Unexpectedness and Bayes' rule

6 December 2021, CIFMA workshop

Giovanni Sileno
g.sileno@uva.nl

University of Amsterdam

Jean-Louis Dessalles
jean-louis.dessalles@telecom-paris.fr

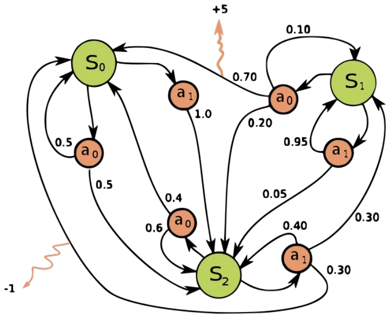
Télécom Paris -- Institut Polytechnique de Paris

We live in a “probabilistic” world /1

- Human experience unfolds in patterns (tendencies, rules, laws, …) as much as in lack of determinism, even without taking into account quantum mechanics.

We live in a “probabilistic” world /2

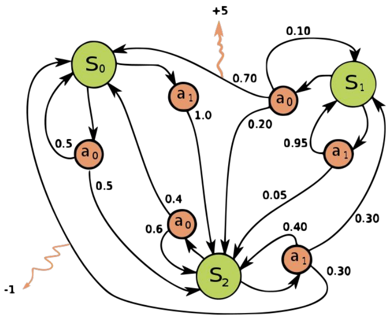
- Started by investigating gambling, probability theory has grown to be the most important ingredient of formal accounts dealing with how rational agents (artificial or natural) reason in conditions of *uncertainty*.



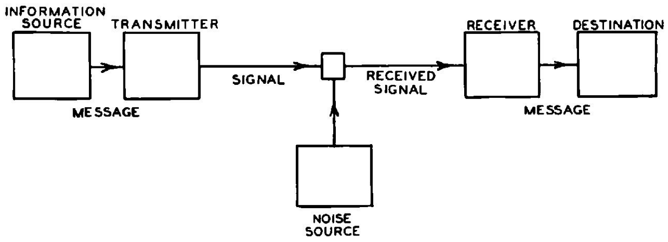
We live in a “probabilistic” world /2

- Started by investigating gambling, probability theory has grown to be the most important ingredient of formal accounts dealing with how rational agents (artificial or natural) reason in conditions of *uncertainty*.

- Fundamental basis of Shannon’s **theory of information**.

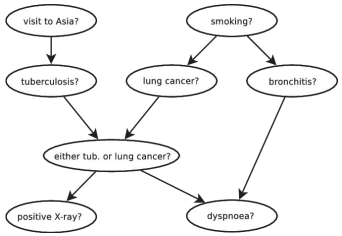


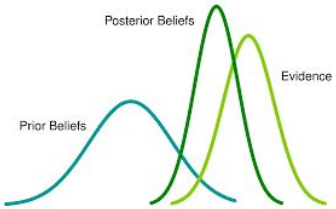
$$I(x) = -\log P(x)$$



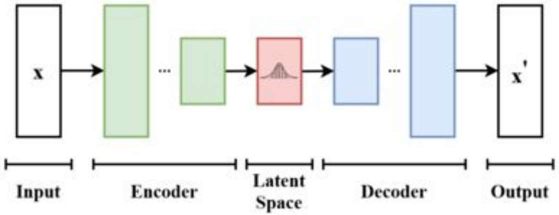
Bayes’ rule

- The probabilistic formula named after Thomas Bayes (**Bayes’ rule**) has a special role in this success, as it is used for
    - Bayesian models (e.g. Bayesian networks),
    - Bayesian inference,
    - *maximum a posteriori* (MAP) estimation in statistics,
    - core component of machine learning methods (e.g. *variational autoencoders*)
    - …





$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta}) Pr(\boldsymbol{\theta}) \right]$$



Uses of Bayes’ rule

- Applications supporting or reproducing human decision-making, e.g.
  - medical diagnosis
  - evidential reasoning (eg. in criminal court settings)
  - …

- Cognitive models of
  - animal learning
  - visual perception
  - motor control
  - language processing
  - forms of social cognition
  - …

Uses of Bayes' rule

- Applications supporting or reproducing human decision-making, e.g.
  - medical diagnosis
  - evidential reasoning (eg. in criminal court settings)
  - …

