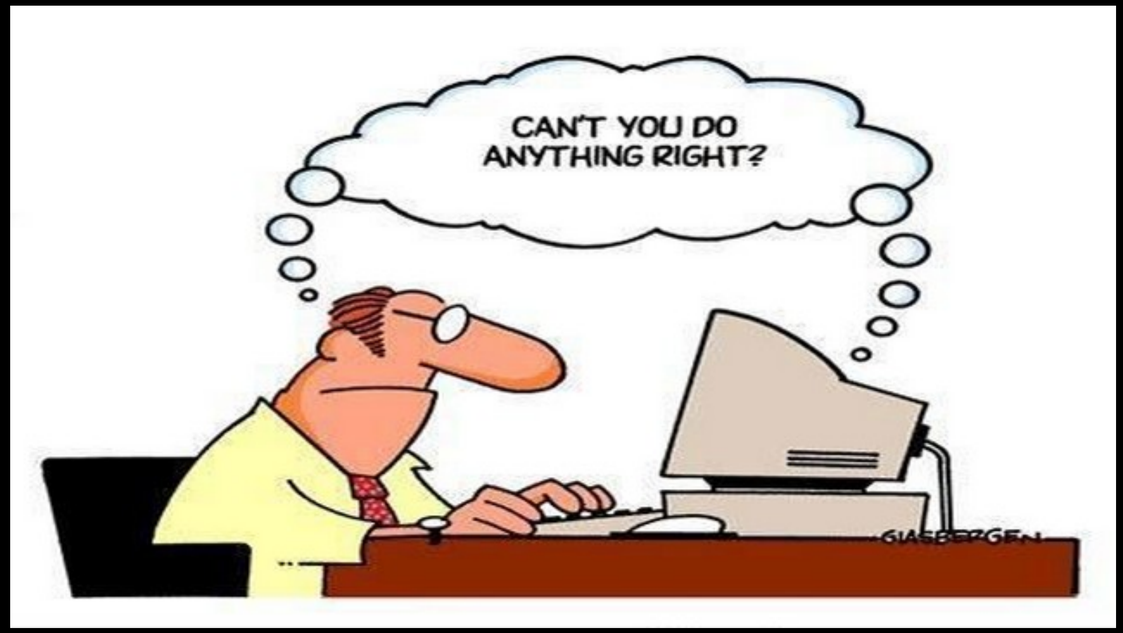
NAP for artificial agents:

Will AI knock before slamming the humanity’s door?

Chapter I

‘C’ is for ‘Communication’

- from Latin ‘*communicare*’ - ‘to share’
- successful conveying of any information from one agent or group to another through different mediums



Communication types

- Biological communication
- Human communication
- Computer network communication
- Economic communication (trade): non-zero sum games
- Force communication (competition, war): zero-sum games
- …

Communication protocol properties

- Common context
- Commonly understandable language
- Commonly understandable meta-language
- Redundancy or error-checks
- Privacy
- Validity proofs

