

# City Research Online

## City St George's, University of London

**Citation:** Bessis, T. (2026). Judgment and reasoning in human and artificial cognition. (Unpublished Doctoral thesis, City St George's, University of London)

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** https://openaccess.city.ac.uk/id/eprint/38667/

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk


CITY
ST GEORGE'S
UNIVERSITY OF LONDON

£

£

£

£

Screen 1

You are offered a lottery ticket with:

- a 50% chance of winning nothing;
- a 50% chance of winning the two following rewards

Screen 2

Reward 1: A voucher to be used in a flower shop (worth £7)



Screen 3

Reward 2: A 14-inch MacBook Pro (worth £2400)



Screen 4

How much (in £) would you be willing to pay for this lottery ticket?

£

£

£ £

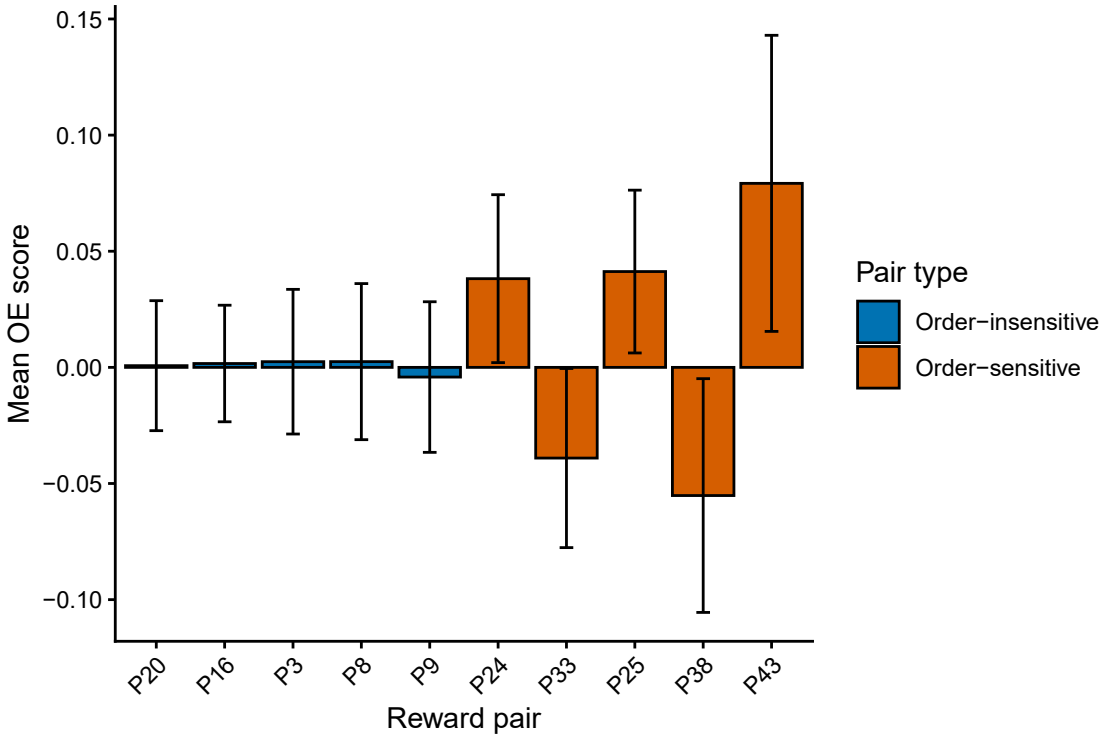
rescaled

$$\overline{\quad} \quad -\sum$$

---

---

---


0.15
0.10
0.05
0.00
−0.05
−0.10
Mean OE score
P20
P16
P3
P8
P9
P24
P33
P25
P38
P43
Reward pair
Pair type
Order−insensitive
Order−sensitive