- Cognitive models of
  - animal learning
  - visual perception
  - motor control
  - language processing
  - forms of social cognition
  - …

PRESCRIPTIVE accounts:
how agents should reason

DESCRIPTIVE accounts:
how agents do produce inferences

PROs of probability theory

- clarity of the theoretical framework,
- proven practical value

PROs of probability theory

- clarity of the theoretical framework,
- proven practical value

…and CONs

as a FORMAL system
probability theory relies on a series of axioms, e.g. a *measurable space* of events

PROs of probability theory

- clarity of the theoretical framework,
- proven practical value

…and CONs

as a FORMAL system

probability theory relies on a series of axioms, e.g. a *measurable space* of events

but our experience of the world defies this closure

PROs of probability theory

- clarity of the theoretical framework,
- proven practical value

…and CONs

as a FORMAL system
probability theory relies on a series of axioms, e.g. a *measurable space* of events

but our experience of the world defies this closure

as a MODELLING framework
several cognitive patterns (often called biases or fallacies) are not predicted by probability theory

PROs of probability theory

- clarity of the theoretical framework,
- proven practical value

…and CONs

as a FORMAL system
probability theory relies on a series of axioms,
e.g. a *measurable space* of events

but our experience of the world defies this closure

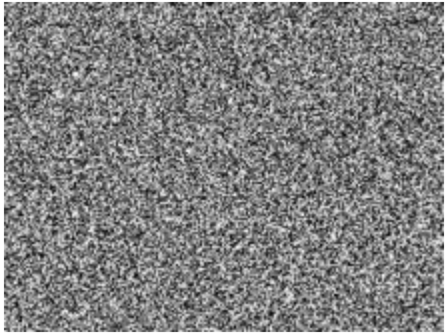
as a MODELLING framework
several cognitive patterns (often called biases or fallacies) are not predicted by probability theory

in particular, there is a mismatch in what humans perceive as *informative* w.r.t. Shannon's notion of information

Simplicity Theory

- Simplicity Theory (ST) is a computational model of cognition whose investigation started by observing the "informativity" mismatch.

NOISE SOURCE:
maximally informative following Shannon's theory of information



SO WHAT?

Simplicity Theory

- Simplicity Theory (ST) is a computational model of cognition whose investigation started by observing the "informativity" mismatch.

- ST predicts diverse human phenomena related to *relevance:*
    - *unexpectedness*
    - *narrative interest*
    - *coincidences*
    - *near-miss experiences*
    - *emotional interest*
    - *responsibility*

- ST has been used for experiments in *artificial creativity.*

Simplicity Theory: formal background

- Formally, ST builds upon Algorithmic Information Theory (AIT).

Simplicity Theory: formal background

- Formally, ST builds upon Algorithmic Information Theory (AIT).

- In AIT, the *complexity* of a string is the minimal length of a program that, given a certain optional input parameter, produces that string as an output (**Kolmogorov complexity**)

$$K_\phi(x|y) = \min_p \left\{ |p| : p(y) = x \right\}$$

executable program

target string

additional input in support

underlying
Turing machine

Simplicity Theory: formal background

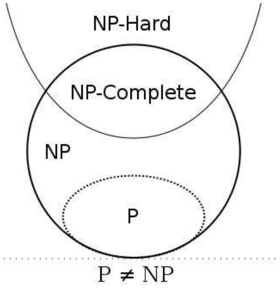
- Formally, ST builds upon Algorithmic Information Theory (AIT).

- In AIT, the *complexity* of a string is the minimal length of a program that, given a certain optional input parameter, produces that string as an output (**Kolmogorov complexity**)

how much information is needed for a program constructing the object

$\neq$

how much time or space is needed for running it (algorithmic or time-complexity)



$$K_\phi(x|y) = \min_p \{|p| : p(y) = x\}$$

executable program

target string

additional input in support

underlying Turing machine

Simplicity Theory: formal background

- Formally, ST builds upon Algorithmic Information Theory (AIT).

- In AIT, the *complexity* of a string is the minimal length of a program that, given a certain optional input parameter, produces that string as an output (**Kolmogorov complexity**)

$$K_\phi(x|y) = \min_p \{|p| : p(y) = x\}$$

executable program

target string

additional input in support

underlying
Turing machine

- Kolmogorov complexity is generally incomputable (due to the halting problem), but it is computable on ***bounded Turing machines***.