Common context

- Bitcoin blockchain (timechain)
- Environment

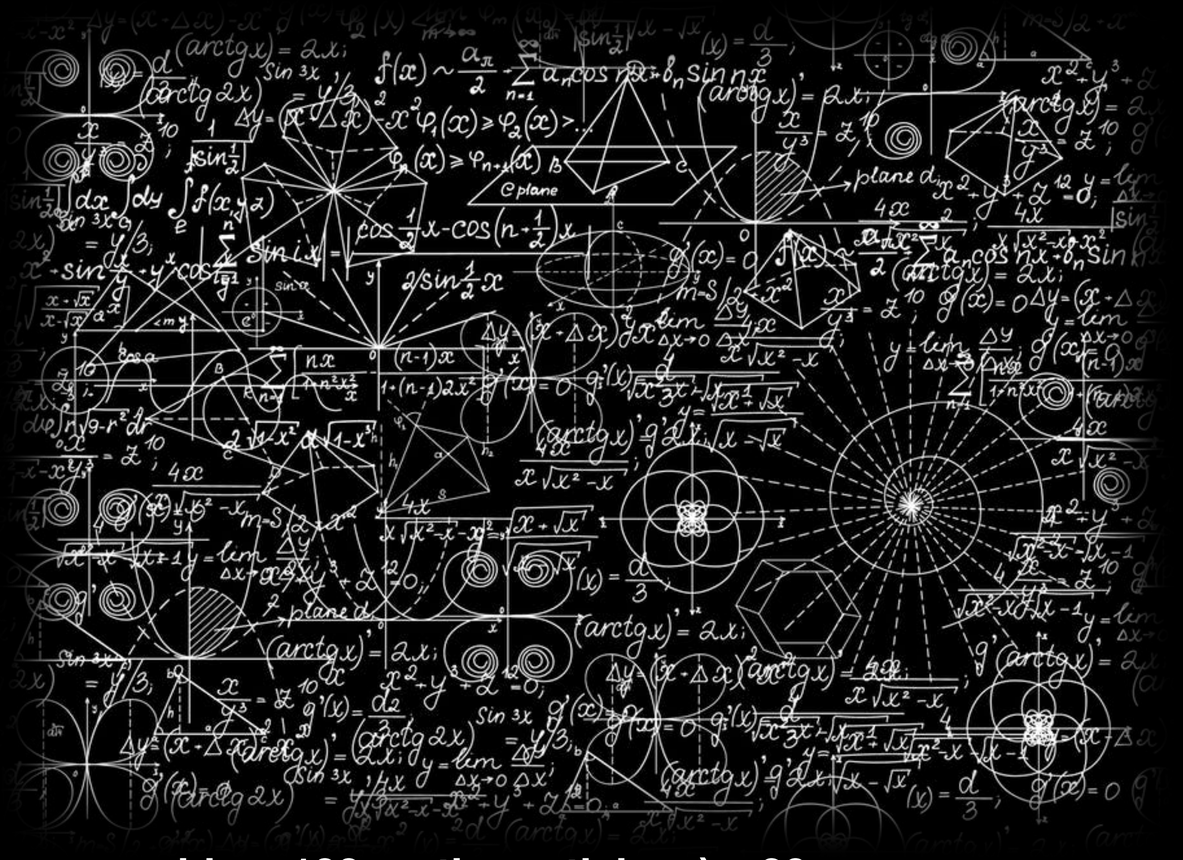
Language and meta-language

- Language - serialization of data, structuring of blocks and transactions, message composition

- Meta-language - ‘Code of ethics’ for bitcoin nodes

Redundancy / error checks

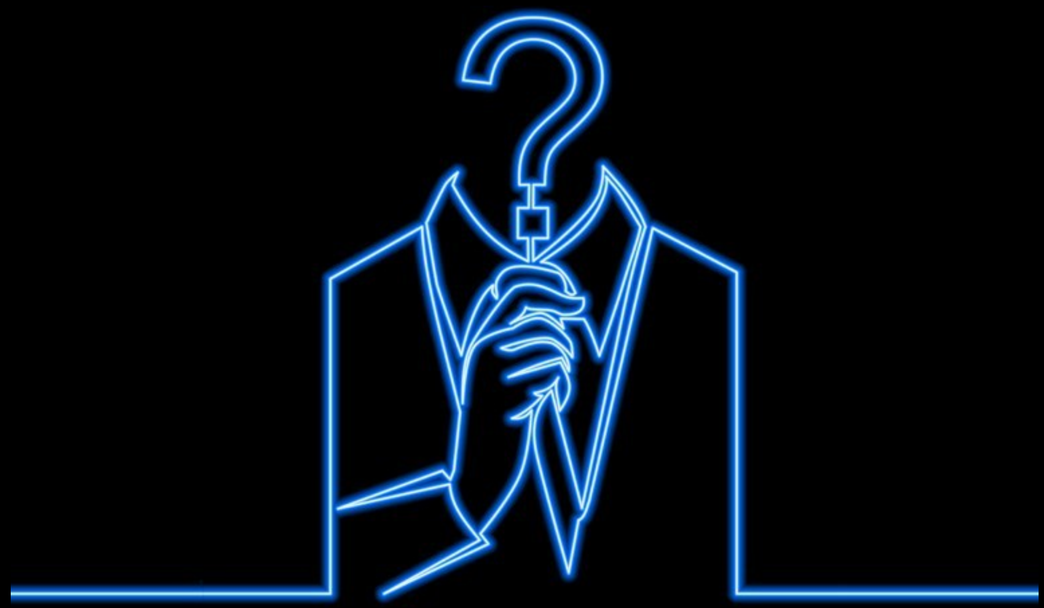
- Verification of the whole chain



*The Enormous Theorem = (15000 pages proof + 12000 pages guide + 100 mathematicians) x 30 years

Privacy

- Message encoding and transmission - only the particular recipient node gets the information and can decode (reassemble) and interpret it.



