (Artificial Life – Simulated Evolution)

Visual simulations of theories of sentience can be used to understand and evaluate the different theories and metaphysical hypotheses in relation to sentience. In this way we could have a more precise idea of how much sentience there is and where, which would allow us to be more effective in reducing suffering.

Variable	Value
Estado	Funcionando
Total Hormigas	23
Sexos	10
Energía Total	13
Nº mueren	0
Nº mueren Vejez	0
Nº Pelean	0
Duración de los ciclos (en segundos)	0,1689060

Evolutionary algorithms: evolving pieces of code



Estado: **Funcionando**

0.27920770 Máximo

0.00000000 Mínimo

22 Ciclo

7 Agente

12.6363636 Duración de los ciclos (en segundos)

1999-7-16 Fecha Comienzo

13:1:20 Hora Comienzo

1999-7-16 Fecha Fin

13:5:58 Hora Fin

0 Años

0 Meses

0 Días

0 Horas

4 Minutos

38 Segundos

278 Segundos totales

Código generado correspondiente al primer agente

```
Function YourFenotype(myEnvironment As UTEEnvironment) as Double
-----
' This code was automatically created at 16/07/99 13:03:36
-----
' This code is the fenotype (meaning) of this genotype (compressed program):
' Genotype: (      -3;      -52.7;      -41.4;          1.5;          9.4;          23;
' Genotype: (      ?+?;          k;          x;          k;          k;          ?+?;
-----
Dim ret as Double
Dim temp1 as double
Dim temp2 as double
Dim temp3 as double
temp3 = myEnvironment.d(1)
temp2 = -41.4
temp1 = temp2 + temp3
ret = temp1
YourFenotype = ret
-----
' After execution, the genotype has been adjusted to
' Genotype: (      -3;      -52.7;      -41.4;          1.5;          9.4;          23;
' Genotype: (      ?+?;          k;          x;          k;          k;          ?+?;
-----
End Function
```

```
3;      38.5;      21;
?+?;          x;          k;
```

HOW TO DEMONSTRATE SENTIENCE?

- METHODOLOGY: how to address the problem
- Three ways to solve the problem of recognizing sentience:
 - I – THEORIES OF SENTIENCE (DEDUCTION)
 - II – SIMILARITY WITH MYSELF (INDUCTION)
 - III - THE BEST POSSIBLE EXPLANATION (ABDUCTIVE REASONING)

HOW TO DEMONSTRATE SENTIENCE?

II - SIMILARITY WITH MYSELF (INDUCTION)

- Similar external appearance
- Similar internal constitution
- Similar behavior
- Similar (evolutionary) origin
- Similar genetics (genetic proximity)
- Similar “utility” (or inevitability).

I DO FEEL

WHO OTHERS FEEL?

Practical approach: "brain". Yes, but how do we really do it?
All are forms of "likeness".

- Maximum evidence: [1.] I feel

- "I feel, therefore I exist "

- Interpolation between individuals: [2.] Similar appearance and [3.] similar behavior

" If it looks like me and behaves like me, it will feel like me "

- Interpolation between species: [4.] Same origin (evolutionary) and [5.] Genetic proximity

" If you have been created like me, you will feel like me "

- Utility or need: [6.] Evolutionary utility (or inevitability)

"Evolution is testing things and keeps those that serve for something" (as has happened with me)