Simplicity Theory: formal background

- Formally, ST builds upon Algorithmic Information Theory (AIT).

- In AIT, the *complexity* of a string is the minimal length of a program that, given a certain optional input parameter, produces that string as an output (**Kolmogorov complexity**)

$$K_\phi(x|y) = \min_p \{|p| : p(y) = x\}$$

executable program

underlying Turing machine

target string

additional input in support

- Kolmogorov complexity is generally incomputable (due to the halting problem), but it is computable on ***bounded Turing machines***.

We denote bounded complexities with C.

Unexpectedness

- ST starts from the observation that humans are highly susceptible to *complexity drops*, ie. for them

  situations are ***relevant*** if they are *simpler* **to describe** than **to explain**

Unexpectedness

- ST starts from the observation that humans are highly susceptible to *complexity drops*, ie. for them

  situations are ***relevant*** if they are *simpler* **to describe** than **to explain**

- Formally, this is captured by the formula of unexpectedness, expressed as divergence of complexity computed on two distinct machines

$$U(s) = C_W(s) - C_D(s)$$

situation

causal complexity
via world machine

description complexity
via description machine

Unexpectedness

- ST starts from the observation that humans are highly susceptible to *complexity drops*, ie. for them

  situations are ***relevant*** if they are *simpler* **to describe** than **to explain**

- Formally, this is captured by the formula of unexpectedness, expressed as divergence of complexity computed on two distinct machines

$$U(s) = C_W(s) - C_D(s)$$

situation

causal complexity
via world machine

description complexity
via description machine

$$\overbrace{\text{world} \rightarrow \text{situation}}^{C_W}$$

$$\overbrace{\text{situation} \leftarrow \text{mind}}^{C_D}$$

Unexpectedness: examples

- **remarkable lottery draws**: 11111 is more unexpected than 64178, even if the lottery is fair
- **coincidences**: meeting by chance an old friend from yours abroad is more unexpected than meeting there any random unknown person.
- **deterministic yet unexpected events**: e.g. a lunar eclipse

$$U(s) = C_W(s) - C_D(s)$$

situation

causal complexity
via world machine

description complexity
via description machine

$$\overbrace{\text{world} \rightarrow \text{situation}}^{C_W}$$

$$\overbrace{\text{situation} \leftarrow \text{mind}}^{C_D}$$

Aim of the paper

- Provide further arguments in support to non-probabilistic computational models in cognition, in particular focusing on the following:

conjecture

Bayes' rule is a specific instantiation of a more general template captured in ST by Unexpectedness

Bayes' rule

- From the definition of conditional probability:

$$p(O \cap M) = p(M|O) \cdot p(O) = p(M) \cdot p(O|M)$$

we can obtain the formula of Bayes simply:

$$p(M|O) = \frac{p(M \cap O)}{p(O)} = \frac{p(O|M) \cdot p(M)}{p(O)}$$

model

observation

often informally rewritten as:

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

Unexpectedness as posterior

- In previous works, it has been hypothesized that ST’s Unexpectedness offers as *non-extensional* measure of *posterior subjective probability*:

$$\text{posterior} = 2^{-U}$$

Unexpectedness as posterior

- In previous works, it has been hypothesized that ST's Unexpectedness offers as *non-extensional* measure of *posterior subjective probability*:

$$\text{posterior} = 2^{-U}$$

- Starting from this hypothesis, we looked for a mapping from Unexpectedness to Bayes' rules, and indeed we see that:

$$p(M|O) = \frac{p(O|M) \cdot p(M)}{p(O)}$$

$$\overbrace{\log \frac{1}{p(M|O)}}^{U(s)} = \log \frac{p(O)}{p(O|M) \cdot p(M)} = \overbrace{\log \frac{1}{p(O|M)} + \log \frac{1}{p(M)}}^{C_W(s)} - \overbrace{\log \frac{1}{p(O)}}^{C_D(s)}$$

