EVIDENTIAL REASONING

Posterior probability and impact

All ***inductive arguments*** (intended here as all *non-valid arguments*, thus including not only *descriptive* inductions but also *explanatory* ones, i.e., the so-called *abductive arguments*) can be analyzed from (at least) **two complementary perspectives**, which focus either on the hypothesis of interest or on the available evidence, respectively

Evidence → *e*

Hypothesis → *h*

(Note: for our purposes, it is not relevant whether the conclusion of this argument is a *generalization* or an *explanation*, but that it does not follow logically from the premise)

Within a Bayesian framework, these issues have been identified with two related but distinct notions:

(i) the ***degree of belief*** in a hypothesis *h* in light of evidence *e* $p(h|e)$

(ii) the ***degree of support*** provided by evidence *e* to a hypothesis *h* $c(h, e)$ or $imp(h, e)$

Qualitative notion of impact

The support of evidence *e* on hypothesis *h* is expressed by the notion of **BAYESIAN CONFIRMATION** (or **IMPACT**)

$$c(h, e) \begin{cases} c(h, e) > 0 \quad iff \quad p(h|e) > p(h) & [confirmation] \\ c(h, e) = 0 \quad iff \quad p(h|e) = p(h) & [neutrality] \\ c(h, e) < 0 \quad iff \quad p(h|e) < p(h) & [disconfirmation] \end{cases}$$

Confirmation… a somewhat poor terminological choice (much better impact) because:

- the technical meaning of *confirmation* departs from that of natural language, in which it usually implies to validate or ascertain
- *confirmation* only conveys the idea of positive support while, of course, impact can be negative as well (*disconfirmation*)
- in the psychological literature, the term *confirmation* has gained a negative connotation because of so-called *confirmation bias* (Nickerson, 1998), a tendency to suboptimal reasoning depending on one's target hypothesis

Quantifying impact

$$\mathrm{d}(h,e) = \mathrm{p}(h|e) - \mathrm{p}(h)$$

(Eells, 1982; Jeffrey, 1992)

$$\mathrm{l}(h,e) = \log \frac{\mathrm{p}(e|h)}{\mathrm{p}(e|\neg h)}$$

(Good, 1950; 1984)

$$\mathrm{r}(h,e) = \log \frac{\mathrm{p}(h|e)}{\mathrm{p}(h)}$$

(Keynes, 1921; Horwich, 1982)

$$\mathrm{z}(h,e) = \begin{cases} \dfrac{\mathrm{p}(h|e) - \mathrm{p}(h)}{\mathrm{p}(\neg h)} & \text{if } \mathrm{p}(h|e) \geq \mathrm{p}(h) \\ \dfrac{\mathrm{p}(h|e) - \mathrm{p}(h)}{\mathrm{p}(h)} & \text{if } \mathrm{p}(h|e) < \mathrm{p}(h) \end{cases}$$

(Crupi, Tentori, & Gonzalez, 2007)

(and many others...)

Impact and posterior probability can be dissociated

Example (from Tentori, Chater, & Crupi, 2016)

Imagine a sample of 200 students, 100 males and 100 females

You draw at random one of these students, let's call this student "X"

	X is a male --- X likes cigars		X is a male --- X likes going to the cinema
UCL sample	$pr(h\|e) = .38$	<	$pr(h\|e) = .95$
	$pr(h\|\neg e) = .12$	<<	$pr(h\|\neg e) = .95$
	$c(h,e) > 0$	>	$c(h,e) = 0$

Impact and posterior probability are both necessary to account for how inductive reasoning operates

For example, consider a murder trial

- for a verdict to be reached, the probability of the hypothesis “guilty” in light of all the evidence presented must be judged

- however, a juror may also be interested in estimating the *probative value* of a given piece of evidence (e.g., infidelity) in order to decide on its admissibility

  In legal terms, “probative” refers to evidence which, if true, would increase the likelihood of a conclusion (Good, 1996; David & Follette, 2002) and this is a matter of evidence assessment (cannot be expressed by a single probability value)

How are people's impact judgements?