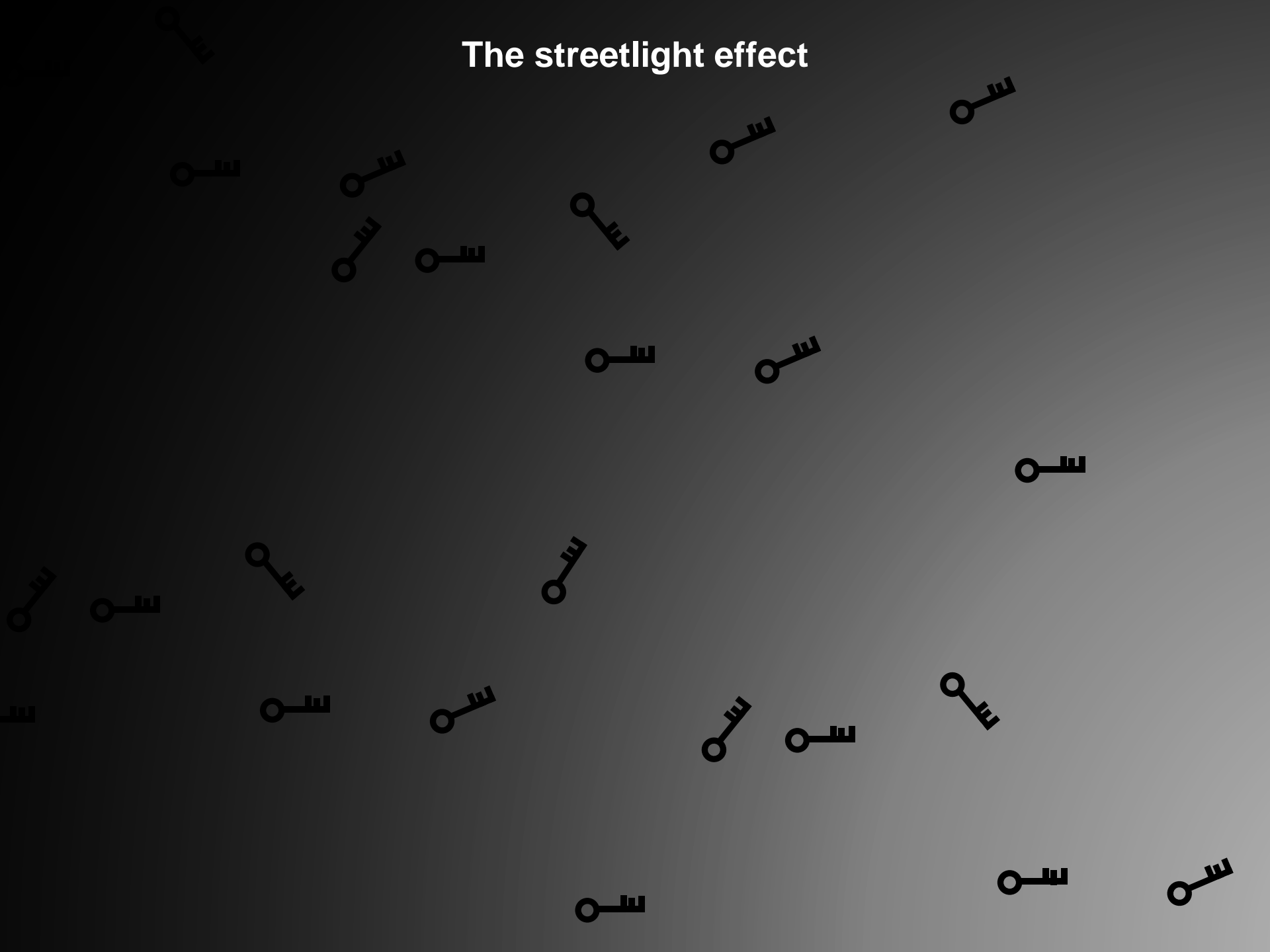
HOW TO DEMONSTRATE SENTIENCE?

III - THE BEST POSSIBLE EXPLANATION (ABDUCTIVE REASONING)

We can use abductive reasoning, in the style of Sherlock Holmes: unless it is an intentional deception, if a being seems sentient, it surely is. In this sense:

- If it seems sentient, probably it is. But
- If it has been built simply to seem sentient, then probably it is not.
- If it doesn't seem sentient, it could still be. In this case, one way to clear this mystery is the reasoning by analogy (similarity). But this can be unfair to other beings very different from us because of the streetlight effect.

The streetlight effect



The streetlight effect

We look for sentience not where it is more probable to be, but where is more probable to find (by similarity to us).

We must look for sentience where is more probable to reduce the biggest amounts of intense suffering.

We need to research the different sentience paradigms and not just believe in only one because it could lead to an astronomical moral catastrophe.

HOW TO DEMONSTRATE SENTIENCE?