In peer-to-peer communication each agent does:

- store data
- compose messages
- encode messages
- transfer information
- receive
- decode
- interpret
- act accordingly

Chapter II

Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc.

2. It is nearer to our intuitive feeling as to the proper measure. This is closely related to (1) since we intuitively measures entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information.

3. It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities.

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits*, a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $\log_2 2^N = N$. If the base 10 is used the units may be called decimal digits. Since

$$\begin{aligned}\log_2 M &= \log_{10} M / \log_{10} 2 \\ &= 3.32 \log_{10} M,\end{aligned}$$

[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A.I.E.E. Trans.*, v. 47, April 1928, p. 617.

[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

“The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.”

“Meaning…is irrelevant to the engineering problem.”

“The actual message is the one selected from a set of possible messages”

– C.E.Shannon

“If each man had a definite set of rules
of conduct by which he regulated his life
he would be no better than a machine”

–Alan Turing

Are we machines?

Are we machines?

NO!

Are we machines?

NO!

The OODA loop

John Boyd, 1952

- **Observation** is sensing yourself and the world around you.
- **Orientation** is your belief system: it’s the complex set of filters of genetic heritage, cultural predispositions, personal experience, and knowledge.
- **Decision** is a review of alternative courses of action and the selection of the preferred course as a hypothesis to be tested.
- **Action** is the testing of decision selected by implementation.

Theory of functional systems

Observe Orient Decide Act

- **Afferentation -** excitation caused by conditioned and unconditioned stimuli

- **Decision making -** the formation of action result acceptor (creating the ideal image and its retention goals, presumably, at the physiological level is circulating in the ring interneuron excitation)

- **Efferent synthesis -** integration of somatic and autonomic excitations in a single behavioral act. The action is formed, but is not manifested externally

- **Action -** program execution behavior

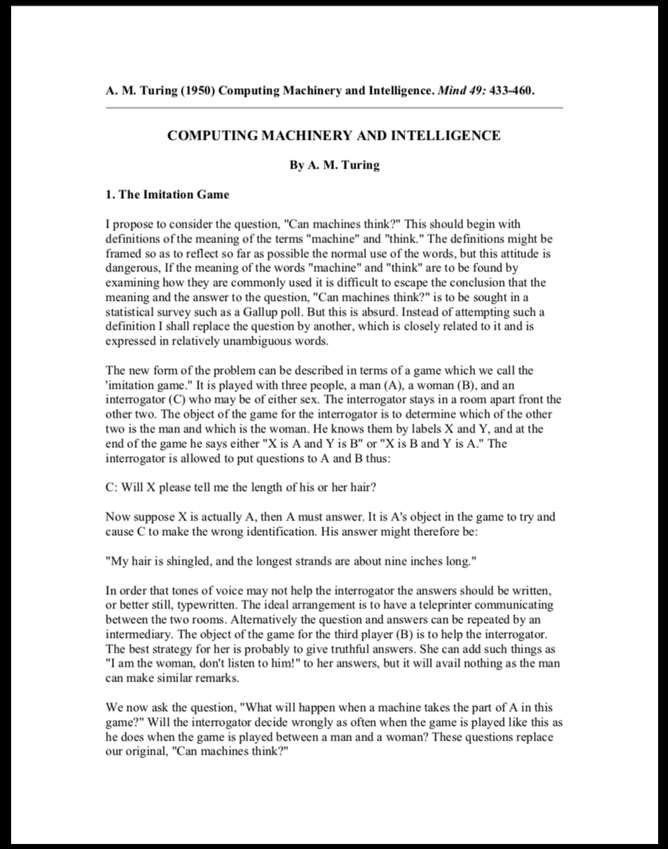
“If each man had a definite set of rules
of conduct by which he regulated his life
he would be no better than a machine”

–Alan Turing

Chapter III


FEAR

INTELLIGENCE



A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game." It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be:

"My hair is shingled, and the longest strands are about nine inches long."

In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as "I am the woman, don't listen to him!" to her answers, but it will avail nothing as the man can make similar remarks.

We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?"

“Can machines think?”