Assessing impact

Good news

Participants properly estimate the impact of certain evidence on a given hypothesis (even when their probability judgments are far from correct)

Participants properly estimate the impact of uncertain evidence

Moreover, classic fallacies in probabilistic reasoning can be explained in terms of confirmation (see slides above)

A new general hypothesis

In dealing with everyday uncertainty, **people may rely more on detecting relations of inductive confirmation** (i.e., the net impact of new evidence on the credibility of the hypotheses concerned) **than on values of posterior probability** (i.e., the overall credibility of hypotheses as updated on all given evidence)

Tentori, K., Chater, N., & Crupi, V. (2016). Judging the probability of hypotheses versus the impact of evidence: Which form of inductive inference is more accurate and time-consistent? *Cognitive Science*, *40*, 758–778.

Aim of the study

To compare the **accuracy** and the **reliability** of confirmation vs. probability judgments in a test-retest experiment

Experimental questions, dependent variables and predictions

1. Within-subjects consistency (test-retest reliability)

   Which between confirmation and probability judgments are more reliable over time?

   ⇨ Confirmation judgments should be more consistent

2. Accuracy

   Which between confirmation and probability judgments are more accurate?

   ⇨ Confirmation judgments should be more accurate

Experimental procedure and stimuli

Preliminary phase

We asked 200 undergraduates (100 females and 100 males), randomly selected at UCL, to fill in a survey with various personal questions, as the followings:

Do you have a driving license? Do you have freckles? Can you ski? Do you own (at least) one plant? Do you own (at least) one videogame console? Do you support any football team? Do you like going to the cinema? Etc.

Responses were used to compute frequency based probabilities (e.g., the probability of owning (at least) one videogame console in light of the evidence of being male/female) and corresponding objective confirmation values (e.g., the impact of the evidence of being male/female on the hypothesis of owning (at least) one videogame console) according to the following models from the literature: the likelihood ratio (l), the probability ratio (r) and the relative distance measure (z)

Experimental procedure and stimuli

Stimuli construction

2 pieces of evidence
e = male or ¬*e* = female

selection of
28 hypotheses

} 56 arguments

Example

evidence → *X is a male*

X owns (at least) one videogame console

hypothesis →

Stimuli

posteriors vs. impacts

	$Pr(h \mid e) > .5$ and $Pr(h \mid \neg e) > .5$	$Pr(h \mid e) > .5$ and $Pr(h \mid \neg e) < .5$	$Pr(h \mid e) < .5$ and $Pr(h \mid \neg e) > .5$	$Pr(h \mid e) < .5$ and $Pr(h \mid \neg e) < .5$
$C(h,e) > 0$ $C(h,\neg e) < 0$	e = X is a **male** h= X has a **driving licence** h= X owns (at least) one **bike** h = X can play **volleyball** e = X is a **female** h = X likes **tea** h = X likes **carrots** h = X likes **shopping**	e = X is a **male** h = X can play **poker** h = X supports a **football team** h = X likes **beer** h = X can play **football** h = X own (at least) a **videogame console** h = X can play **basketball** e = X is a **female** h= X likes **ice-figure skating** h = X likes **candles** h = X worked as a **babysitter** h = X own (at least) one **cuddle toy** h = X likes reading **fashion magazines** h = X can **dance**		e = X is a **male** h = X likes **cigars** h = X can **surf** h = X **snores** e = X is a **female** h = X owns (at least) one **plant** h = X has **freckles** h = X owns (at least) one **weighting scale**
$C(h,e) = 0$ $C(h,\neg e) = 0$	e = X is a **male** h = X owns (at least) a **mp3 player** h = X likes going to the **cinema** e = X is a **female** h= X owns (at least) a **mp3 player** h = X likes going to the **cinema**			e = X is a **male** h = X has his/her **own website** h = X has (at least) **3 siblings** e = X is a **female** h = X has his/her **own website** h = X has (at least) **3 siblings**
$C(h,e) < 0$ $C(h,\neg e) > 0$	e = X is a **female** h = X has a **driving licence** h = X owns (at least) one **bike** h = X can play **volleyball** e = X is a **male** h = X likes **tea** h = X likes **carrots** h = X likes **shopping**		e = X is a **female** h = X can play **poker** h = X supports a **football team** h = X likes **beer** h = X can play **football** h = X own (at least) a **videogame console** h = X can play **basketball** e = X is a **male** h = X likes **ice-figure skating** h = X likes **candles** h = X worked as a **babysitter** h = X own (at least) one **cuddle toy** h = X likes reading **fashion magazines** h = X can **dance**	e = X is a **female** h = X likes **cigars** h = X can **surf** h = X **snores** e = X is a **male** h = X owns(at least) one **plant** h = X has **freckles** h = X owns (at least) one **weighting scale**