Unexpectedness as posterior

- In previous works, it has been hypothesized that ST’s Unexpectedness offers as *non-extensional* measure of *posterior subjective probability*:

$$\text{posterior} = 2^{-U}$$

- Starting from this hypothesis, we looked for a mapping from Unexpectedness to Bayes’ rules, and indeed we see that:

$$p(M|O) = \frac{p(O|M)\cdot p(M)}{p(O)}$$

problem: 1 parameter with unexpectednes, 2 with posterior

$$\overbrace{\log\frac{1}{p(M|O)}}^{U(s)} = \log\frac{p(O)}{p(O|M)\cdot p(M)} = \overbrace{\log\frac{1}{p(O|M)} + \log\frac{1}{p(M)}}^{C_W(s)} - \overbrace{\log\frac{1}{p(O)}}^{C_D(s)}$$

Unexpectedness as posterior

- In previous works, it has been hypothesized that ST's Unexpectedness offers as *non-extensional* measure of *posterior subjective probability*:

$$\text{posterior} = 2^{-U}$$

- Starting from this hypothesis, we looked for a mapping from Unexpectedness to Bayes' rules, and indeed we see that:

$$p(M|O) = \frac{p(O|M) \cdot p(M)}{p(O)}$$

problem: 1 parameter with unexpectednes, 2 with posterior

$$\overbrace{\log \frac{1}{p(M|O)}}^{U(s)} = \log \frac{p(O)}{p(O|M) \cdot p(M)} = \overbrace{\log \frac{1}{p(O|M)} + \log \frac{1}{p(M)}}^{C_W(s)} - \overbrace{\log \frac{1}{p(O)}}^{C_D(s)}$$

let's investigate these two terms…

Causal complexity

$$\overbrace{\log\frac{1}{p(O|M)}+\log\frac{1}{p(M)}}^{C_W(s)}-\overbrace{\log\frac{1}{p(O)}}^{C_D(s)}$$

- The causal complexity is the length in bits of the shortest path that, according to the agent's world model, ***produces*** the situation.

Causal complexity

$$\overbrace{\log\frac{1}{p(O|M)} + \log\frac{1}{p(M)}}^{C_W(s)} - \overbrace{\log\frac{1}{p(O)}}^{C_D(s)}$$

- The causal complexity is the length in bits of the shortest path that, according to the agent's world model, ***produces*** the situation.

- The causal path is temporally unfolded. The chain rule has the form:

$$C_W(c * s) = C_W(s||c) + C_W(c)$$

sequential composition **causal link**

(implicit: from the current situation)

Causal complexity

$$\overbrace{\log\frac{1}{p(O|M)}+\log\frac{1}{p(M)}}^{C_W(s)}-\overbrace{\log\frac{1}{p(O)}}^{C_D(s)}$$

- The causal complexity is the length in bits of the shortest path that, according to the agent's world model, ***produces*** the situation.

- The causal path is temporally unfolded. The chain rule has the form:

$$C_W(c*s)=C_W(s||c)+C_W(c)$$

sequential composition **causal link**

(implicit: from the current situation)

- Being a Kolmogorov complexity, the cause can be omitted if it lies on the shortest path

$$C_W(s)=\min_c C_W(c*s)=\min_c\left[C_W(s||c)+C_W(c)\right]$$

Causal complexity

$$\overbrace{\log \frac{1}{p(O|M)} + \log \frac{1}{p(M)}}^{C_W(s)} - \overbrace{\log \frac{1}{p(O)}}^{C_D(s)}$$

- The causal complexity is the length in bits of the shortest path that, according to the agent's world model, ***produces*** the situation.

- The causal path is temporally unfolded. The chain rule has the form:

$$C_W(c * s) = C_W(s||c) + C_W(c)$$

sequential composition **causal link**

(implicit: from the current situation)

- Being a Kolmogorov complexity, the cause can be omitted if it lies on the shortest path

$$C_W(s) = \min_c C_W(c * s) = \min_c [C_W(s||c) + C_W(c)]$$

➡ *the Unexpectedness formula abstracts the **causally explanatory** factor*

Description complexity

$$\underbrace{\log\frac{1}{p(O|M)} + \log\frac{1}{p(M)}}_{C_W(s)} - \underbrace{\log\frac{1}{p(O)}}_{C_D(s)}$$

- The description complexity is the length in bits of the shortest program that, leveraging mental resources, ***determines*** the situation

Description complexity

$$\underbrace{\log \frac{1}{p(O|M)} + \log \frac{1}{p(M)}}_{C_W(s)} - \underbrace{\log \frac{1}{p(O)}}_{C_D(s)}$$