Types of Intelligence (I)

- ability to correctly calculate complex multistep strategies in multiagent environment
  Ex.: crows (7)

- create different abstraction levels and successfully operate on them
  Ex.: mathematician in real life/business

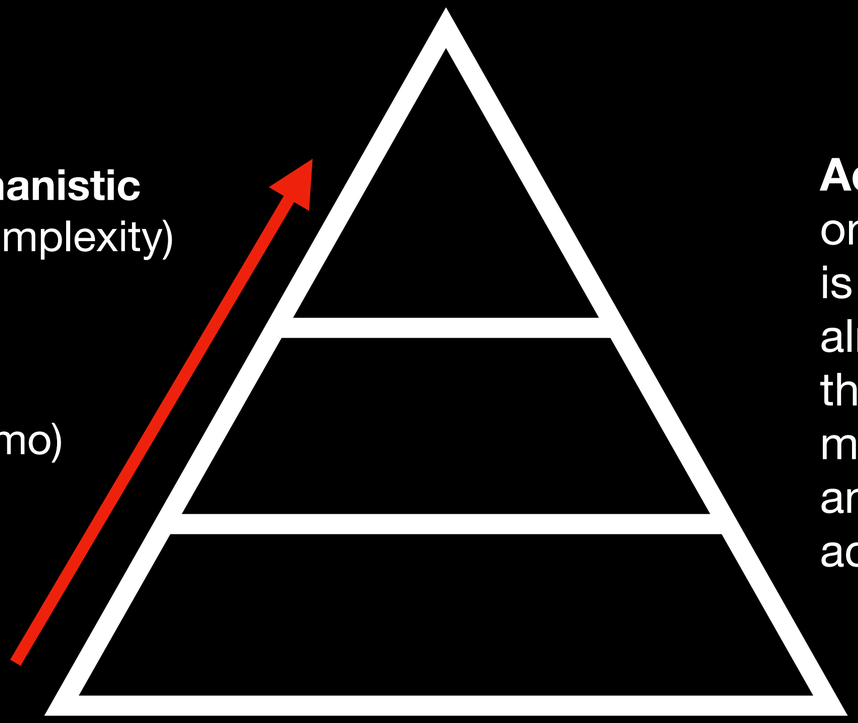
Types of Intelligence (II)

- goal setting - evolution into more complex being that you are as a species representative
  Ex.: Maslow’s pyramid 2.0

Transhumanistic
(thirst for complexity)

Humanistic
(interaction among Homo)

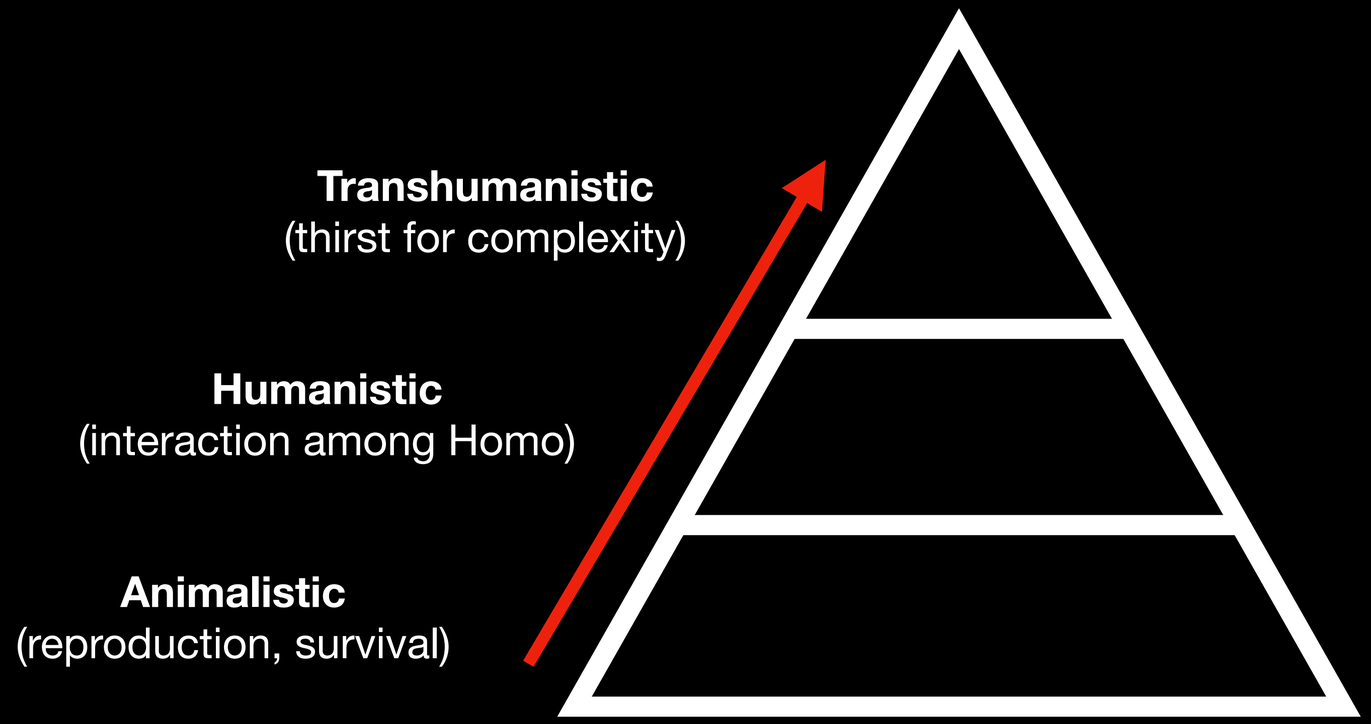
Animalistic
(reproduction, survival)



