Reasoning as confirmation-seeking hypothesis testing

Mathias Sablé Meyer & Salvador Mascarenhas

December 2017

You can find a digital version of this handout here:
http://web-risc.ens.fr/~msable/handout_XPhi_19-Dec-2017.pdf

Reasoning fallacies

Introductory examples

Access your intuitions and don't try to outsmart the problems:

- "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball."

  How much does the ball cost?

- Let me introduce you to Mary:

  – Mary has met every king or every queen of Europe.
  – Mary has met the king of the Netherlands.

  Does it follow that *Mary has met the king of Spain*?

- Alice is looking at Bob but Bob is looking at Carol. Alice is married but Carol is not.

  – Is any *married* person looking at any *unmarried* person?

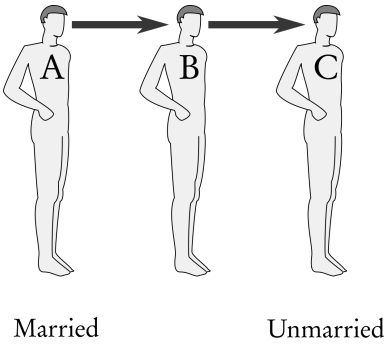

Figure 1: Alice, Bob and Carol

Distinction

- **Compelling fallacies** are (classically) *invalid* inference patterns that we often *accept*.
- **Repugnant validities** are (classically) *valid* inference patterns that we often *reject*.

An interesting case: Linda

> "Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."[1]

Which is more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

[1] Tversky and Kahneman (1983)

Notice that probability theory tells us that

$$\forall A, \forall B, (A \subset B) \Rightarrow P(B) > P(A)$$

i.e. if A is a subset of B, then the probability of B is greater than the one of A[2]

[2] Intuitively, this is saying that if you're a bank teller and active in a feminist movement, then in particular you are a bank teller: truth maker for proposition 2. also satisfy 1. while the converse is not true

Illusory Inference From Disjunction (IIFD)

Our variation on a fallacy discovered by Johnson-Laird:

- p_1 : Either John speaks English and Marie speaks French or Bill speaks German.
- p_2 : John speaks English.
- q : Does it follow that Marie speaks French?

Acceptance rate $\approx 85\%$[3], and it's **not valid**.

[3] While modus ponens has an acceptance rate $\approx 90\%$!

Introducting ETR

ETR: Erotetic[4] Theory of Reasoning — an account of non classicality in human reasoning: logical *or* forks worlds and raises questions.

[4] From Ancient Greek ἐρωτητικός. Of or pertaining to questioning. See: Interrogatory.

Intuitive idea

ETR is a set of reasoning rules[5] that model a reasoner having the following property: propositions are treated sequentially and whenever an *or* appears, it is taken as a question that needs answering.

Such a set of rules can have various application strategy, the big picture is that the *laziest* ones lead to (many) fallacies whereas the most expansive ones are operationally similar to *classical logic*.

[5] see appendix

IIFD

- Lean posterior/rational gambler:

  "Picks whatever maximises the posterior"

- Dynamic posterior:

  "Picks the option whose posterior increased the most during the update"

- Lean confirmation:

  "Picks whatever confirms more the hypotheses"

- Dynamic confirmation:

  "Picks whatever increased most during the learning process of beliefs about the world."

Confirmation you say?

Answers the question "how much does some information p *confirms* some hypothesis q?"[6]

[6] Intuition: p confirms q if knowing p increases the belief you have in q — to what extent is to be defined

Traditional measures

$$
\begin{aligned}
D(h,e) &= p(h|e) - p(h) \\
R(h,e) &= \ln\left(\frac{p(h|e)}{p(h)}\right) \\
L(h,e) &= \ln\left(\frac{p(e|h)}{p(e|\neg h)}\right) \\
C(h,e) &= p(e \wedge h) - p(h) \times p(e) \\
S(h,e) &= p(h|e) - p(h|\neg e) \\
Z(h,e) &= \begin{cases} \frac{p(h|e) - p(h)}{1 - p(h)} & \text{if } p(h|e) \geq p(h) \\ \frac{p(h|e) - p(h)}{p(h)} & \text{otherwise} \end{cases}
\end{aligned}
$$

"Confirmation measures" assign a value to what we want to naïvely call confirmation. Even if something is very unlikely, it may be that some other piece of evidence *confirms* it.

Therefore they share the idea that they should look at variations around difference/ratios between the hypothesis and the hypothesis knowing the evidence — see their structure.