III - THE BEST POSSIBLE EXPLANATION (ABDUCTIVE REASONING)

- If it has been built simply to seem sentient, then probably it is not.
- But if we had very powerful computers, and we let a simulation run the equivalent of 4,000 million human years, and a character in the simulation, when he is nailed says "Ouch!", we must conclude that it feels.

HOW TO DEMONSTRATE SENTIENCE? FORMULA

$$S = K_1 \sum CT_n PT_n + K_2 \sum CL_n PL_n + K_3 PE$$

S is the sentience, the capacity to feel of an object.

K_1 , K_2 and K_3 are the relative weights (or importance) that we assign to different groups of criteria to recognize sentience. Its sum must give 1. Each weight K corresponds to a set of criteria to recognize sentience. This formula contemplates three groups of criteria:

K_1 is the importance we give to theories about sentience (T)

K_2 is the importance we attach to likeness, resemblance or closeness (L)

K_3 is the importance we give to the best possible explanation (E)



Sentience is a fact

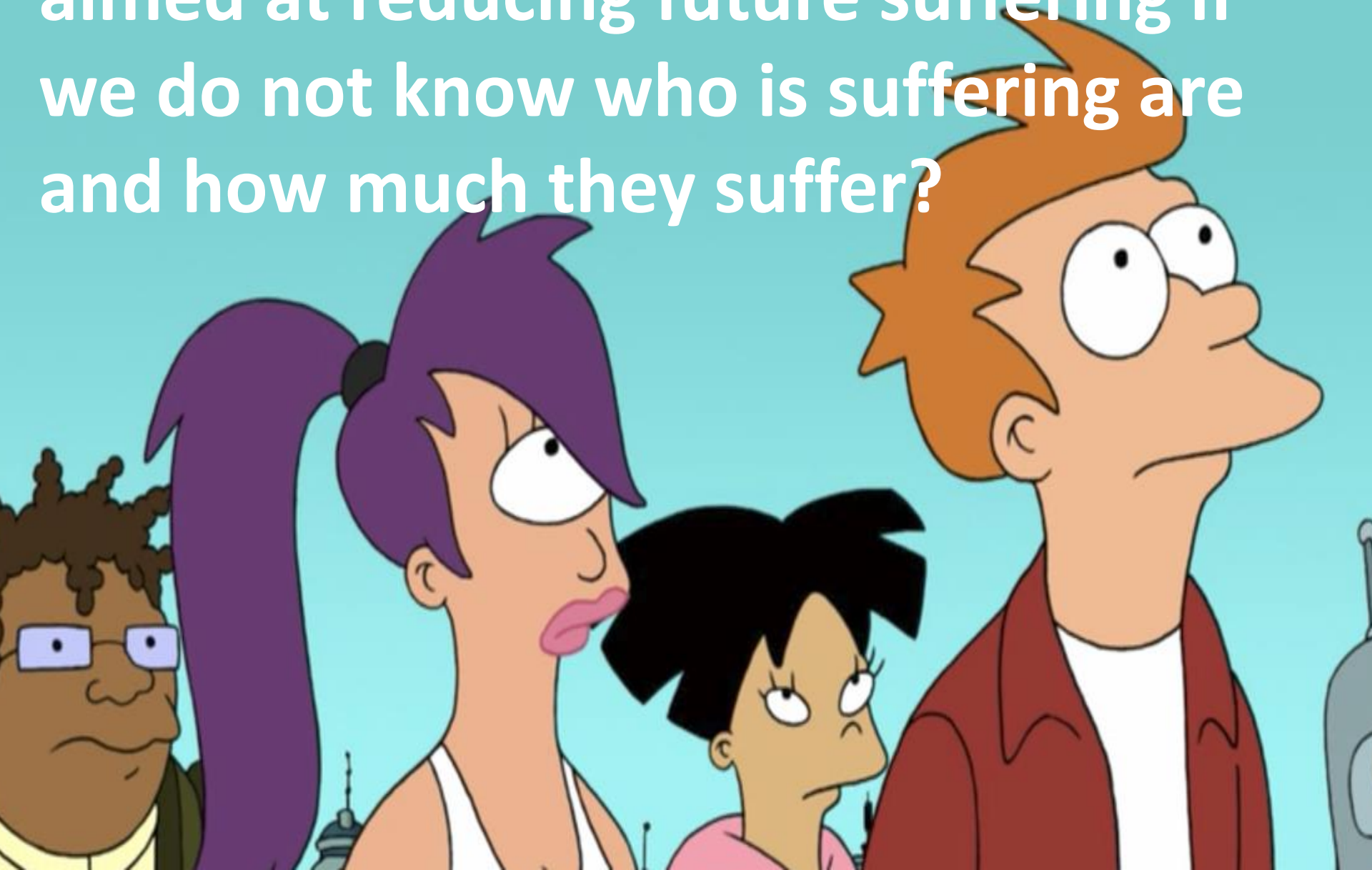
**We recognize the
sentience in others
because of their
resemblance to us**



