behavioral and neural foundations of framing-effects

Sacha Bourgeois-Gironde, Élise Payzan, Raphael Giraud

Behavioral and Neural foundations of Framing Effects

Sacha Bourgeois-Gironde[1] Élise Payzan[2]

[1]École Normale Supérieure LSH (Lyon)
Institut Jean Nicod (Paris)

[2]École Normale Supérieure (Paris)
Economics Department

April 2005

Two main experimental setups

- Series of experiments to investigate the neural basis for violation of a rationality norm; our envisioned experimental design is twofold:
  - A **two-step procedure** resting on the «*bat and ball problem*» proposed by Shane Frederick (henceforth **BB**): the standard condition (duplication of Shane Frederick's experiment) plus a second condition with cues to trigger the analytical response
  - A **two-step procedure** using an adapted version of Kahneman and Tversky «*asian-disease problem*» (henceforth **FE**)
- Use of event-related fMRI methodology to pin down the neural mechanisms associated with specific behavioral patterns («wrong» responses)
- Comparison of the fMRI results: wrong response in BB and violation of invariance rule in FE $\longrightarrow$ Are similar neural mechanisms involved?

Focus on a possible feeling of «irrationality»

- Zero in on these patterns: in FE the subject retrieves her previous (inconsistent) response; in BB the intuitive response is provided after some hesitation $\longrightarrow$ hypothesis of a **«specific feeling of irrationality»** – that we label **«epistemic emotion»** – and search for possible neural correlates
- This **«epistemic emotion»** – if exists – should be *task-related*; must be distinguished with the purely *content-related* **«moral emotion»** triggered by some special features of the depicted situation $\longrightarrow$ We envision to minimize the latter hence use besides the original version of asian-disease (involving direct human death) some more neutral scenarios

Pin down the nature of error in BB

- **Assume** that we have replicated the aggregate result of Shane Frederick in our own two-step «*within-subjects*» design BB, namely we observe a massive rate of wrong responses
- How to interpret this? The «*dual process theory*» is likely to be an insightful benchmark ▸ Flesh out
- Natural candidate to explain the BB-type error: the «override failure» mechanism: insufficient activation of RPFC to correct the impulsive intuitive response $\longrightarrow$ We label it **the cognitive illusion hypothesis**

- «Competing» potential rationale for BB-type error: focus on a «computational defect» (low analytical capabilities) $\longrightarrow$ We call it **the cognitive limitation hypothesis**

Validity of the «*dual process theory*»

- **Assume** that we observe in our two-step FE experiment a massive rate of «*framing effect*»
- How to interpret this? As in BB the «override failure» (insufficient activation of RPFC to correct the impulsive intuitive response) might entail the observed error $\longrightarrow$ we define this as **the bounded rationality hypothesis**
- Alternative hypothesis: «*framing effect*» in FE is not due to a cognitive defect, but to a subjective non equivalence between the two options $\longrightarrow$ we call this **the «broadened rationality» hypothesis**
- Corresponding behavioral model: incorporate moral and psychological consequences in the set of choices of the subject
- Envisioned experimental devices (three) to try to answer this tricky question in this peculiar case

Definitions

- Henceforth – for more convenience – we will sum up the rationality debate in this (provisional) fashion:
  - Regarding BB: **bounded rationality hypothesis** Vs **limited rationality hypothesis** (cognitive limitations)
  - Regarding FE: **bounded rationality hypothesis** Vs **«broadened rationality» hypothesis**

- **«Bounded rationality»** $\Leftrightarrow$ «cognitive illusion»: we will not tend to classify as rationally bounded performances that are either inherently cognitively limited or emotionally driven

▸ More detail

- **«Broadened rationality»** as «broadened» consequentialism

Behavioral data in the BB problem: experimental design

- We contrast **two conditions**:
  - **Condition 1**: subjects give directly an answer to the BB task
  - **Condition 2**: subjects choose between three results which one is correct: 1.10 1 1.05
- Each subject performs the two tasks **in this order**
- **Condition 2** is expected to elicit a task-construal by S2