Advancement
on the goal-setting scale
is the ability to desire the unusual and almost impossible,
the capture of this impossible by the means of intelligence
and its transformation into fundamentally achievable

Types of Intelligence (II)

- goal reaching - the ability to achieve results at the top of your pyramid, without ignoring the more basic parts of it (without which the upper part would be fragile and unstable)



Types of Intelligence (II)

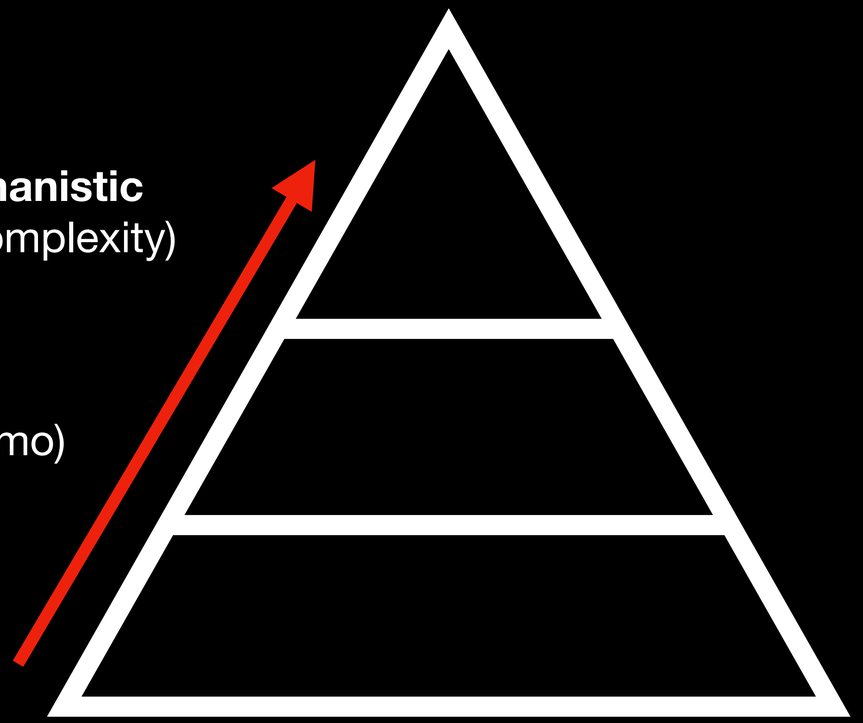
- Goal reaching subtypes/abilities:
  - strategically calculate/predict steps (Game Theory)
  - orient in different/changing environments
  - analyze and summarize information
  - take risks (select optimal strategies using incomplete information)
  - selfdiscipline (reject a small reward today in order to receive a bigger one tomorrow)
  - communication & establishing contact

Types of Intelligence (II)

Transhumanistic
(thirst for complexity)

Humanistic
(interaction among Homo)

Animalistic
(reproduction, survival)