Experimental procedure and stimuli

Experimental phase

A new random sample of 35 UCL undergraduates (mean age 22.43 years; 21 females) participated in the experiment; they received £10 for their participation

Participants came twice to the laboratory, and on both occasions were asked to make 56 probability and 56 confirmation judgments

Participants were randomly divided in two groups, 19 were presented with a discrete probability scale and a continuous confirmation scale, while the other 16 were presented with a continuous probability scale and a discrete confirmation scale

To control for possible carry-over effects, the order of probability and confirmation questions was balanced across participants

Group 1:

Confirmation judgments - continuous scale

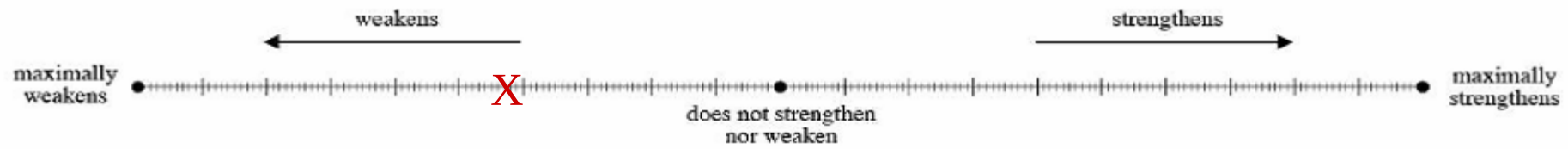
Consider a group of 200 students, 100 males and 100 females, randomly selected at UCL.
Imagine we draw at random one of these 200 students. Let's call this student X.

Consider the following hypothesis (possibly true or false) concerning X:
X owns (at least) one videogame console.

Now you are given a new piece of information (surely true) concerning X:
X is female.

How does this new piece of information (i.e., that *X is female*) affect
the hypothesis under consideration (i.e., that *X owns (at least) one videogame console*)?

The information that
*X is **FEMALE***



the hypothesis that
*X **OWNS (AT LEAST) ONE VIDEOGAME CONSOLE***

Group 1:

Probability judgments - discrete scale

Consider a group of 200 students, 100 males and 100 females, randomly selected at UCL.

How many of the 100 **female** students ***OWN (AT LEAST) ONE VIDEOGAME CONSOLE***? 27

Group 2:

Confirmation judgments - discrete scale

Consider a sample of 200 students, 100 males and 100 females, randomly selected at UCL.
Imagine we draw at random some students from this sample (one at a time, with replacement).

For each drawn student you will be presented with a hypothesis (possibly true or false).
After that, you will be given a new piece of information (surely true).

Your task is to indicate the impact of this new piece of information on the hypothesis under consideration.

Express your opinion about the impact of the given information on the hypothesis under consideration indicating a number between **- 50** ("the information **maximally weakens** the hypothesis") and **+ 50** ("the information **maximally strengthens** the hypothesis").
Use **0** to indicate **no impact at all** ("the information **does not weaken or strengthen** the hypothesis **even a little**").