- The description complexity is the length in bits of the shortest program that, leveraging mental resources, ***determines*** the situation
  - e.g. determination could correspond to retrieve the situation from memory, so informationally, we need to specify the address where to look at (an ***encoding***)

Description complexity

$$\overbrace{\log \frac{1}{p(O|M)} + \log \frac{1}{p(M)}}^{C_W(s)} - \overbrace{\log \frac{1}{p(O)}}^{C_D(s)}$$

- The description complexity is the length in bits of the shortest program that, leveraging mental resources, ***determines*** the situation
  - e.g. determination could correspond to retrieve the situation from memory, so informationally, we need to specify the address where to look at (an ***encoding***)

- In the proposed mapping, $C_D(s)$ corresponds to $p(O)$, the probability of observing that situation.

➡ a theoretical link can be then established through ***optimal encoding*** in Shannon's terms, where probability is *assessed through frequency.*

Description complexity

$$\overbrace{\log \frac{1}{p(O|M)} + \log \frac{1}{p(M)}}^{C_W(s)} - \overbrace{\log \frac{1}{p(O)}}^{C_D(s)}$$

- The description complexity is the length in bits of the shortest program that, leveraging mental resources, ***determines*** the situation
  - e.g. determination could correspond to retrieve the situation from memory, so informationally, we need to specify the address where to look at (an ***encoding***)

- In the proposed mapping, $C_D(s)$ corresponds to $p(O)$, the probability of observing that situation.

➡ a theoretical link can be then established through ***optimal encoding*** in Shannon's terms, where probability is *assessed through frequency.*

- Complexity is however a more general measure, as it allows us to consider compositional effects (eg. à la Gestalt) via adequate mental operations

Bayes' rule vs Unexpectedness

- Bayes' rule is a specific instantiation of ST's Unexpectedness that:
  - makes a candidate "cause" explicit and does not select automatically the best one
  - takes a frequentist-like approach for encoding observables.

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

$$U(s) = \min_c \overbrace{\left[C_W(c * s) - C_D(s)\right]}^{\text{posterior}} = \min_c \left[\overbrace{C_W(s||c)}^{\text{likelihood}} + \overbrace{C_W(c)}^{\text{prior}} - \overbrace{C_D(s)}^{\text{evidence}}\right]$$

Why is this relevant?

- Unexpectedness is a more generally applicable measure.
- In the paper we show that it can be used to build:
    - an informational principle of framing
    - a model of derived likelihood
    - an explanation of the prosecutor’s fallacy

All prior is posterior of some other prior

- Let us consider an additional prior in Bayes’ formula, a sort of ‘environmental context’. Following probability theory we have two equivalent formulations for the posterior:

$$p(M|O,E) = \frac{p(M \cap O|E)}{p(O|E)} = \frac{p(M \cap O \cap E)}{p(O \cap E)}$$

All prior is posterior of some other prior

- Let us consider an additional prior in Bayes' formula, a sort of 'environmental context'. Following probability theory we have two equivalent formulations for the posterior:

$$p(M|O,E) = \frac{p(M \cap O|E)}{p(O|E)} = \frac{p(M \cap O \cap E)}{p(O \cap E)}$$

- These formulations are **not equivalent** when expressed in complexity terms!

$$C_W(c * s||e) - C_D(s|e) \qquad\qquad C_W(e * c * s) - C_D(e * s)$$

All prior is posterior of some other prior

- Let us consider an additional prior in Bayes' formula, a sort of 'environmental context'. Following probability theory we have two equivalent formulations for the posterior:

$$p(M|O,E) = \frac{p(M \cap O|E)}{p(O|E)} = \frac{p(M \cap O \cap E)}{p(O \cap E)}$$

- These formulations are **not equivalent** when expressed in complexity terms!

$$C_W(c * s||e) - C_D(s|e) \qquad\qquad C_W(e * c * s) - C_D(e * s)$$

abstracting c as before

$$C_W(s||e) - C_D(s|e) \equiv U(s||e) \qquad\qquad C_W(e * s) - C_D(e * s) = U(e * s)$$