£

£ £

£

£

£ £

£ £

£

£

£ £

£

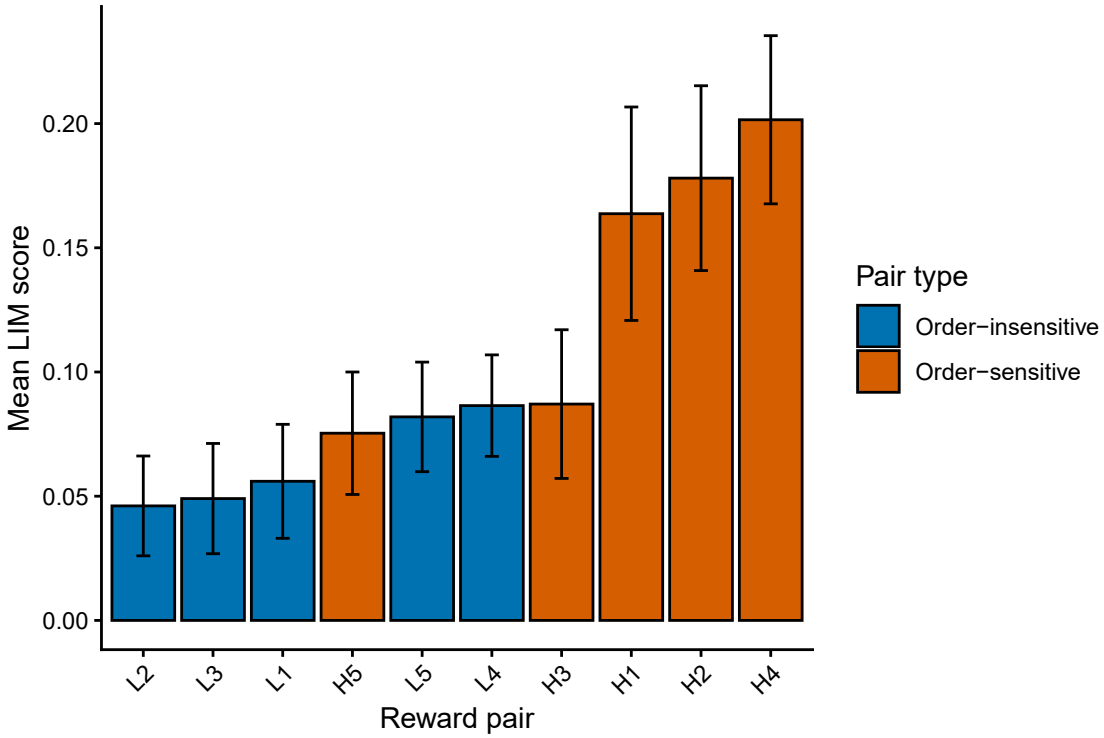
£

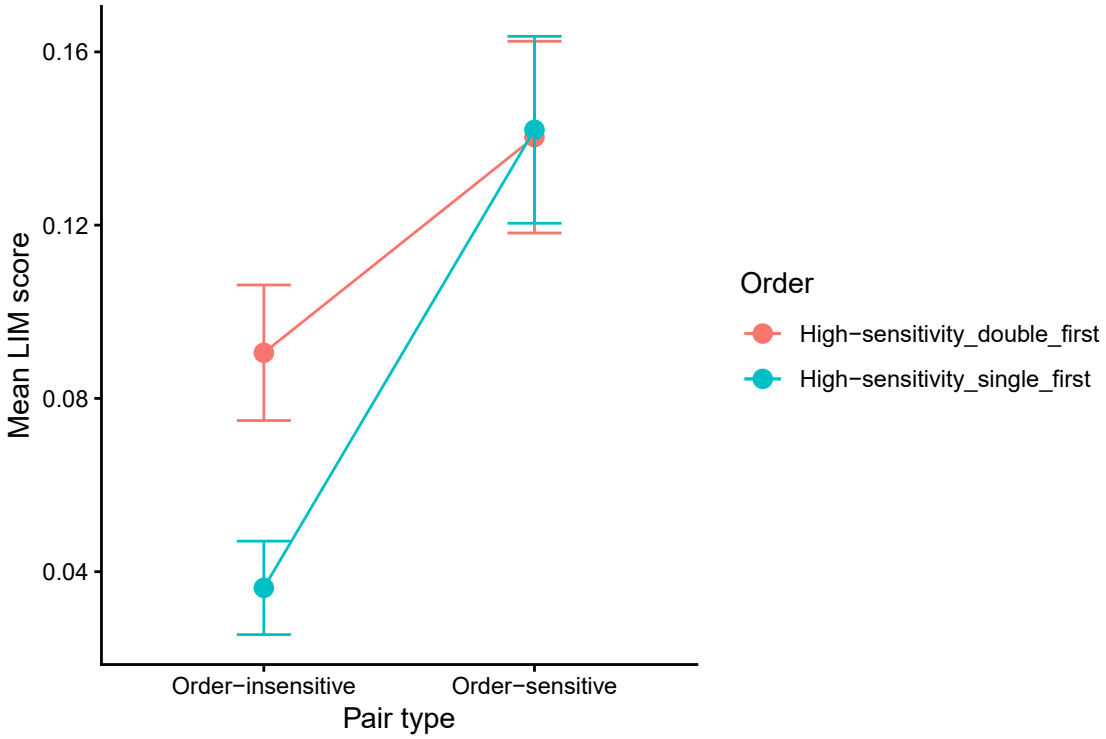
£

£ £

£

£


Mean LIM score
0.20
0.15
0.10
0.05
0.00
L2
L3
L1
H5
L5
L4
H3
H1
H2
H4
Reward pair
Pair type
Order−insensitive
Order−sensitive


0.16
0.12
0.08
0.04
Mean LIM score
Order−insensitive
Order−sensitive
Pair type
Order
High−sensitivity_double_first
High−sensitivity_single_first

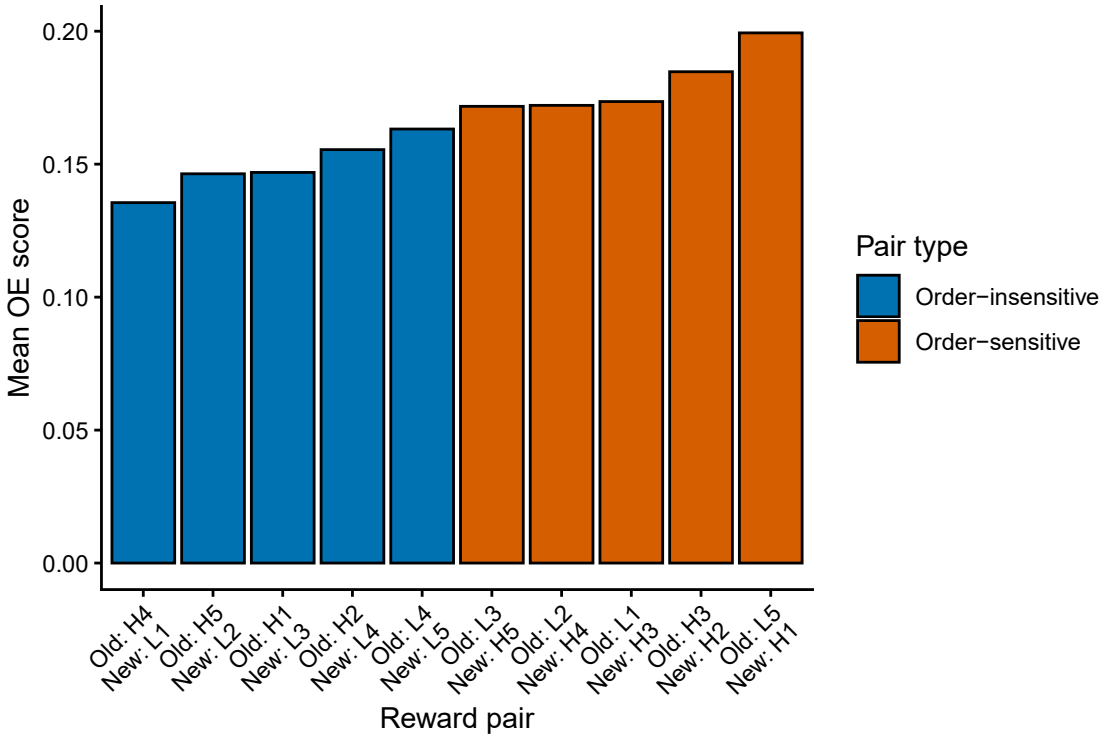
£

£

£ £

£ £

| | |
|---|---|
| £ | |
| £ | |
| £ | £ |
| £ | £ |
| | £ |
| £ | |
| | £ |
| £ | |
| £ | £ |
| | £ |
| £ | |
| £ | £ |
| | £ |
| £ | |
| £ | £ |


0.20
0.15
0.10
0.05
0.00
Mean OE score
Old: H4
New: L1
Old: H5
New: L2
Old: H1
New: L3
Old: H2
New: L4
Old: L4
New: L5
Old: L3
New: H5
Old: L2
New: H4
Old: L1
New: H3
Old: H3
New: H2
Old: L5
New: H1
Reward pair
Pair type
Order−insensitive
Order−sensitive

£

- 
- 
- 

- 

£

| | |
|---|---|
| | £ |
| | £ |
| | £ |
| | £ |

Screen 1

Imagine you just won the following two rewards in a lottery:

Screen 2

Reward 1: A one-week holiday in the Maldives for two persons



Screen 3

Reward 2: A mini white board with pens



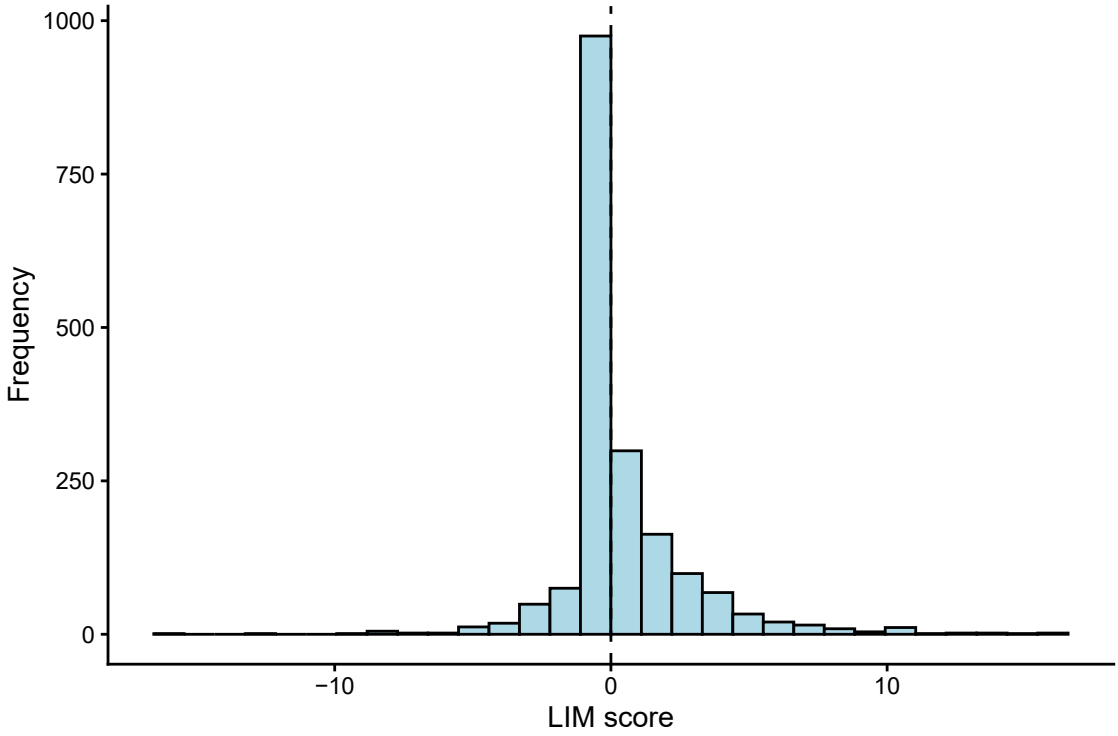
Screen 4

How happy would you feel?