A large, detailed illustration of an octopus, likely a cuttlefish or squid, with its tentacles spread out against a blue background. The octopus is brown and green, with a textured, bumpy skin. Its tentacles are long and have small suckers. The background is a solid blue color.

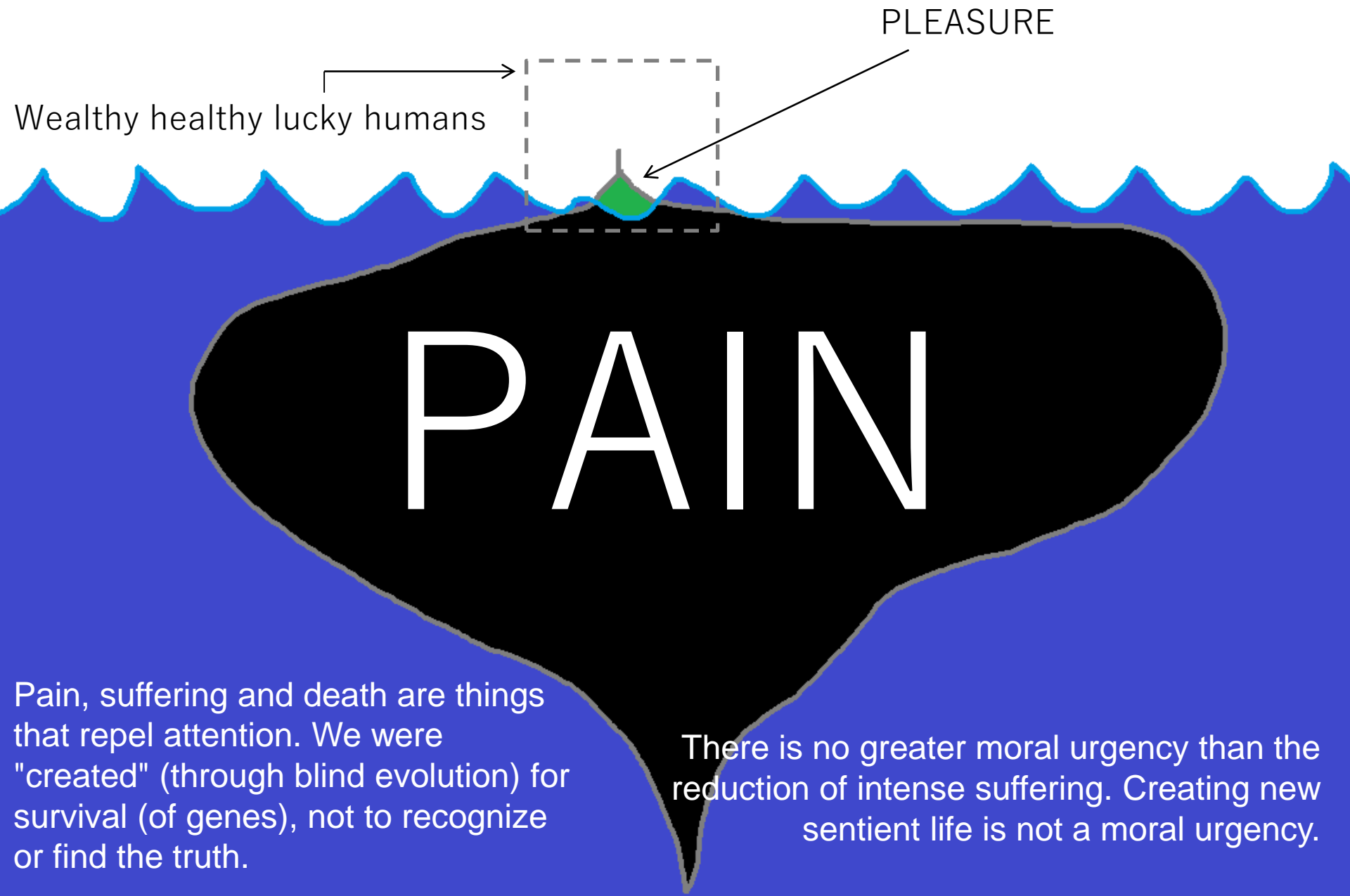
**But how can we know if other
objects very different from us are
sentient?**