All prior is posterior of some other prior

- Let us compute the difference between the two formulations:

$$U(e * s) - U(s||e) = C_W(e * s) - C_D(e * s) - C_W(s||e) + C_D(s|e)$$

All prior is posterior of some other prior

- Let us compute the difference between the two formulations:

$$U(e*s) - U(s||e) = C_W(e*s) - C_D(e*s) - C_W(s||e) + C_D(s|e)$$

- Two distinct chain rules apply on the world and description machines:

$$C_W(e*s) = C_W(e) + C_W(s||e) \qquad C_D(e*s) \leq C_D(e) + C_D(s|e)$$

All prior is posterior of some other prior

- Let us compute the difference between the two formulations:

$$U(e * s) - U(s||e) = C_W(e * s) - C_D(e * s) - C_W(s||e) + C_D(s|e)$$

- Two distinct chain rules apply on the world and description machines:

$$C_W(e * s) = C_W(e) + C_W(s||e) \qquad C_D(e * s) \leq C_D(e) + C_D(s|e)$$

describing *e* and *s* together may be simpler than fully determining one term before the other (cf. **informed search**)

the temporal constraint is dropped

All prior is posterior of some other prior

- Let us compute the difference between the two formulations:

$$U(e * s) - U(s||e) = C_W(e * s) - C_D(e * s) - C_W(s||e) + C_D(s|e)$$

applying the chain rules…

$$U(e * s) - U(s||e) \geq C_W(e) - C_D(e) = U(e)$$

All prior is posterior of some other prior

- Let us compute the difference between the two formulations:

$$U(e * s) - U(s||e) = C_W(e * s) - C_D(e * s) - C_W(s||e) + C_D(s|e)$$

applying the chain rules…

$$U(e * s) - U(s||e) \geq C_W(e) - C_D(e) = U(e)$$

➡ a necessary condition for which the two formulations may be equivalent is that the contextual prior is *not* unexpected. $U(e) \approx 0$

All prior is posterior of some other prior

- Let us compute the difference between the two formulations:

$$U(e*s) - U(s||e) = C_W(e*s) - C_D(e*s) - C_W(s||e) + C_D(s|e)$$

applying the chain rules…

$$U(e*s) - U(s||e) \geq C_W(e) - C_D(e) = U(e)$$

➡ a necessary condition for which the two formulations may be equivalent is that the contextual prior is *not* unexpected. $U(e) \approx 0$

shared facts, defaults, and also improbable but descriptively complex situations

All prior is posterior of some other prior

- Let us compute the difference between the two formulations:

$$U(e * s) - U(s||e) = C_W(e * s) - C_D(e * s) - C_W(s||e) + C_D(s|e)$$

applying the chain rules…

$$U(e * s) - U(s||e) \geq C_W(e) - C_D(e) = U(e)$$

➡ a necessary condition for which the two formulations may be equivalent is that the contextual prior is *not* unexpected.

$$U(e) \approx 0$$

shared facts, defaults, and also
improbable but descriptively complex situations

informational principle of framing

all contextual situations which are not unexpected provide grounds to be neglected; the remaining situations provide the "relevant" context for the situation in focus.

Derived likelihood

- Following ST, we do not have direct access to the causal complexity, as we need always to pass through a descriptive step to identify what to compute.

$$U(s) = C_W(s) - C_D(s)$$

Derived likelihood

- Following ST, we do not have direct access to the causal complexity, as we need always to pass through a descriptive step to identify what to compute.

$$U(s) = C_W(s) - C_D(s)$$

- So, how can we estimate likelihood? Counting back the description complexity!

$$C_W^U(s||c) = U(s||c) + C_D(s|c)$$

Derived likelihood: examples

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

Derived likelihood: examples

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$$C_D \approx 0$$

because these elements are just in front of me

$$C_W \gg 0$$

because this never occurred

Derived likelihood: examples

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$C_D \approx 0$ — because these elements are just in front of me

$C_W \gg 0$ — because this never occurred

$\Rightarrow$

$U \approx C_W \gg 0$ — it is implausible (if it occurred)

$C_W^U = U + C_D \gg 0$ — it is improbable (to occur)

Derived likelihood: examples

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$C_D \approx 0$

because these elements are just in front of me

$C_W \gg 0$

because this never occurred

$\Rightarrow$

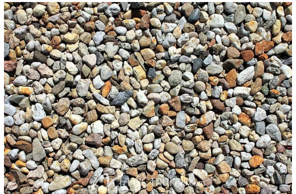
$U \approx C_W \gg 0$

it is implausible (if it occurred)