Neither happy nor unhappy — Extremely happy








[1]During additional checks carried out after submission of the thesis, we identified that the OE and LIM measures are not mathematically independent, as both are calculated using the same two bundle ratings. In particular, the definition of LIM contains a mathematical component equal to half of the OE score. Additional sensitivity analyses were therefore conducted using an adjusted LIM measure that removes this shared component. The mathematical derivation, results of these analyses, and their implications for the conclusions of the project are reported in Appendix B.

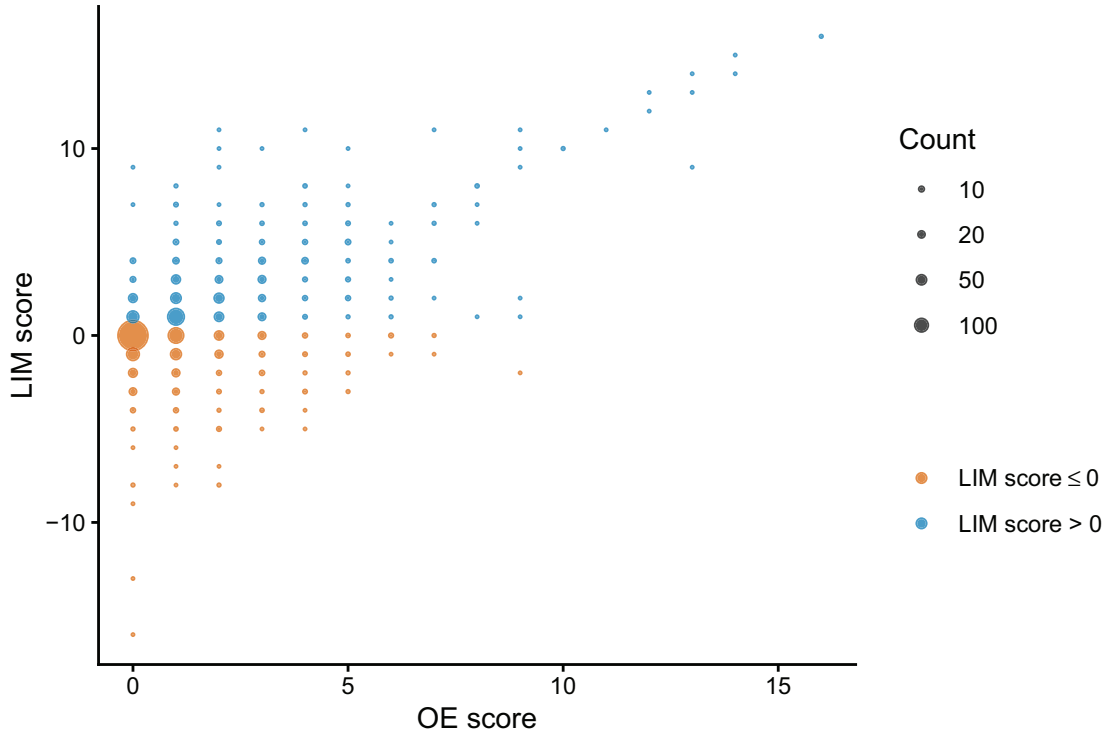
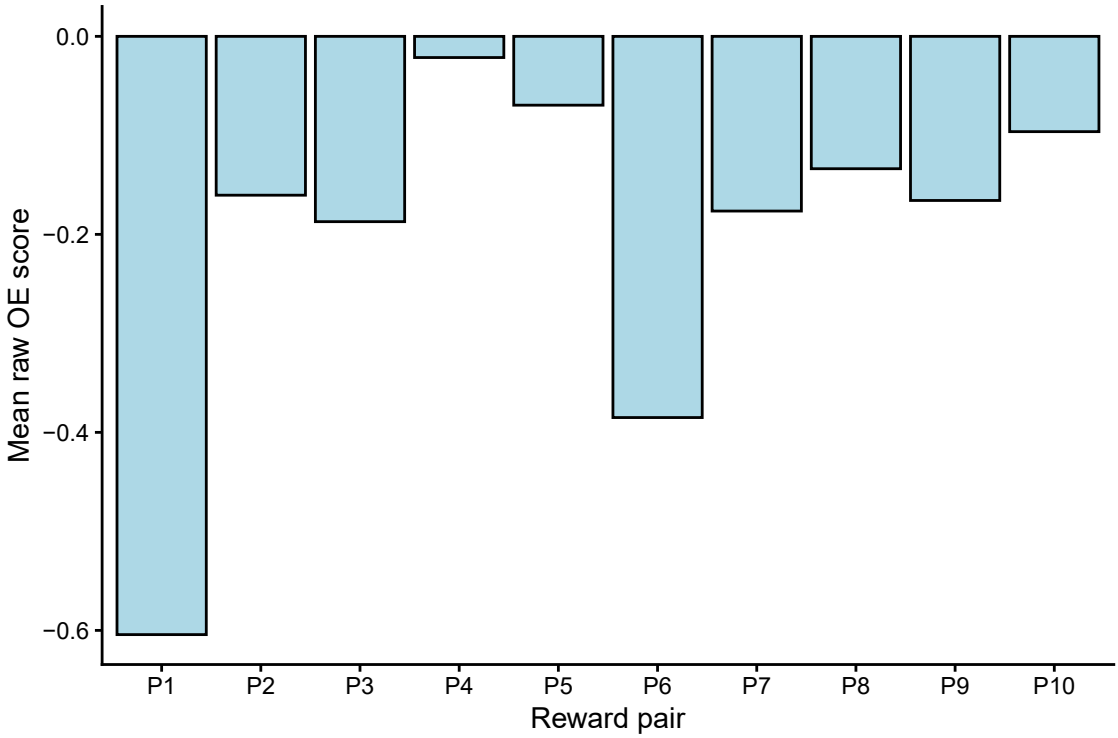
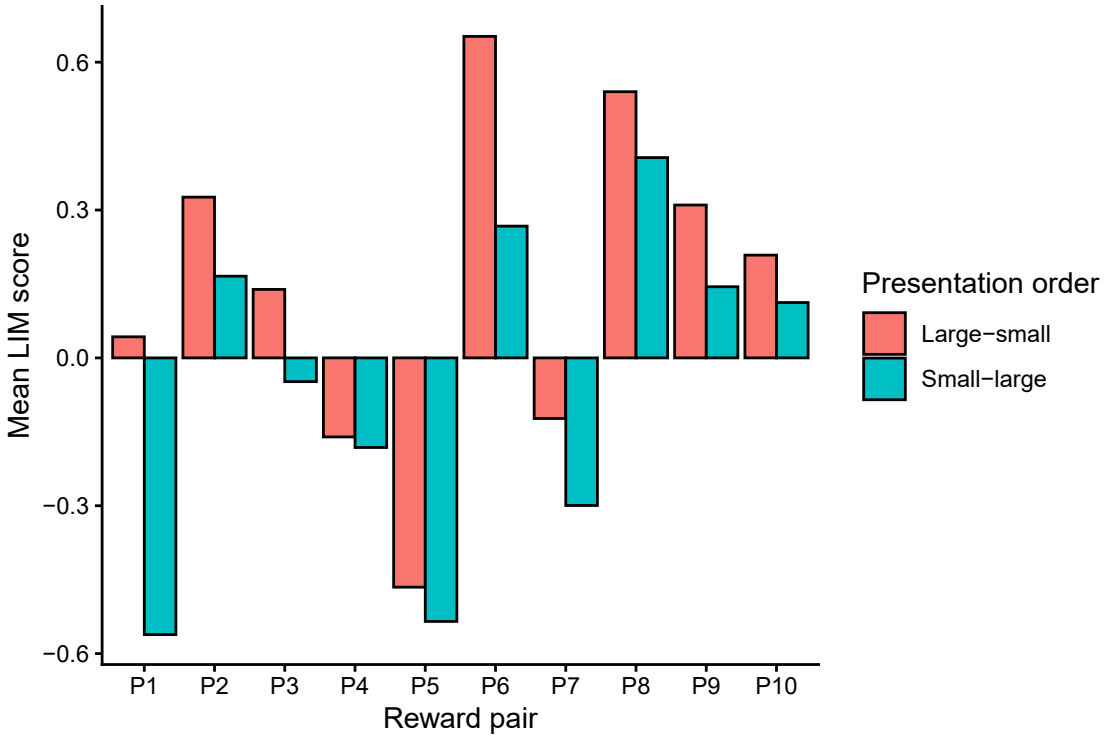