Let's apply this to the limit case of Linda's problem:

- let's assume that being a bank teller and being a feminist are independent events
- and that the introductory text is a *perfect* confirmation of "being a feminist" but is completely orthogonal to "being a bank teller".

We compare $C(b \wedge f, s)$ to $C(b, s)$ — where b is "bank teller", f is "feminist" and s is the story.

For D this leads to the following equalities:

$$
\begin{aligned}
D(b \wedge f, s) &= P(b \wedge f \mid s) - P(b \wedge f) \\
&= P(b) - P(b)P(f) \\
&= P(b)(1 - P(f))
\end{aligned}
\qquad\qquad
\begin{aligned}
D(b, s) &= P(b \mid s) - P(b) \\
&= P(b) - P(b) \\
&= 0
\end{aligned}
$$

Thus as long as $P(b) \neq 0$ and $P(f) \neq 1$, D's account for Linda's problem matches expectations.

The same intuition applies to R:

$$
\begin{aligned}
R(b \wedge f, s) &= \log\left(\frac{P(b \wedge f \mid s)}{P(b \wedge f)}\right) \\
&= \log\left(\frac{P(b)}{P(b)P(f)}\right) \\
&= \log\left(P(f)^{-1}\right) \\
&= -\log(P(f))
\end{aligned}
\qquad
\begin{aligned}
R(b, s) &= \log\left(\frac{P(b \mid s)}{P(b)}\right) \\
&= \log\left(\frac{P(b)}{P(b)}\right) \\
&= \log(1) \\
&= 0
\end{aligned}
$$

$P(f) \in [0,1] \Rightarrow -\log(P(f)) \in [0,\infty]$ thus as long as $P(f) \neq 1$ the result holds for R too.

The remaining cases are left as an exercise to the reader, with the expected result that they all matches observed reasoning behaviour and favour $b \wedge f$.

IIFD — An erotetic account

There is a crucial difference between confirmation theories and the Erotetic one. In the most general case, the structure of the problem is the following:

1. $\varphi \vee \psi$
2. θ

- χ? χ'?

Since the ETR sees *or*s as questions, it choses, for some confirmation measure $C(\cdot,\cdot)$, between $C(\varphi,\theta)$ and $C(\psi,\theta)$, *irrelevant of the question.*

Conversely, confirmation measure are of two kind:

- The *lean* ones that try to pick the best between $C(\chi,(\varphi \wedge \psi) \wedge \theta)$ and $C(\chi',(\varphi \wedge \psi) \wedge \theta)$[7]
- The *dynamic* ones that try to pick the best between $C(\chi,(\varphi \wedge \psi) \wedge \theta) - C(\chi,(\varphi \wedge \psi))$ and $C(\chi', p_1 \wedge \theta) - C(\chi', p_1)$[8]

[7] Intuition: what conclusion best confirms the data?

[8] Intuition: what conclusion increased most during the processing of the information — **pragmatics** comes into play!

Erotetics: do we need them?

	Regular		Reversed	
	Dynamic	Flat	Dynamic	Flat
Exp. Design				
p_1	$(A \wedge H) \vee B$	$(A \wedge H) \vee (A \wedge B)$	$A \wedge \neg C$	$(A \wedge H \wedge \neg C) \vee (A \wedge \neg C \wedge B)$
p_2	A		$(A \wedge H) \vee B$	
q	$H? \vee C?$	$H? \vee C?$	$H? \vee S?$	$H? \vee S?$
Predictions				
Erotetic				
$\mathcal{H}_{ERT}$	$\boldsymbol{H_{Ceil}}$	∅	$\boldsymbol{H}$	∅
Posterior				
$\mathcal{H}_{P,Lean}$	∅	∅	∅	∅
$\mathcal{H}_{P,Dyna}$	$\boldsymbol{H}$	∅	∅	∅
Con. Lean				
$\mathcal{H}_{C=D,Lean}$	∅	∅	∅	∅
$\mathcal{H}_{C=R,Lean}$	∅	∅	∅	∅
Con. Dynamic				
$\mathcal{H}_{C=D,Dyna}$	$\boldsymbol{H}$	∅	∅	∅
$\mathcal{H}_{C=R,Dyna}$	$\boldsymbol{H}$	∅	∅	∅

- In the "Regular" one all *lean* theories fail to predict any difference
- In the "Reversed" one all *but* ETR fail to predict any difference

Thus experimental/behavioural difference will rule out theories.