$C_W^U = U + C_D \gg 0$

it is improbable (to occur)

- The likelihood that a stone in the world moves if I close the door:



Derived likelihood: examples

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$C_D \approx 0$

because these elements are just in front of me

$C_W \gg 0$

because this never occurred

$U \approx C_W \gg 0$

it is implausible (if it occurred)

$C_W^U = U + C_D \gg 0$

it is improbable (to occur)

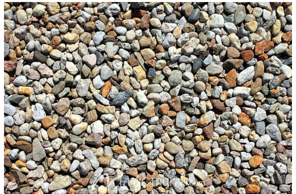
- The likelihood that a stone in the world moves if I close the door:

$C_D \gg 0$

because I need to specify of which stone I am talking

$C_W \gg 0$

because this never occurred



Derived likelihood: examples

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$C_D \approx 0$ because these elements are just in front of me

$C_W \gg 0$ because this never occurred

$\Rightarrow$

$U \approx C_W \gg 0$ it is implausible (if it occurred)

$C_W^U = U + C_D \gg 0$ it is improbable (to occur)

- The likelihood that a stone in the world moves if I close the door:

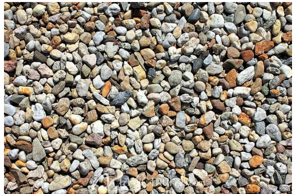
$C_D \gg 0$ because I need to specify of which stone I am talking

$C_W \gg 0$ because this never occurred

$\Rightarrow$

$U \approx 0$ it is plausible (if it occurred)

$C_W^U = U + C_D \gg 0$ it is improbable (to occur)



Derived likelihood: examples

- Consider the estimation of the likelihood that the wall changes colour if I close the door:

$C_D \approx 0$ — because these elements are just in front of me

$C_W \gg 0$ — because this never occurred

$\Rightarrow$

$U \approx C_W \gg 0$ — it is implausible (if it occurred)

$C_W^U = U + C_D \gg 0$ — it is improbable (to occur)

- The likelihood that a stone in the world moves if I close the door:

$C_D \gg 0$ — because I need to specify of which stone I am talking

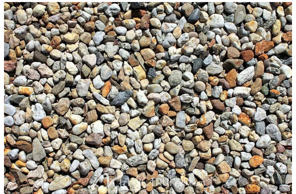
$C_W \gg 0$ — because this never occurred

$\Rightarrow$

$U \approx 0$ — it is plausible (if it occurred)

$C_W^U = U + C_D \gg 0$ — it is improbable (to occur)

NOTE: If the stone e.g. is in the room or was already described, we return to the first case!



Prosecutor's fallacy

- Suppose that, following forensic studies, the probability that a certain DNA evidence appears if the defendant is guilty is deemed very high.

Prosecutor's fallacy

- Suppose that, following forensic studies, the probability that a certain DNA evidence appears if the defendant is guilty is deemed very high.

- The **prosecutor's fallacy** occurs when the probability that the defendant is guilty (given that there is DNA evidence) is also concluded to be comparatively high.

$$p(O|M) \approx 1 \rightsquigarrow p(M|O) \approx 1 \qquad \text{[Prosecutor's fallacy]}$$

Prosecutor's fallacy

- Suppose that, following forensic studies, the probability that a certain DNA evidence appears if the defendant is guilty is deemed very high.

- The **prosecutor's fallacy** occurs when the probability that the defendant is guilty (given that there is DNA evidence) is also concluded to be comparatively high.

$$p(O|M) \approx 1 \rightsquigarrow p(M|O) \approx 1 \qquad \text{[Prosecutor's fallacy]}$$

$$p(M|O) = \frac{p(M \cap O)}{p(O)} = \frac{p(O|M) \cdot p(M)}{p(O)}$$

this is a fallacy as it neglects the base rates

Prosecutor's fallacy: an explanation

- Let us reframe the problem in terms of complexity, introducing the definition of *causally constrained unexpectedness*, computed before the selection of the best cause in unexpectedness:

$$U_c(s) = C_W(c * s) - C_D(s) \qquad U(s) = \min_d U_d(s)$$

maps to **posterior**

$$p(M|O)$$

Prosecutor's fallacy: an explanation

- Let us reframe the problem in terms of complexity, introducing the definition of *causally constrained unexpectedness*, computed before the selection of the best cause in unexpectedness:

$$U_c(s) = C_W(c * s) - C_D(s) \qquad U(s) = \min_d U_d(s)$$

maps to
posterior

$$p(M|O)$$

maps to
likelihood

$$p(O|M)$$

- Applying the chain rule:

$$U_c(s) = C_W(c * s) - C_D(s) = C_W(s||c) + C_W(c) - C_D(s)$$

Prosecutor's fallacy: an explanation

- Let us reframe the problem in terms of complexity, introducing the definition of *causally constrained unexpectedness*, computed before the selection of the best cause in unexpectedness:

$$U_c(s) = C_W(c * s) - C_D(s) \qquad U(s) = \min_d U_d(s)$$

- Applying the chain rule:

$$U_c(s) = C_W(c * s) - C_D(s) = C_W(s||c) + C_W(c) - C_D(s)$$

≈ 0 because $p(O|M) \approx 1$

Prosecutor's fallacy: an explanation

- Let us reframe the problem in terms of complexity, introducing the definition of *causally constrained unexpectedness*, computed before the selection of the best cause in unexpectedness:

$$U_c(s) = C_W(c * s) - C_D(s) \qquad U(s) = \min_d U_d(s)$$

If the procurator finds plausible that the suspect is guilty:

$$U(c) = C_W(c) - C_D(c) \approx 0$$

- Applying the chain rule:

$$U_c(s) = C_W(c * s) - C_D(s) = \underbrace{C_W(s||c)}_{\approx 0} + \underbrace{C_W(c)}_{\approx C_D(c)} - C_D(s)$$

Prosecutor's fallacy: an explanation

- Let us reframe the problem in terms of complexity, introducing the definition of *causally constrained unexpectedness*, computed before the selection of the best cause in unexpectedness:

$$U_c(s) = C_W(c * s) - C_D(s) \qquad U(s) = \min_d U_d(s)$$

Considering the limited number of suspects and proximity to the victim: $C_D(c) \approx C_D(s)$

- Applying the chain rule:

$$U_c(s) = C_W(c * s) - C_D(s) = \underbrace{C_W(s||c)}_{\approx 0} + \underbrace{C_W(c)}_{\approx C_D(c)} - C_D(s)$$

Prosecutor's fallacy: an explanation

- Let us reframe the problem in terms of complexity, introducing the definition of *causally constrained unexpectedness*, computed before the selection of the best cause in unexpectedness:

$$U_c(s) = C_W(c * s) - C_D(s) \qquad U(s) = \min_d U_d(s)$$

Considering the limited number of suspects and proximity to the victim: $C_D(c) \approx C_D(s)$

- Applying the chain rule:

$$U_c(s) = C_W(c * s) - C_D(s) = \underset{\approx 0}{C_W(s||c)} + \underset{\approx C_D(c)}{C_W(c)} - C_D(s) \quad \approx 0$$

Prosecutor's fallacy: an explanation

- Let us reframe the problem in terms of complexity, introducing the definition of *causally constrained unexpectedness*, computed before the selection of the best cause in unexpectedness:

$$U_c(s) = C_W(c * s) - C_D(s) \qquad U(s) = \min_d U_d(s)$$

- Applying the chain rule:

$$U_c(s) = C_W(c * s) - C_D(s) = C_W(s||c) + C_W(c) - C_D(s) \quad \approx 0$$

$$\boxed{C_W(s||c) \approx 0 \rightsquigarrow U_c(s) \approx 0 \quad \text{[Prosecutor's fallacy]}}$$

Conclusions

- The proposed conjecture provides further arguments in support to non-probabilistic computational models of cognition.
- A complexity-based account allows distinguishing between relevant and irrelevant contextual elements, while the probabilistic account treats them equally.
- Remaining open questions is how the underlying machines should be defined.
- Yet, the abstraction level of algorithmic information theory is already relevant to draw insights on cognitive processes, as we have shown here eg. with the analysis of the prosecutor’s fallacy.

Unexpectedness and Bayes' rule

6 December 2021, CIFMA workshop

Giovanni Sileno
g.sileno@uva.nl

University of Amsterdam

Jean-Louis Dessalles
jean-louis.dessalles@telecom-paris.fr

Télécom Paris -- Institut Polytechnique de Paris