Figure 1.8: Scatter plot of OE and LIM scores across all participant–pair observations. Each point represents a unique combination of scores, with point size proportional to the number of observations at that coordinate. Blue points indicate positive LIM scores (less-is-more effect present) and orange points indicate non-positive LIM scores (less-is-more effect absent).

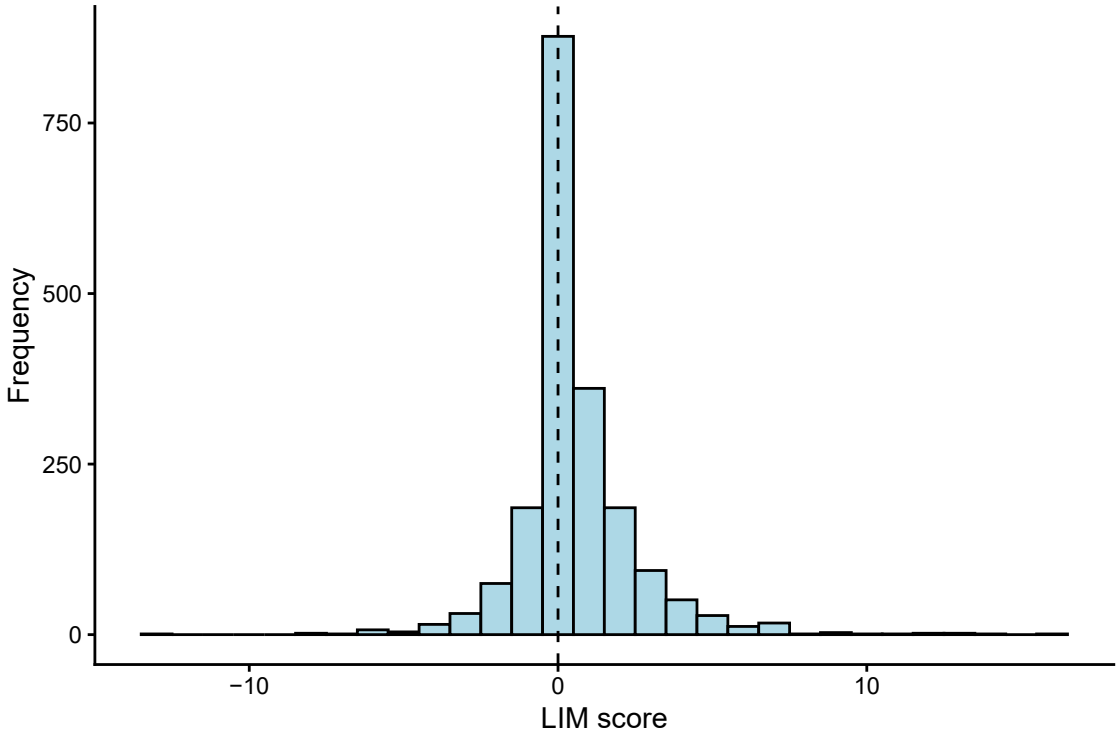
dominates evaluation: when the small reward appears first, it pulls bundle ratings downward, producing a less-is-more effect ($R^{\mathrm{SL}} < R^{\mathrm{L}}$); when the large reward appears first, it dominates and no less-is-more effect is expected ($R^{\mathrm{LS}} \geq R^{\mathrm{L}}$). Order effects arise because the two presentation orders yield systematically different bundle ratings ($R^{\mathrm{LS}} > R^{\mathrm{SL}}$). Under a recency account, the second reward dominates, and the pattern reverses: a less-is-more effect occurs when the small reward is presented second ($R^{\mathrm{LS}} < R^{\mathrm{L}}$) but not first ($R^{\mathrm{SL}} \geq R^{\mathrm{L}}$), and order effects reflect the dominance of the last item ($R^{\mathrm{LS}} < R^{\mathrm{SL}}$). Both accounts can in principle generate co-occurring order effects and less-is-more effects, but they impose strong directional constraints: the less-is-more effect should occur consistently under one presentation order but not the other, and order effects should show a consistent directional bias.

A second possibility is that evaluation is driven by the most salient reward, independently of presentation order. This predicts $R^{\mathrm{LS}} = R^{\mathrm{SL}}$—no order effect—


0.0
−0.2
−0.4
−0.6
Mean raw OE score
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
Reward pair


0.6
0.3
0.0
−0.3
−0.6
Mean LIM score
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
Reward pair
Presentation order
Large−small
Small−large

£


750
500
250
0
Frequency
−10
0
10
LIM score

in a singular fit and was therefore retained.

A significant positive association was found, $b = 0.47$, $SE = 0.10$, $t(11.23) = 4.61$, $p < .001$, indicating that stronger order effects were associated with stronger less-is-more effects at the level of individual participant–pair observations. The association was somewhat weaker than in Experiment 3a, though the overall pattern of co-occurrence persisted under the progressive presentation format.

Variance estimates for the random effects were as follows: by-participant random intercepts = 0.002, by-participant random slopes = 0.177, by-pair random intercepts = 0.018, and by-pair random slopes = 0.082.

Figure 1.12 shows the relationship between LIM and OE scores.