A student X is drawn at random from a sample of 200 UCL students, 100 males and 100 females

Hypothesis (possibly true or false): *X owns (at least) one videogame console*

Information (surely true): *X is female*

From – 50 to + 50, the impact of the information that *X is* ***FEMALE***
on the hypothesis that *X* ***OWNS (AT LEAST) ONE VIDEOGAME CONSOLE*** is: - 22

Group 2:

Probability judgments - continuous scale

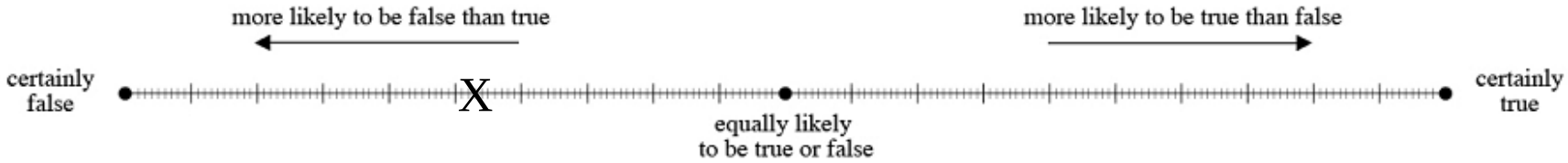
Consider a group of 200 students, 100 males and 100 females, randomly selected at UCL.
Imagine we draw at random one of these 200 students. Let's call this student X.

Consider the following hypothesis (possibly true or false) concerning X:
X owns (at least) one videogame console.

Now you are given a new piece of information (surely true) concerning X:
X is female.

In light of this new piece of information (i.e., that *X is female*),
what is the probability of the hypothesis under consideration (i.e., that *X owns (at least) one videogame console*)?

In light of the information that *X is* ***FEMALE***,
the hypothesis that *X* ***OWNS (AT LEAST) ONE VIDEOGAME CONSOLE*** is



posteriors vs. impacts

Results: **consistency**

As a general consistency measure, we correlated each participant's judgments in the first session with his/her corresponding judgments in the second session (N= 56)

Average correlations

Participants	**Confirmation**	**Probability**
Group 1 (1-19)	.90	.87
Group 2 (20-35)	.91	.85
All (1-35)	**.91**	**.86**

$t(34)=3.72 \quad (p < .01)$

This result suggests that both judgments are generally consistent, however confirmation judgments are more consistent than probability ones

(For both probability and confirmation judgments, there is no gender difference in consistency, neither with regards to the 56 judgments taken altogether nor the 28 judgments which employ "male" vs. "female" as evidence)

Results: accuracy 1

As a general accuracy measure, we **correlated** each participant's confirmation and probability judgments in the first session **with the corresponding objective values** (N= 56)

Average correlations

Participants	Confirmation			Probability
	l	z	r	
Group 1 (1-19)	.83	.80	.77	.67
Group 2 (20-35)	.82	.79	.76	.49
All (1-35)	**.82**	**.80**	**.77**	**.59**

t(34)= 11.14 t(34)= 10.16 t(34)= 8.06

(for all p<.001)

Correlations are much higher for confirmation (no matter which measure is employed to quantify it) than probability judgments

Note: l and z are the models that better predicted judgments of evidential impact in previous studies (e.g., Tentori et al, 2007; Crupi, et al, 2007)

Results: accuracy 2

We also computed the **absolute average error** in confirmation and probability judgments made by each participant (N= 56)

Average error (on a 100-point scale)

Participants	Confirmation			Probability
	l	**z**	**r**	
Group 1 (1-19)	9.59	9.85	11.67	18.79
Group 2 (20-35)	11.95	12.29	14.83	21.61
All (1-35)	**10.67**	**10.97**	**13.11**	**20.08**

t(34)= 16.79 t(34)= 17.13 t(34)= 9.73

(for all p<.001)

Results: accuracy 3

We also quantified the degree of agreement between probability errors and the direction of impact judgments

We considered, for each participant, all the arguments (out of the 56) whose impact s/he had judged as different from zero