How to prioritize limited resources aimed at reducing future suffering if we do not know who is suffering are and how much they suffer?

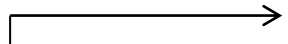


The situation could be even worse than it seems. Not only we need to know who the suffering beings are and how much they suffer. We also need a better understanding about what sentience is. For example, according to some paradigms, the idea of "being who suffers" apart from everything else could be wrong. That is, we not only ignore the answers, but we could be asking some wrong questions.

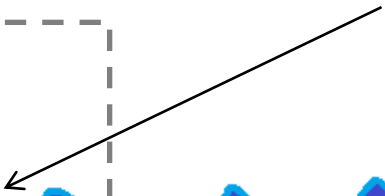
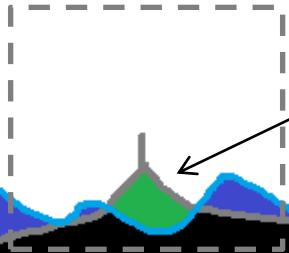




PLEASURE



Wealthy healthy lucky humans



PAIN

Pain, suffering and death are things that repel attention. We were "created" (through blind evolution) for survival (of genes), not to recognize or find the truth.

There is no greater moral urgency than the reduction of intense suffering. Creating new sentient life is not a moral urgency.

CONCLUSIONS

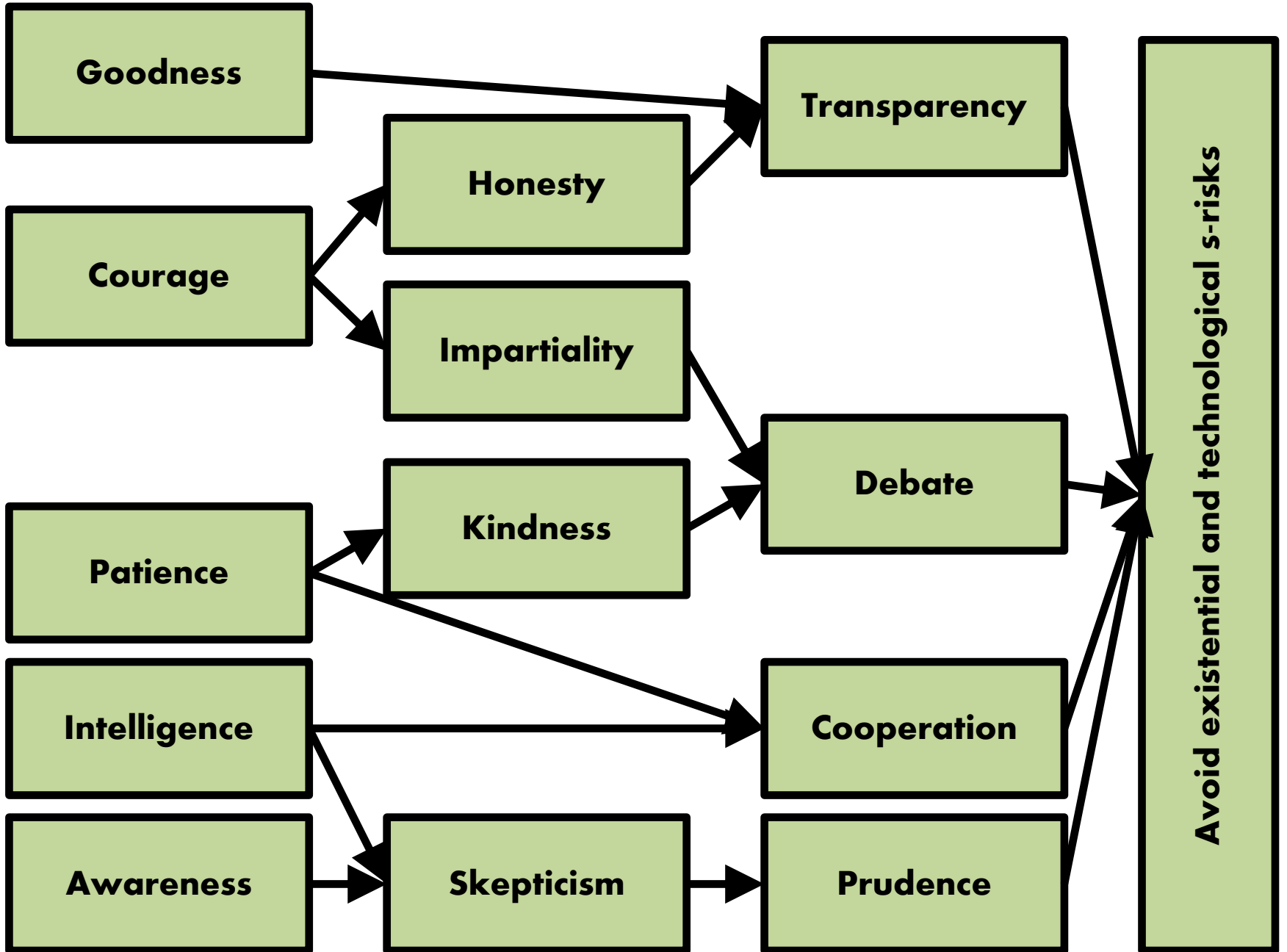
- It is no crystal clear of what sentience / consciousness is: neither of what generates / invokes it, nor of the conditions necessary for it to happen.
- From different philosophical perspectives, machines can be sentient. What changes are the conditions that are supposed to be necessary (v.g. biological machines).
- The creation of machines or simulations can cause a moral catastrophe of an astronomical magnitude.
- We should not rule out hypotheses simply because they seem far-fetched or anti-intuitive. Its moral implications could be extraordinary.
- We must continue carefully contemplating, understanding and researching on various hypotheses of sentience.

PRECAUTIONS AND RECOMMENDATIONS

- Slow down the technological development until it is equated with moral development.
- Convergence of values. Search for agreements.
- International / intercultural / inter-axiological collaboration.
- Transparency in research.
- Patience, humility, indulgence, strategy.
- Activism against experimentation in neuronal biological substrates.
- Development of the scientific attitude:
 - Open and skeptical mind (about other theories and mine).
 - Impartiality, honesty and skepticism.
 - Recognition of motivation.