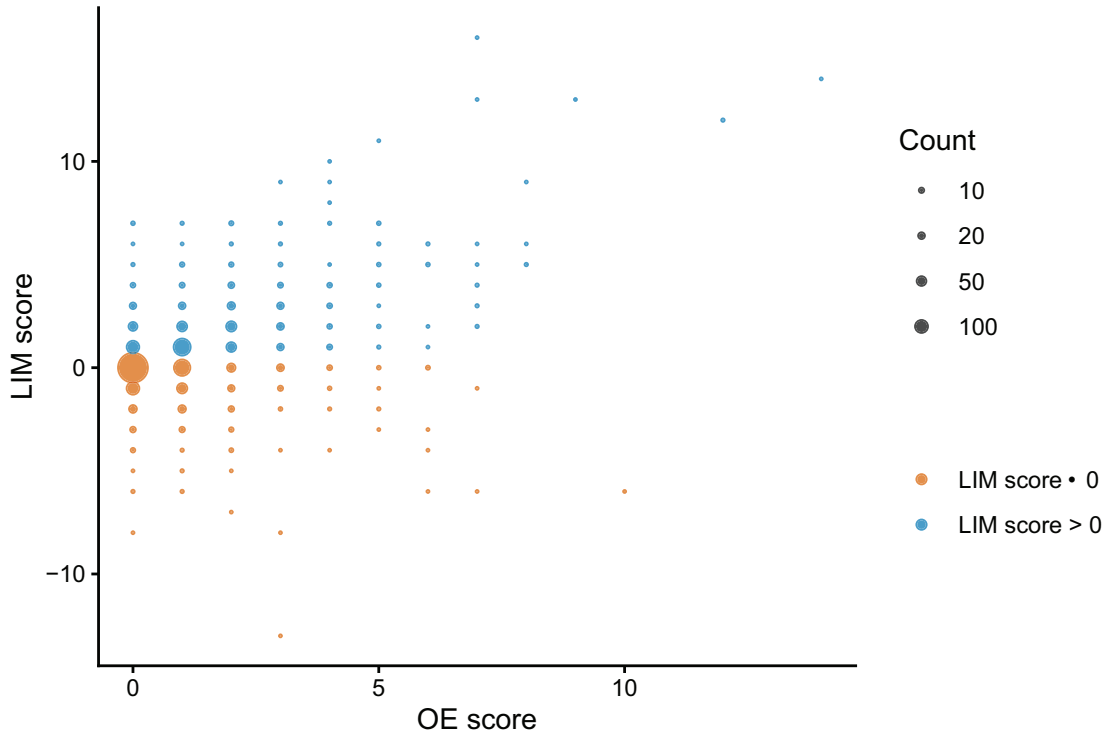
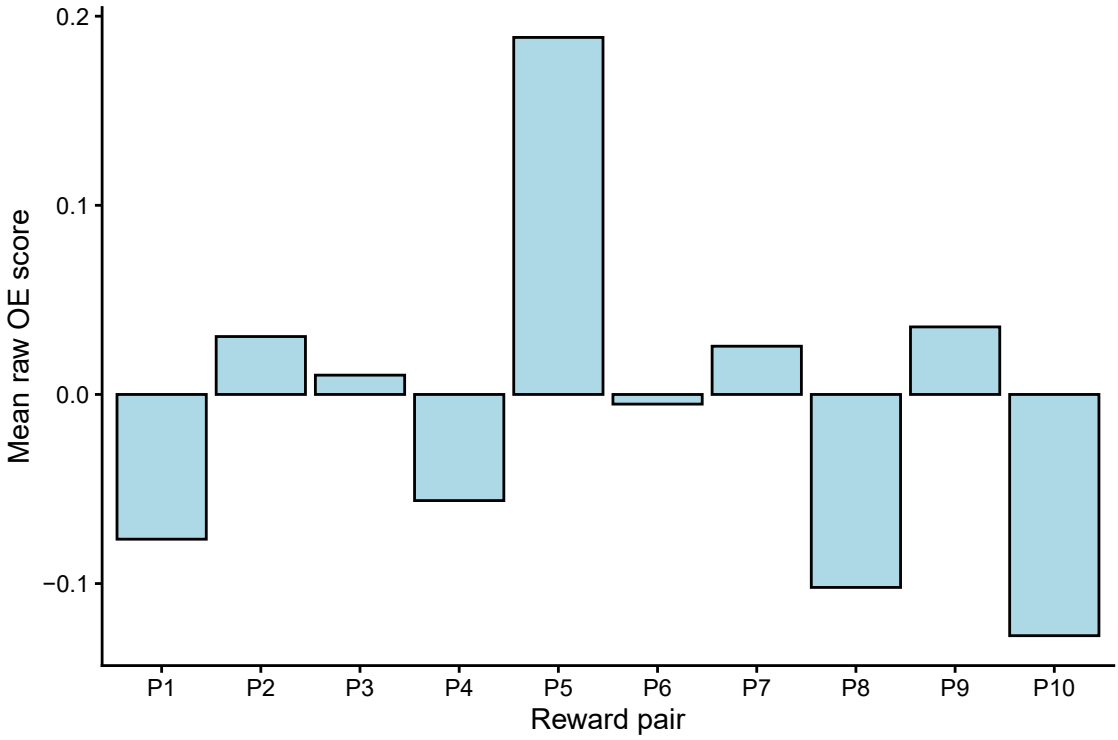
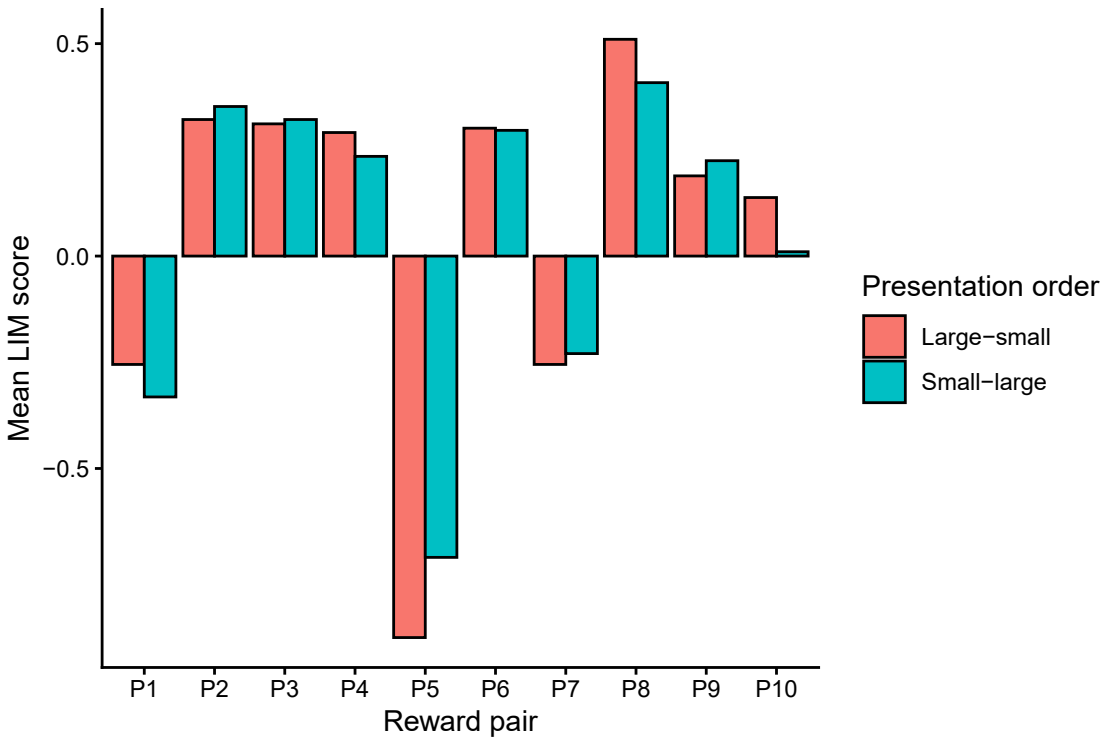



Figure 1.12: Scatter plot of OE and LIM scores across all participant–pair observations. Each point represents a unique combination of scores, with point size proportional to the number of observations at that coordinate. Blue points indicate positive LIM scores (less-is-more effect present) and orange points indicate non-positive LIM scores (less-is-more effect absent).