Then, we computed the absolute difference $|Pr_{obj} - Pr_{subj}|$ and assigned it a positive sign whenever the participant had judged the impact as positive [negative] and had overestimated [underestimated] the objective probability, a negative sign otherwise

Finally, we averaged all these differences taken with their sign

For the great majority (80%) of participants the index was positive

The sample mean across all the 35 participants is +6.1, which is statistically different from the assumed null value of 0 (one-sample t-test, $t(34)=6.087$, $p<.01$)

Thus, **when impact is positive [negative], on average, participants overestimated [underestimate] the corresponding posterior probability of 6%**

To wrap up:

- confirmation judgments are systematically more accurate and reliable than probability judgments
- confirmation judgments also predict of the direction of the errors in probability judgments

These results do not depend on the specific measure employed to quantify impact nor on the specific scale used to collect the judgments

Although, in the Bayesian tradition, confirmation is formally expressed as a function of probability, its cognitive assessment may be a primitive kind of judgment which is more reliable than that of probability

Why are people sensitive to confirmation relations while often inaccurate in probability estimates?

Some (complementary?) conjectures

a) The assessment of confirmation is crucial for many tasks, and in particular, those in which the value of information or the soundness of arguments has to be considered

b) Confirmation captures the relation between two variables, while probability is an absolute judgment

   Confirmation is therefore more suitable than probability to track more or less direct causal dependencies and, as a consequence, might achieve a higher degree of stability in response to changing background knowledge

c) ???

Imagine asking yourself, for example:

What is the probability of the hypothesis that “a person living in a certain place (e.g., UK, Sudan, …) has a tuberculosis infection (TBI)” given the evidence that “she is coughing up blood”?

To provide a precise answer seems quite hard, because even if you are aware that TBI can be rather common in patients who cough up blood, to quantify the exact probability requires more information about the spread of TBI and alternative diseases compatible with that symptom in that specific population.

What is the impact of the evidence that “a person living in a certain place (e.g., UK, Sudan, …) is coughing up blood” on the hypothesis that “she has a tuberculosis infection (TBI)”?

You know that, in most (if not all) circumstances, to cough up blood is valuable (albeit inconclusive) supporting evidence for the hypothesis of having TBI. Therefore, even without being expert on the specific population under consideration, you can conclude that this piece of evidence has a positive and relevant impact on the hypothesis at issue.

Ubiquity of evidential reasoning

Qualitative Adams' thesis:
A simple indicative conditional "if *A*, *B*" is *acceptable* to a person iff her degree of belief in *B* given *A*, $p(B|A)$, is high

Problematic sentences:

If Brazil ends first in the next FIFA world cup,
there will be a heads in the first 10 million tosses with this fair coin.

Douven & Verbrugge's (2012) **evidential support thesis**: A simple indicative conditional "if *A*, *B*" is acceptable to a person iff her degree of belief in *B* given *A*, $p(B|A)$, is not only high but also higher than $p(B)$.

Task: Judgments $p(B)$, of $p(B|A)$, and of the acceptability of "if *A*, *B*"

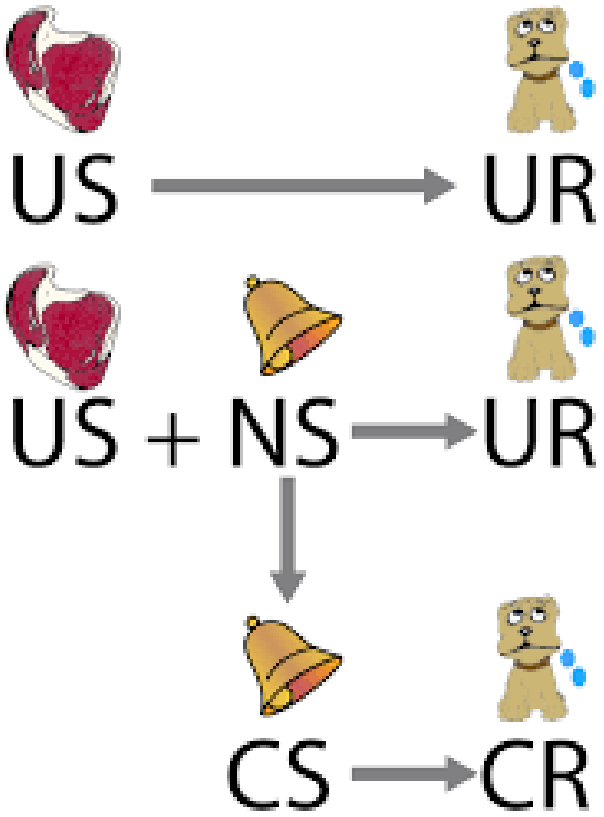
Results:

- a high value of $p(B|A)$ is not sufficient for the acceptability of "if *A*, *B*"
- there is a strong association between qualitative and quantitative confirmation (as measured with $d(B, A)$), and the acceptability of the conditional "if *A*, *B*"

Krzyżanowska, Collins & Hahn (2017) clarified that the oddity of so-called missing-link conditionals does not depend on natural language pragmatics as much as on **probabilistic relevance** (i.e., a conditional is assertable only if its antecedent is relevant for the consequent)

Learning: Classical Conditioning

Before conditioning



After conditioning

Rescorla (1968)

What is the **critical variable** for classical conditioning to occur?
The CS had to be a useful predictor of the US… but **what makes the CS a useful predictor**?

«**Contiguity**»

The number of times the CS is paired with the UCS

«**Contingency**» (information provided by the CS about the US)

Positive contingency: the CS signals an increase in the probability that the US will occur (compared to before the CS)
→ **excitatory conditioning**: the subject learns to perform a certain response, like salivating when the bell is rung

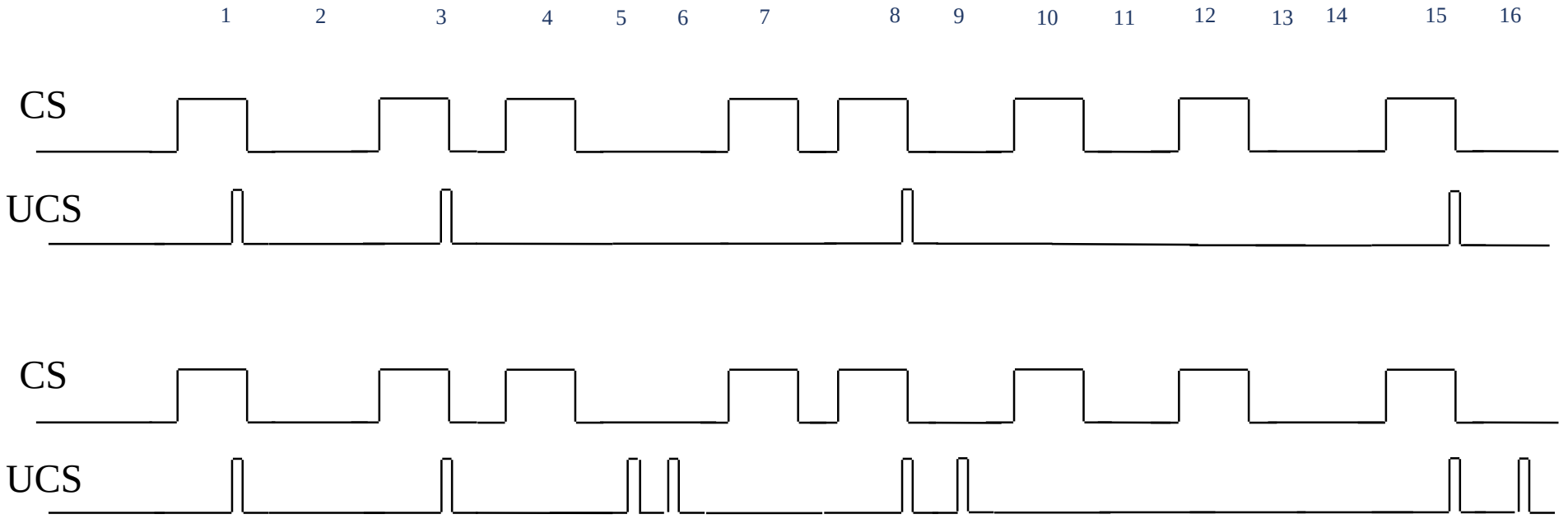
Zero contingency: CS predicts neither an increase nor a decrease in the probability of the US