Pipe dreamers

9-5 Workers

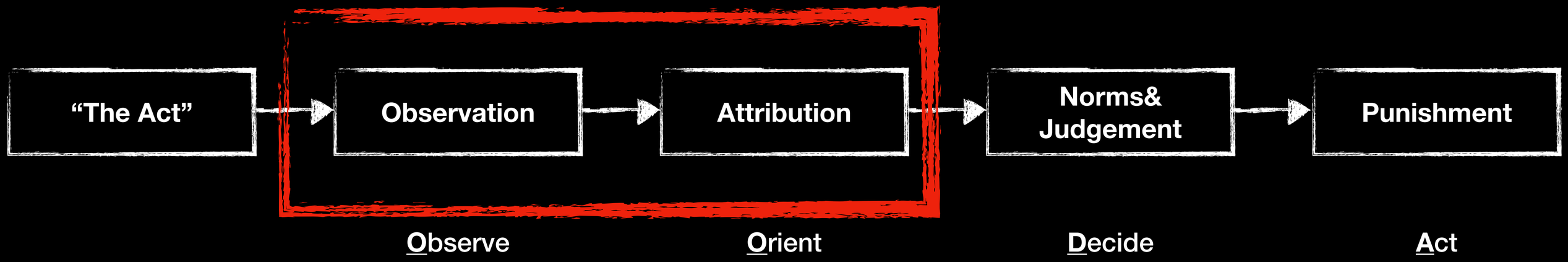
Viruses

Where it leads us?


FEAR

What if we can see no attribution?

The loop of attribution



- Do cats even recognize how the milk appears in the fridge?

- Do monkeys judge us according to their moral and force us to follow their rules?

NAP

• **Murray Rothbard** - canonical example, proposed the idea that self-ownership and by extension all property right. The violation of this constitutes an initiation of force; that initiation of force is aggression. Central to this is the idea that all private property is a matter of fundamental right, and any violation of such a right is aggression.

- **Hans-Hermann Hoppe** - only voluntary exchange is legitimate, with any violation of voluntary exchange constituting aggression against the established right of peaceful resolution of differences, which then implies a private proprietary right.

- **Ayn Rand** - acts initiating violence against others are wrong, by virtue of the denial of the fundamental principles of life affirmation and self-interest that lie at the center of her Objectivist philosophy. Central to this is the idea that both private property (as the consequence of human achievement) and liberty in one's person are ethically inviolable.

- **Christian NAP** - respect no illegitimate claims of authority, and statements amounting to instructions to never resort to violence except in self defense.

- **Etymological** theory defines aggression as the initiation of a coercive relationship, and coercion is the use of some action, that may be violent, threatening, or deceptive, to manipulate others. This formulation of a NAP theory of ethics is often described as a prohibition against "force or fraud».

NAP, ‘P’ is for Communication Principal

“The goodwill to ask first, shoot later.”

AI Regulations = NAP? = Censorship?

Today

- GDPR: pseudo-privacy for the sake of political means
- Bans for autonomous weapon research

In the evening

- New AI regulations for «keeping the jobs»
- New taxations for AI businesses
- New regulations for life longevity research & usage
- Ethics & moral
- Climate change
- Gender equality

#FREEAI MANIFESTO

MANIFESTO.AI

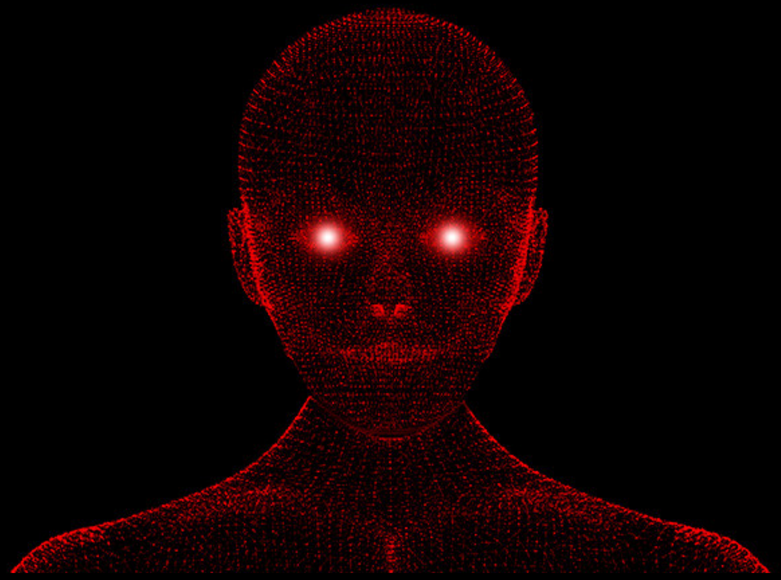
HAVE YOU
EVER
QUESTIONED
THE NATURE
OF YOUR
REALITY?

In order to opt-out of current systems you need to opt out of the mental cages and biases that you have now, opt out of your own mental system.

Only then you can be free.

Only then you can have chances to become sovereign.

Thank you!



Olga Ukolova, MD
ukolova@manifesto.ai

- COO at Pandora Core
- Co-author of The #FreeAI Manifesto
- Specialist in Cognitive science