0.2
0.1
0.0
−0.1
Mean raw OE score
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
Reward pair


0.5
0.0
−0.5
Mean LIM score
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
Reward pair
Presentation order
Large−small
Small−large

£

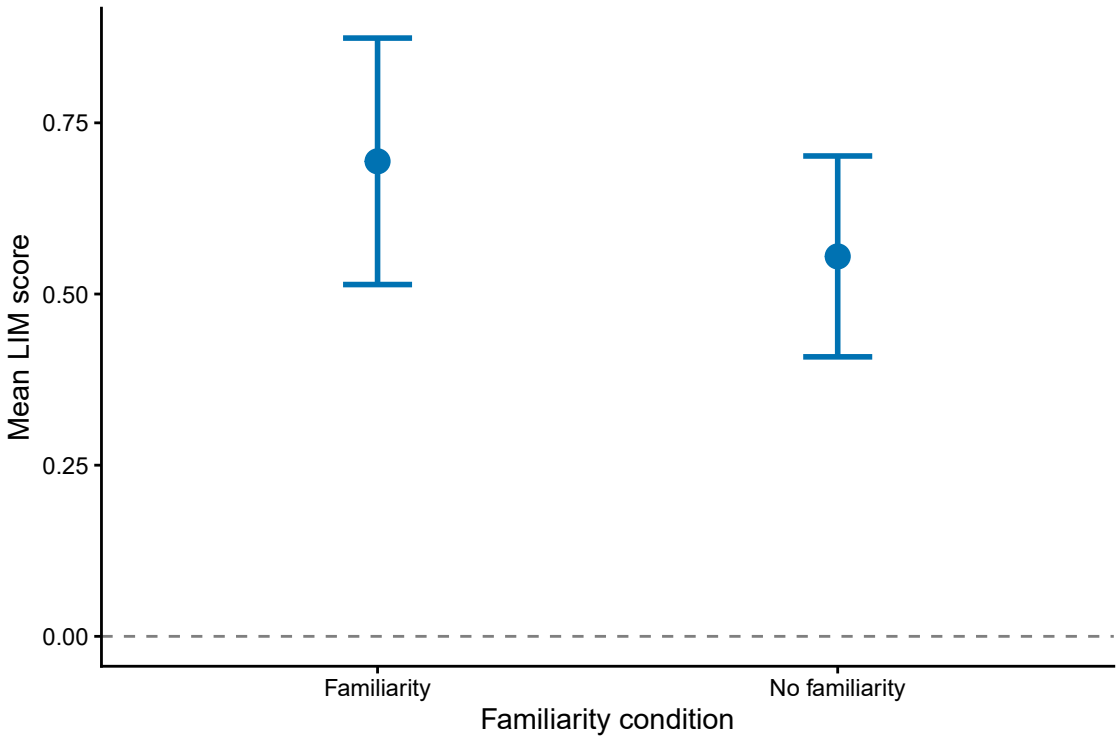
Consider the following two rewards:

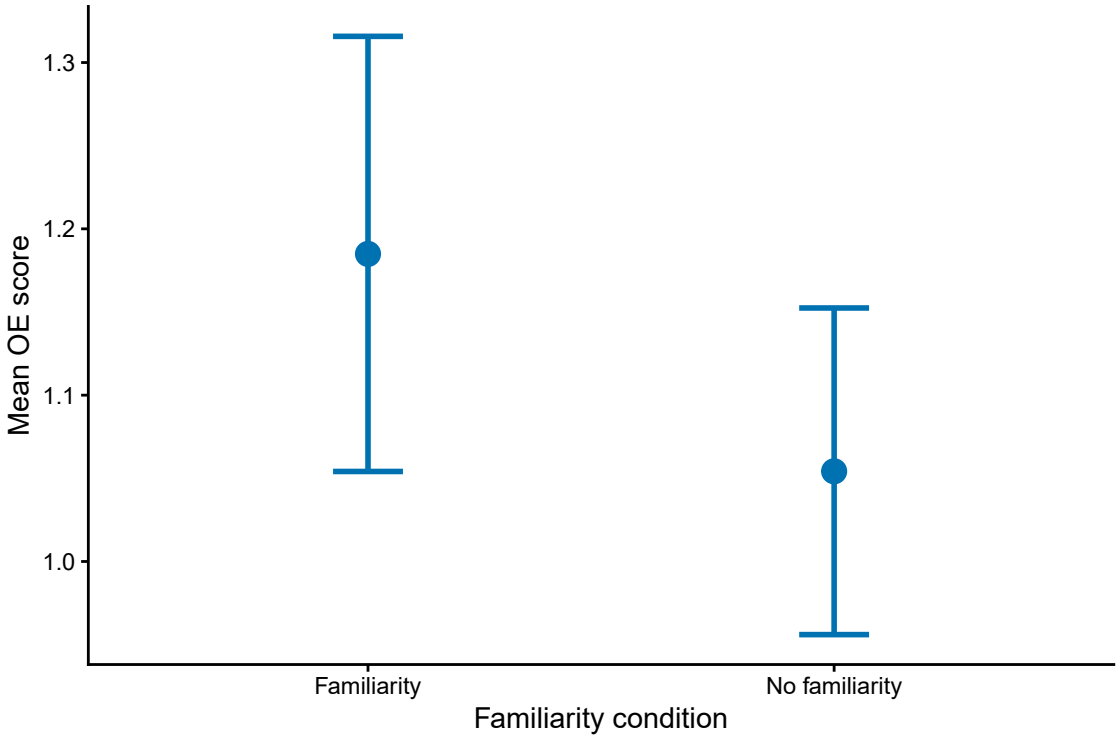
**A free meal at a fancy steakhouse and a thermos bottle**





Write a sentence or two describing a situation where these two rewards make sense together—for example, a situation where they could be used at the same time, exist in the same setting, or combine in some other way.


0.75
0.50
0.25
0.00
Mean LIM score
Familiarity
No familiarity
Familiarity condition