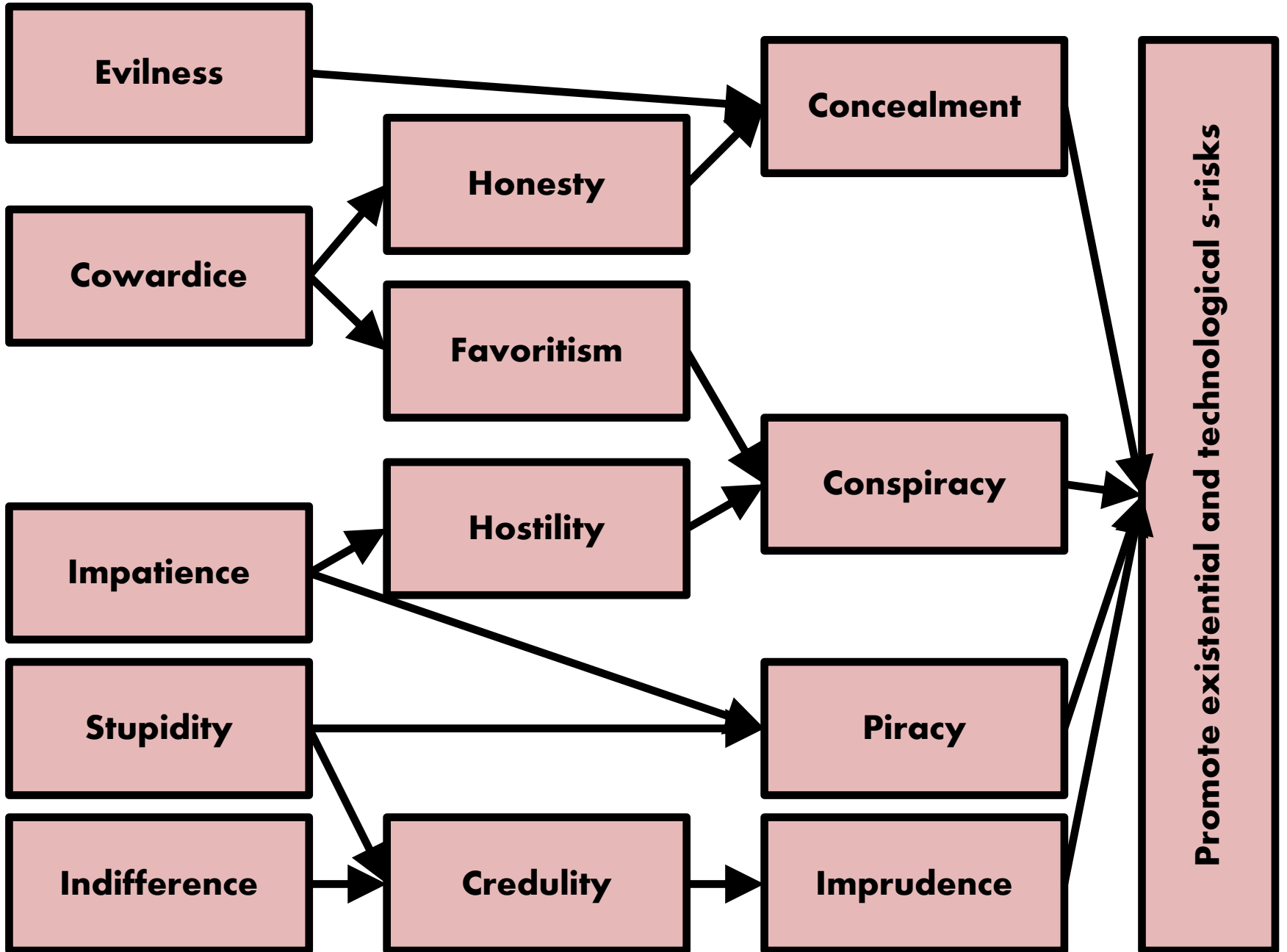
Attitudes and attributes and behaviors

Results



Attitudes and attributes and behaviors

Results



DRAFT PROPOSAL OF ACTION

- Disseminate the ideas and reasoning that lead to the recognition of possible sentience in machines.
- Disseminate the axiology that leads to the consideration of the moral relevance of sentience in machines.
 - Promote an ethic based on the ability to feel.
 - Promote the moral consideration of animals and the reflection of the reasons for this consideration.
- Work on defining a strategy against experimentation in biological substrates.
 - Is it necessary to previously promote and win the debate of animal experimentation?
 - Collaborate with experienced organizations in defense of animals.
- Promote research on sentience (eg. creation and visual simulation of hypotheses for validation)

Thanks!



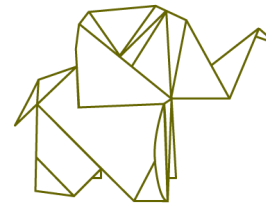
manuherran.com



OPIS

Organisation for the
Prevention of
Intense Suffering

preventsuffering.org



Sentience
RESEARCH

sentience-research.org