Experimental results

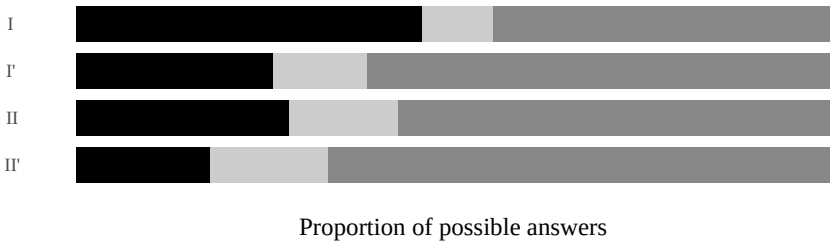

Figure 2: Proportion of answers for each condition. N = 48 tested on mTurk, all participants. Black indicates answers according to ETR, light gray the other answer, and dark gray the escape option "there is no best choice"

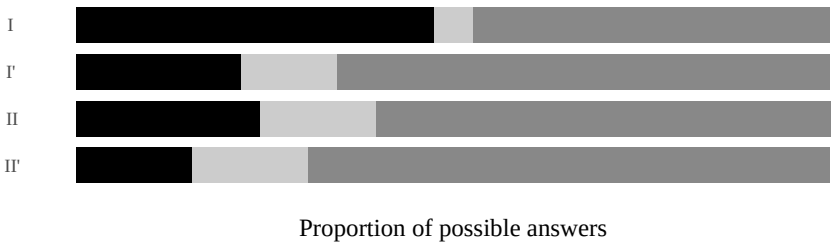

Figure 3: Proportion of answers for each condition. N = 39 tested on mTurk, filtered on questionnaire about cards knowledge.

Important take away results:

- Any lean version is not enough (χ^2-test between I and I' has a p-value < 0.005, not significant for II vs II')

- I' and II' exhibit similar patterns — II' slightly harder
- I to II shows a decrease of about 15% which is expected[9]

[9] Simpler experiments show that reversing the order in the IIFD indeed leads to similar drops.

Main issue: people are bad with cards. We get interesting trends if we rule out people that never gave "crazy" answers[10]

[10] See "In the absence of any relevant information, I'll say that diamond is more likely than black".

Further exploration of IIFD

Another way to address the IIFD is to look at some form of pattern matching of worlds: upon, hearing $a \vee b$, one creates two separate worlds — one for a and one for b — further information will help decide which world to look at.

In the case of the IIFD, it relied and a blurry notion of "pattern matching" — at some level, p_2 points to one side of p_1 more than the other and one goes for it.

An experiment was run to rule out all *low level*[11] accounts of this nature: if pattern matching is right account then it needs to be at a conceptual level.

[11] auditory, syntactical, etc.

The goal was to test whether the following cousin of the IIFD was indeed a fallacy:

1. $(a \wedge b) \vee c$
2. d

- Does it follows that b?

As a function of the strength of $d \Rightarrow a$ *in the absence of any context*[12].

[12] Denise D. Cummins, *Naive theories and causal deduction* in Memory and Cognition 1995, 23

Experiment 1

The first part of the task was to gather the intuitive strength of various entailments of the shape $d \Rightarrow a$, as well as to control for the two associated entailments $a \Rightarrow b$ and $d \Rightarrow b$.

	D ent. A		A ent. B		D ent. B		Max
ID	M.	Std. D.	M.	Std. D.	M.	Std. D.	Std. E.
1	3.66	1.33	0.58	1.08	0.78	1.28	0.09
2	4.34	1.35	0.41	1.05	0.42	1.11	0.09
3	4.89	1.28	0.35	1.04	0.39	1.09	0.08
4	4.36	1.35	0.38	1.02	0.42	1.14	0.09
5	4.37	1.30	0.39	1.00	0.38	1.00	0.08
6	4.78	1.38	0.37	0.99	0.39	1.04	0.09
7	3.33	1.38	0.37	1.11	0.32	0.95	0.09
8	3.05	1.44	0.48	1.04	0.58	1.11	0.09

With ID, for $D \Rightarrow A$, in the following order:

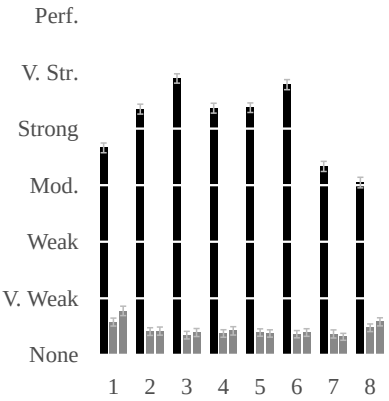

Figure 4: In black are the $D \Rightarrow A$ ratings. The two other are respectively $A \Rightarrow B$ and $D \Rightarrow B$. Black has higher rating and higher variance — by design.