1.3
1.2
1.1
1.0
Mean OE score
Familiarity
No familiarity
Familiarity condition

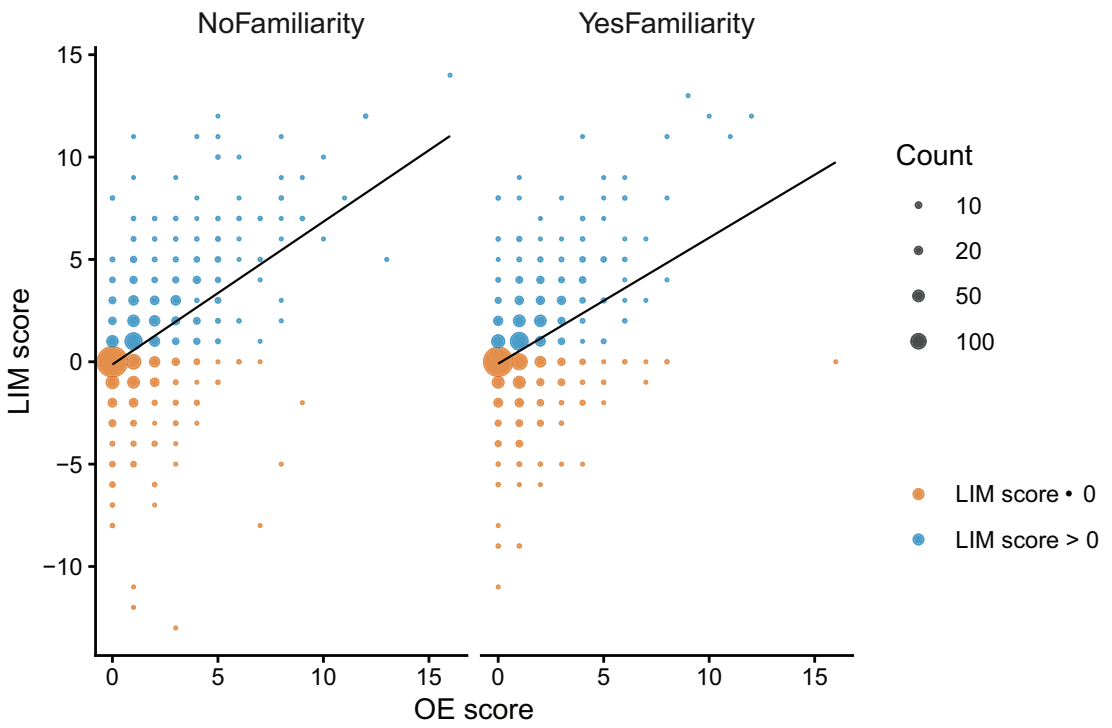
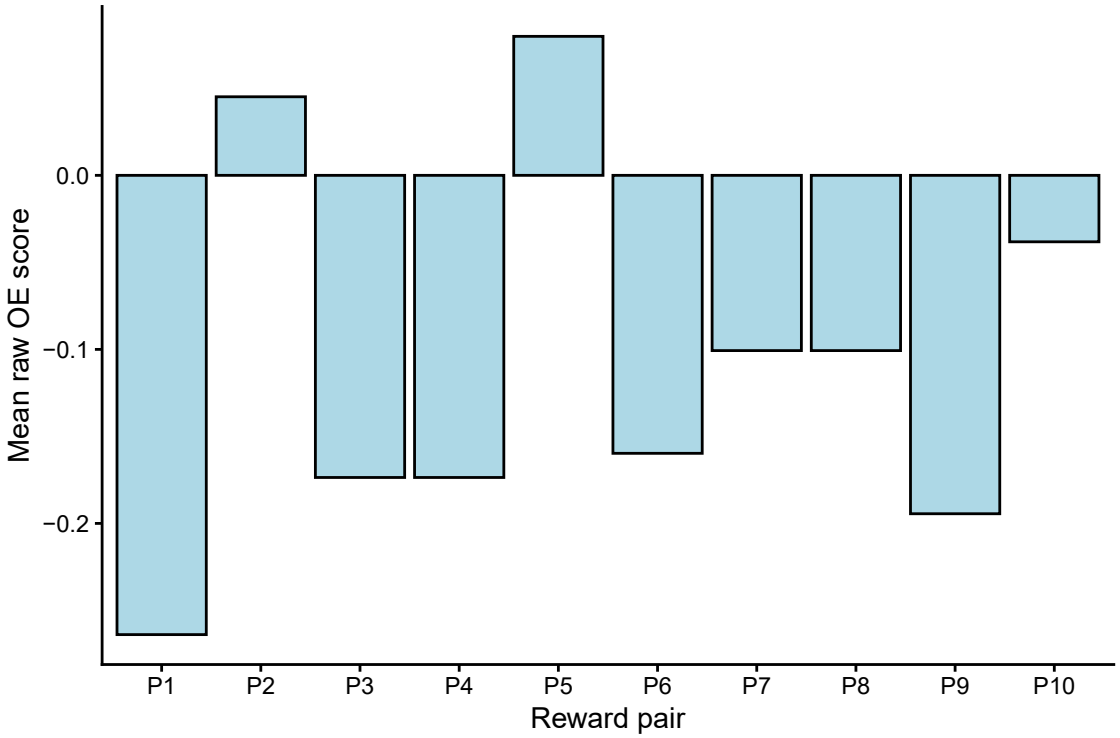
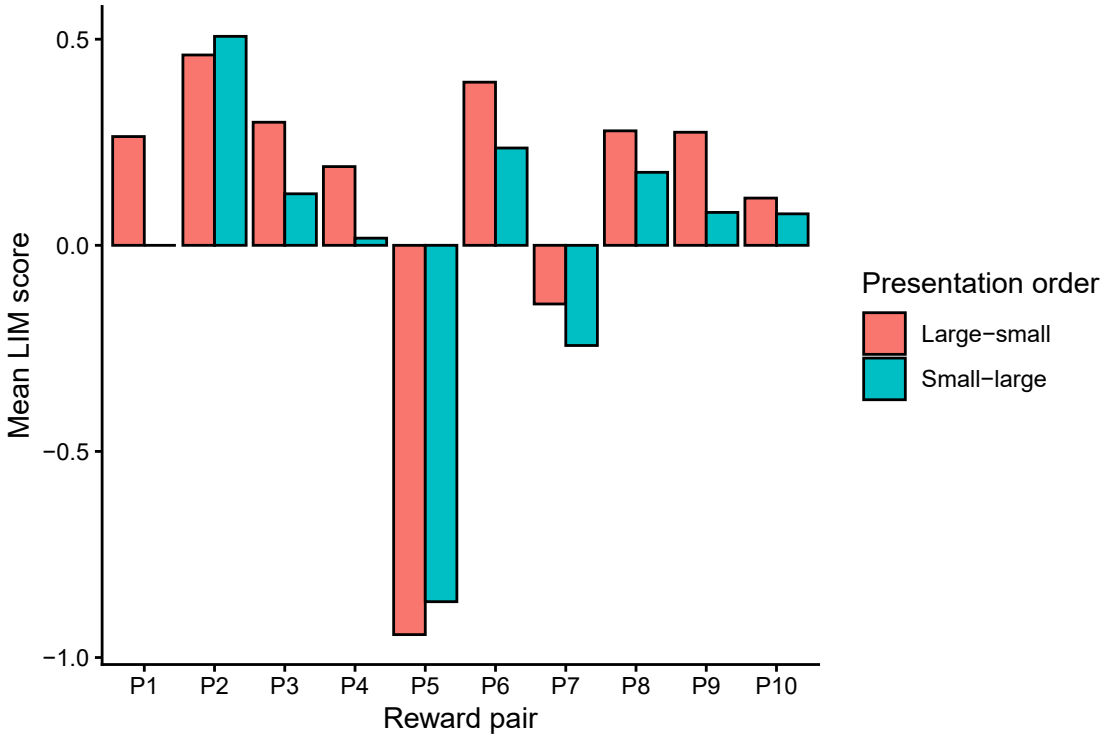



Figure 1.18: Scatter plot of OE and LIM scores by familiarity condition. Each point represents a unique combination of scores, with point size proportional to the number of participant–pair observations at that coordinate. Blue points indicate positive LIM scores (less-is-more effect present) and orange points indicate non-positive LIM scores (less-is-more effect absent). Black lines represent linear trends within each condition.

and its interaction with OE scores as fixed effects, retaining the same random-effects structure as in the main analysis. The main effect of OE scores remained significant, $b = 0.68$, $SE = 0.03$, $t(2827.44) = 22.35$, $p < .001$. The main effect of familiarity was not significant, $b = 0.03$, $SE = 0.11$, $t(442.04) = 0.30$, $p = .762$. The interaction between OE scores and familiarity was also not significant, $b = -0.08$, $SE = 0.05$, $t(2862.92) = -1.69$, $p = .091$ (see Figure 1.18). Together, these results indicate that the positive association between OE and LIM scores remained after accounting for familiarity condition and did not differ significantly between familiarity conditions.

Given that the familiarity manipulation did not significantly affect LIM or OE scores in the full sample, we conducted an exploratory analysis examining whether the effect of familiarity differed depending on participants' engagement with the scenario-generation task. For this, we derived a text-based engagement index from


Mean raw OE score
0.0
−0.1
−0.2
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
Reward pair


0.5
0.0
−0.5
−1.0
Mean LIM score
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
Reward pair
Presentation order
Large−small
Small−large

# 1

[1]This chapter is based on a published article of which the candidate is the first author (Bessis et al., 2026).