Behavioral data in the BB problem: prediction

- Test of the «cognitive illusion»: according to the **bounded rationality hypothesis**, subjects don't show any cognitive limitation when they answer intuitively the BB problem: they naturally substract 1 to the sum 1.10
- Direct implication: if true, then subjects should be **sensitive** to the alternative presentation of the problem
- Hence we claim that subjects **will perform better in experiment 2** if they are not victim of a «cognitive limitation» effect but of a «cognitive illusion» one

Neural data in the BB problem: factorial design and predictions (1 out of 3)

- The overall task (Cond 1 and Cond 2) is performed under fMRI and investigated through an event-related protocol
- We expect 3 **types of answers in Condition 2** and observe their neural substrates:
  - **intuitive response is immediately given**
  - **Conflict** between S1 ans S2:
    - **intuitive response is given, with a delay** $\longrightarrow$ detection without correction: S2 does not override
    - **«right» analytical response is given**: 1.05 $\longrightarrow$ detection and correction of the error: S2 overrides

Neural data in the BB problem: factorial design and predictions (2 out of 3)

- We predict steady activation of S1's specific substrates across the two tasks
- S1 activation is function of the nature of the stimuli and following Dehaesne and Spelke, we predict a significant activation of regions located in the **parietal lobe**, since **mathematical approximations** tasks are performed in BB.

Neural data in the BB problem: factorial design and predictions (3 out of 3)

- Regarding S2 activation, we look at the contrasts entailed by the two conditions:
  - Detection without correction in Cond 2 – Immediate response in Cond 1 $= \Delta_1 \longrightarrow$ we predict **RPFC specific activation** and **ACC significant activation** (conflict) if «dual process theory» verified
  - Detection and correction in Cond 2 – Immediate response in Cond 1 $= \Delta_2 \longrightarrow$ **RPFC and ACC specific activation**
  - Detection and correction in Cond 2 – Detection without correction in Cond 1 $= \Delta_2 \longrightarrow$ **RPFC specific activation**
  - Besides we would like to use a sort of «difference in difference» $\Delta_1 - \Delta_2$ to identify a **specific activation of limbic substrates** (mirroring the **«epistemic feeling»** of uneasiness when coping with the analytical task) $\longrightarrow$ **Right Anterior Insula specific activation?**

Basic strategy

- Hypothesis to be tested: the violation of the invariance rule is due to an extrinsic (moral) factor: neither a «cognitive limitation» effect, nor a «thinking style» bias (toward intuition)
- The procedure should be threefold:
  - **Preliminary step**: run the multi-version FE experiment and distinguish the scenarios such that «*framing effect*» is the most frequent pattern vis-a-vis the ones involving steady preferences across the two frames → see S L.Schneider (2000)
  - Second step: within the versions with «*framing effect*» **test «the robustness» of this violation** of the invariance rule, by making the latter directly accessible to the subject
  - **Statistical analysis** exploiting **individual differences**: positive correlation between an «intuitive thinking» pattern (see below) and the robustness of the violation?

First step: standard FE experiment

- We envision to use a different versions of the «asian-disease problem»: rate of death or live of humans, rate of failure or success of a commercial product → variation in the degree of moral content
- Use of a **«*within-subjects*» procedure**: each subject faces in turn the positive frame (henceforth PF) and the negative frame (NF) **with «filler» tasks** between frame pairs, **and sufficient temporal lag** between each of the two framing tasks; this to try to **bypass the potential carryover effect**

  ▸ Caveat
- In each session they are asked to rank the proposed options using the scale 1→4
- Zero in on the versions such that «framing effect» is the most frequent pattern (if some); second step restricted to these case

Second step: «Norm retrieval» test (1 out of 2)

- Second stage, the «norm retrieval» : the subject faces the overall setup, namely PF and NF are presented simultaneously
- Information delivering: **the experimenter informs the subject of the logical equivalence**
- Then the experimenter asks the subjects **in this order**:
  - How the average person would respond according to you?
  - How would you respond?
- This design aims to **bypass the possible «persistence bias» in the response** ▸ More detail

Second step: «Norm retrieval» test (2 out of 2)