1. If fertilizer was put on the plants, then the plants grew quickly.
2. If the brake was depressed, then the car slowed down.
3. If Mary jumped into the swimming pool, then Mary got wet.
4. If the trigger was pulled, then the gun fired.
5. If Larry grasped the glass with his bare hands, then Larry left fingerprints on his glass.
6. If the gong was struck, then the gong sounded.
7. If John studied hard, then John did well on the test.
8. If the apples were ripe, then the apples fell from the tree.

Experiment 2

This is the IIFD with $p_2 = d$ as shown previously. The results from Exp. 1 were rescaled to range in $[0, 1]$.

	Exp. 2		Exp. 1
ID	Yes	No	Rating
2	0.53	0.47	0.72
3	0.67	0.33	0.81
4	0.64	0.36	0.73
5	0.69	0.31	0.73
6	0.71	0.29	0.80
7	0.45	0.55	0.56
8	0.38	0.62	0.51

The first result is that the proportion of *Yes*s is above (classically) wrong answers in controls:

- $\approx 10\%$ for simple modus ponens controls
- $\approx 20\%$ for controls with the following structure:

  1. $(a \wedge b) \Rightarrow c$
  2. $\neg a$

  – Does it follows that c?

Proportions: 0.581% go for the fallacy. A linear regression of the structure (Fallacy Acceptance $\sim$ Rating) yields a significant p-value for the slope. A sketch is given on the right, and bellow is the table for this regression:

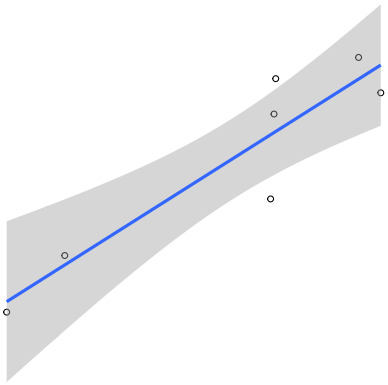

Figure 5: Intuition of the correlation between the rating and the acceptance rate.

	Estimate	Std. Err.	t-value	p-value
(Intercept)	-0.119	0.134	-0.893	0.413
MeanRating	1.010	0.191	5.300	0.003

There is a ceiling effect — where?

Indeed, if we extrapolate this result to the ideal case where the rating would be "Perfect", that is to say the *usual* IIFD with $D = A$, we know that the acceptance rate is around 90%. The present estimation gives rise to a somewhat perfect 605.84% acceptance rate.

Conversely, the intercept is not significant so we should assume 0 to be the acceptance rate of a completely orthogonal case. But there are 10% of participants that *do not* accept modus ponens, we should expect a flooring effect around this.

Is the ceiling effect in the rating[13] or in the IIFD[14]?

[13] people *will always* find a link between any A and D

[14] people do crazy things with logic

Appendix

ETR Derivation examples

Formalising the *IIFD*:

- $p_1 : (a \wedge b) \vee c$
- $p_2 : a$
- q : Is it true that b?

In the ETR, we'll have $\Gamma = \{a \sqcup b, c \sqcup d\}$ and $\Delta = \{a\}$.

$$\begin{array}{lcl} \Gamma[\Delta]^{\mathsf{Q}} & = & \{a \sqcup b\} \\ \Gamma[\Delta]^{\mathsf{Q}}[\{b\}]^{\mathsf{MR}} & = & \{b\} \end{array}$$

ETR: what about success?

There are many possible derivation, let's try another one with the same premises:

$$\begin{array}{lcl} \Gamma[\{a\}]^{\mathsf{Inq}} & = & \{a \sqcup b, c \sqcup d \sqcup a, c \sqcup d \sqcup \neg a, a \sqcup b \sqcup \neg a\} \\ \Gamma[\{a\}]^{\mathsf{Inq}}[\cdot]^{\mathsf{F}} & = & \{a \sqcup b, c \sqcup d \sqcup a, c \sqcup d \sqcup \neg a\} \\ \Gamma[\{a\}]^{\mathsf{Inq}}[\cdot]^{\mathsf{F}}[\{a\}]^{\mathsf{Up}} & = & \{a \sqcup b, c \sqcup d \sqcup a\} \end{array}$$

No $[\cdot]^{\mathsf{MR}}$ will give you $\{b\}$ as there is no way to select the first conjonction with $\{a\}$ nor with Γ.

A way to imagine what is happening here is to consider that if the subject properly looks at all the possible cases, which $[\cdot]^{\mathsf{Inq}}$ forces him to do, then he will make no such logical mistake as deriving $\{b\}$.