Negative contingency: the CS signals a decrease in the probability that the US will occur (compared to before the CS)
→ **inhibitory conditioning:** the subject learns to withhold or suppress a certain response, like stop salivating when the bell rings (he salivates when the bell is not ringing)

According to Rescorla, **learning only takes place with the positive and negative contingencies**

Conditioning

p(UCS | CS) - p(UCS | No CS) = .5 – 0 = .5



Nozick's confirmation measure:

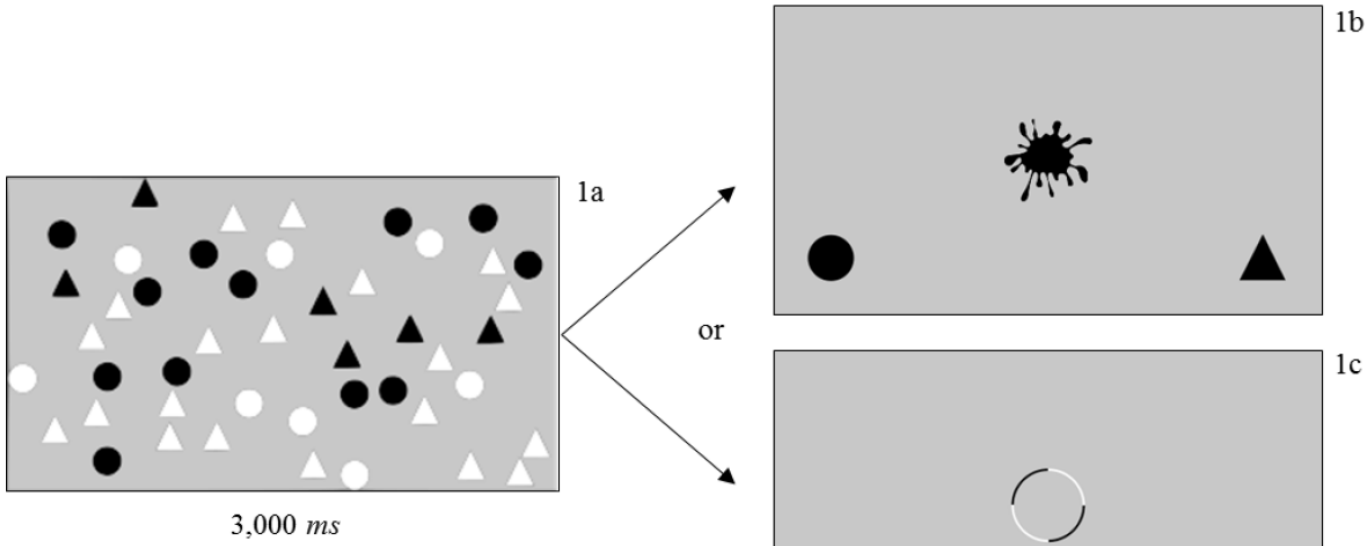
$$\mathrm{n}(h, e) = p(e|h) - p(e|\neg h)$$

p(UCS | CS) - p(UCS | No CS) = .5 – .5 = 0

NO Conditioning

and many other domains...

Perception
(Mangiarulo et al., 2019)



Memory
Bordalo, Coffman, Gennaioli, Schwerter & Shleifer (2019)

Stereotypes
Bathia (2017); Bordalo, Coffman, Gennaioli & Shleifer (2016)

The challenges for the future

1. To see how (and to what extent) the detection of impact relations affects tasks other than traditional inductive reasoning ones

2. To explain why people are more sensitive to impact than probability