- Intuition here: sort of «**Revealed Preference Argument**», echoing Slovic and Tversky's «*Understanding/Accepting principle*»
- We claim that **if framing effects stem from bounded rationality then by making the rational rule accessible, we should observe a correction of the framing effect**
- Conversely, if the rational rule is **not used** as a guide by the subject, **while the rule is available**, this implies that the logical invariance principle is dominated, among the different representations in the set of choices of the subject
- In this setup **the absence of significant retrieval** – given that our setup is designed to exclude the psychological dissonance concern – would be a piece of evidence in favor of the alternative «broadened» rationality hypothesis

Statistical analysis (1 out of 2)

- Use the same pool of subjects in BB and FE and exploit **individual differences** using the BB experiment's outputs
- Assume that the bounded rationality hypothesis was not rejected in BB experiment $\rightarrow$ **Cluster** the subjects of the BB experiment using the patterns of their responses in Cond 1 («analytical types» Vs «cognitive illusion victims»)
- Then look at a potential **positive correlation between the fact of being «analytical» in BB experiment and the propensity to retrieve the initial response in FE** (namely higher sensitivity to the «cue» provided)$\longrightarrow$ We predict the **presence** [resp **absence**] of such a significant link if **bounded rationality** [resp **«broadened» rationality**]

Statistical analysis (2 out of 2)

- Exploit individual differences through **a preliminary test designed to appraise the «thinking style» of the subjects**
- Preliminary test: **«thinking style» measure** through a self-report inventory plus some vignettes tasks inspired of Epstein's REI design
- Hence **cluster** the subjects: more «intuitive-experiential» Vs «more analytical-logical»
- Then look at a possible **positive correlation between the fact of being more «analytical» – according to the inventory – and the propensity to retrieve the initial response in FE** $\longrightarrow$ We predict the **presence** of such a significant link if the **bounded rationality hypothesis** is accurate

What we mean by «*dual process theory*»

- Here the label «*dual process theory*» refers to a coherent **corpus of theories** that provides insights to think of the cognitive defects in BB and FE; basically allows us to model some cases of cognitive defects as the upshot of a **conflict between System 1 and System 2 (S1 and S2)**
  - Kahneman's insights regarding **accessibility**: the error in BB mirrors the relative lack of accessibility of the analytical «guides»/representations (hence S2 does not override S1's response even though there is conflict) $\longrightarrow$ if true make this representation available for the subject and the error should not be robust
  - Error-Related-Negativity (ERN) approach: provides a **neural basis for this conflict** between S1 and S2.
- See the sketchy schema for a brief outline

back

What this definition of bounded rationality implies

- Underlying background: the related hypothesis regarding **massive modularity** («*hyperlocality*») of the mind
- We claim that our vision of bounded rationality calls for an **intermediary position** regarding modularity: observable phenomena of bounded rationality are **task-dependent**, but they never point to particular semantic contents
- Hence **Local property** of bounded rationality: «in-between», **neither absolutely rigid, nor content-related**
- Conversely performances that are inherently cognitively limited (memory, calculus) are completely rigid, and the ones that are emotionally driven are extremely local, namely depending on the content of the task

back

Caveat related to the «*within-subjects*» strategy in the standard FE experiment

- In the first step of the FE experiment, each subject is going to do successively the PF task and the NF one → within-subject procedure
- **Trade-off** regarding the choice of the experimental design: «*between-subjects*» design Vs «*within-subjects*» design
- Relative **gain** of the «*within-subjects*» design: «framing effect» for individuals, not an aggregate outcome; moreover this individual design allows the retrieval task
- Relative **cost**: big risk of bias, if the subject notices **the variation** in the frame (such variation if perceived is a **cue for the subject**, induced to be consistent in our experiment)
- **To minimize this drawback → sufficient temporal lag** between the frame pairs, and **«filler» tasks** presented during the time interval

These particular design is to reveal the «true» preference of the subjects

- We are aware of a **potential problem** associated with the «retrieval test»: if the subject was previously inconsistent in the FE experiment (namely provided a response in PF that is logically inconsistent with the one she gave in NF), might be inclined to provide **the same response pattern**; otherwise will suffer **psychological dissonance** $\longrightarrow$ possible **Persistence bias**
- We claim that **the first answer** (to the question: «*How would the average person answer?*») mirrors the «true» reaction of the subject (to wit the response without the dissonance effect); if **gap** between the first answer and the self-perspective $\longrightarrow$ piece of evidence towards the existence of **psychological dissonance** in this peculiar setup

back