What we said is that mistakes are possible, i.e. there are *wrong* derivations, which does not entails that every derivation leads to logical errors.

ETR rules

C(onjunctive)-Update

$$\begin{array}{lcl} \Gamma[\Delta]^{\mathsf{C}} & = \Gamma \times \Delta \\ & = \{\gamma \sqcup \delta : \gamma \in \Gamma \ \& \ \delta \in \Delta\} \end{array}$$

C-Update pairwise combines each element of Γ with each element of Δ. It incorporates the new information in Δ into Γ.

Q(uestion)-Update

$$\Gamma[\Delta]^{\mathsf{Q}} = \Gamma - \{\gamma \in \Gamma : (\sqcap\Delta) \sqcap \gamma = 0\}$$

Q-Update eliminates from Γ (the "question") all alternatives that have *nothing* in common with *the intersection* of all alternatives in Δ. In other words: take the information in Δ, that is the intersection of all alternatives in Δ. Keep in Γ only those alternatives that share some mental molecule with the information in Δ.

Update

$$\Gamma[\Delta]^{\mathsf{Up}} = \begin{cases} \Gamma[\Delta]^{\mathsf{C}} & \text{if } \Gamma[\Delta]^{Q} = \emptyset \\ \Gamma[\Delta]^{\mathsf{Q}}[\Delta]^{\mathsf{C}} & \text{otherwise} \end{cases}$$

The complete Update procedure first *tests* whether Δ provides an answer to the question in Γ by attempting a Q-Update. If it *doesn't* (i.e.

Q-update returns $\emptyset$), then Update performs a simple C-Update, incorporating the new information in Δ. If it *does*, then Update keeps the (possibly only partly) answered question and C-Updates with Δ, in case Δ provides some new information *beside* providing an answer to Γ.

Molecular Reduction

$$\Gamma[\alpha]^{\mathsf{MR}} = \begin{cases} (\Gamma - \{\gamma \in \Gamma : \alpha \sqsubseteq \gamma\}) \cup \{\alpha\} & \text{if } (\exists\gamma \in \Gamma)\alpha \sqsubseteq \gamma \\ \textit{undefined} & \text{otherwise} \end{cases}$$

Molecular Reduction of Γ on a mental molecule α reduces every alternative in Γ that contains α to α alone. It is undefined in case no alternative in Γ contains α. It amounts to *disjunct simplification* ($(\phi \wedge \psi) \vee \theta \iff \phi \vee \theta$), and as a special case it allows for conjunction elimination.

Filter

$$\Gamma[\cdot]^{\mathsf{F}} = \{\mathrm{dne}(\gamma) : \gamma \in \Gamma \ \&\ \neg\mathrm{contr}(\gamma)\}$$

Filter eliminates all contradictory alternatives in Γ by testing for the presence, within an alternative, of a molecule α and its negation (this is the function $\mathrm{contr}(\cdot)$). Further, it eliminates double negations from the surviving alternatives ($\mathrm{dne}(\cdot)$).

Inquire

$$\Gamma[\Delta]^{\mathsf{Inq}} = \Gamma[\Delta \cup \mathrm{neg}(\Delta)]^{\mathsf{C}}[\cdot]^{\mathsf{F}}$$

Inquire performs a simple conjunctive update (NB: no Q-Update) with a mental model Δ and its negation, followed by filtering out any contradictory alternatives and removing double negations.

Mental Model Negation

For Γ a mental model, notice that $\Gamma = \{\alpha_0, \ldots, \alpha_n\}$ and for each $\alpha_i \in \Gamma$ we have that $\alpha_i = \bigsqcup\{a_{i0}, \ldots, a_{im_i}\}$, for $m_i + 1$ the number of mental model nuclei in α_i. Now,

$$\mathrm{Neg}(\Gamma) = \mathrm{Neg}(\{\alpha_0, \ldots, \alpha_n\}) = \{\neg a_{00}, \ldots, \neg a_{0m_0}\} \times \cdots \times \{\neg a_{n0}, \ldots, \neg a_{nm_n}\}$$

Double negation elimination

$$\mathrm{dne}(a) = \begin{cases} b & \text{if } a = \neg\neg b \text{ for some } b \in \mathit{Atoms}(\mathcal{M}) \\ a & \text{otherwise} \end{cases}$$

$$\mathrm{dne}(\alpha) = \bigsqcup \{\mathrm{dne}(a) : a \in \mathit{Atoms}(\mathcal{M}) \;\&\; a \sqsubseteq \alpha\}$$