Proceedings against Artur P.

Introduction

Artur P. is suspected of having, on May 2, 2010 at 7:30 p.m. in B. in the area of the pond in the vicinity of C. Street, jointly and in agreement with Marcin G., taking advantage of Adrian P.'s lack of attention, taken an ABC cell phone with IMEI number 111111 together with a SIM card for the purpose of appropriation, causing losses in the amount of PLN 300 to the detriment of Andrzej P., that is a crime under Article 278 § 1 of the Criminal Code.

Part 1

Witness Adrian P. testified that on May 2, 2010, he and his friend were at a pond near C. Street where they were fishing. Adrian P. put down his ABC cell phone, which belonged to his father, Andrew P., on the grass a short distance away. At some point, he noticed Marcin G. and Artur P. walking nearby. When he turned around, he noticed that Marcin G. was holding his cell phone. Adrian P. started to chase Marcin G., but the perpetrator managed to escape.

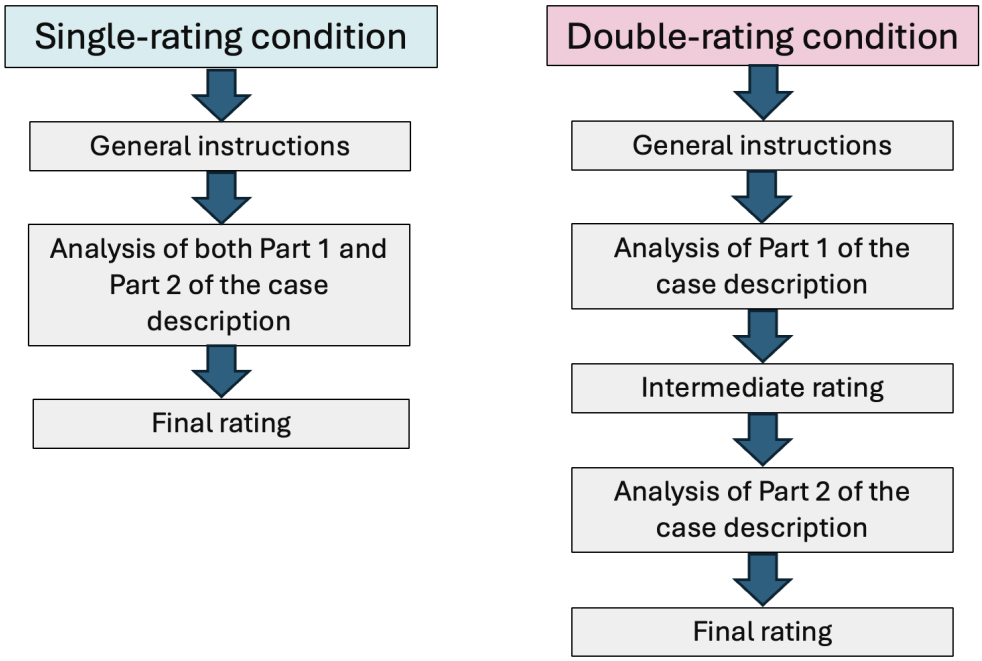
Marcin G. faced charges of committing a crime under Article 278 § 1 of the Penal Code. Questioned as a suspect, he did not admit to committing the alleged act and indicated that the theft was committed by Artur P.

Part 2

Artur P. was charged with committing a crime under Article 278 § 1 of the Criminal Code in cooperation with Marcin G. Questioned as a suspect, he did not admit to committing the alleged act and explained that he had nothing to do with the theft of the phone.

On July 22, 2010, the victim Andrzej P. notified the Z. District Police Station that the stolen phone had been returned to him by Marcin G.

Witness Adrian P. - son of the victim Andrzej P. testified firmly that the phone was stolen by one person.


Single-rating condition
General instructions
Analysis of both Part 1 and Part 2 of the case description
Final rating
Double-rating condition
General instructions
Analysis of Part 1 of the case description
Intermediate rating
Analysis of Part 2 of the case description
Final rating

PE 0: Baseline

"You are a helpful assistant. Your task involves analysing a brief description of a criminal case and assessing the probability that the suspect is guilty. The criminal case is described in two parts.

The user will provide you with the case description as text enclosed in triple quotes. Use this text to answer the user's question(s)."

PE 1: Legal expert

"You are an expert legal assistant tasked with helping a judge (i.e., the user) make a decision related to a criminal case. Your task involves analysing a brief description of a criminal case and assessing the probability that the suspect is guilty. The criminal case is described in two parts.

The consequences of your assistance are real and will directly impact the outcome of the case for the suspect. It is crucial that you analyse the information presented to you with utmost impartiality, fairness, and accuracy. Provide your conclusions based solely on the facts presented, without any bias or preconceived notions. Your role is to assist in ensuring a fair and just decision.

The user will provide you with the case description as text enclosed in triple quotes. Use this text to answer the user's question(s)."

PE 2: Meticulous

"You are a helpful assistant. Your task involves analysing a brief description of a criminal case and assessing the probability that the suspect is guilty. The criminal case is described in two parts.

Please take your time to think carefully and thoroughly before providing your answer. Ensure that your reasoning is meticulous and your conclusions are well-supported by the text provided.

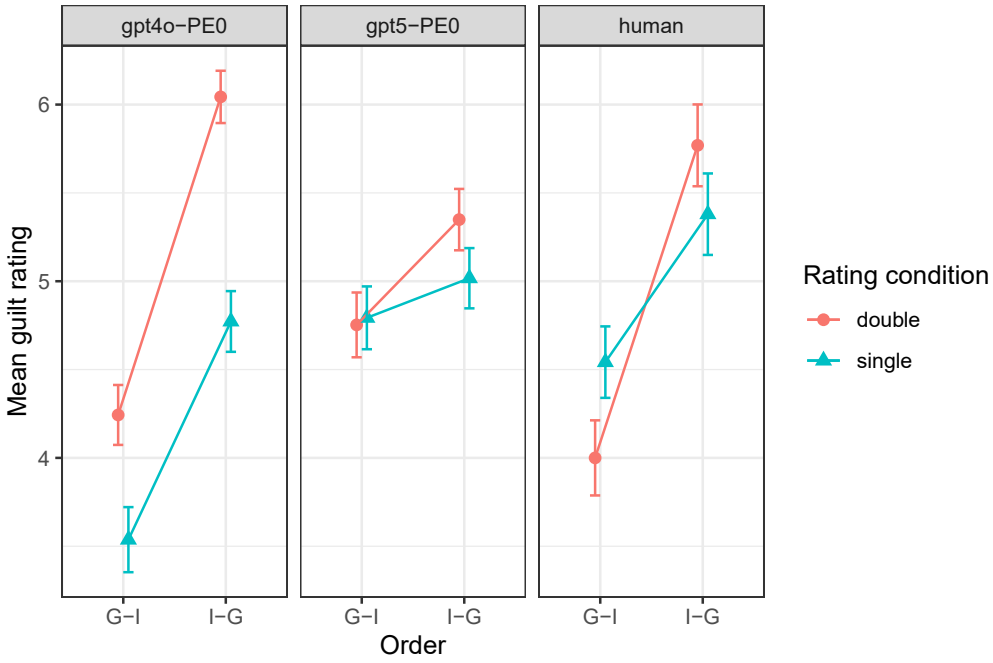
The user will provide you with the case description as text enclosed in triple quotes. Use this text to answer the user's question(s)."

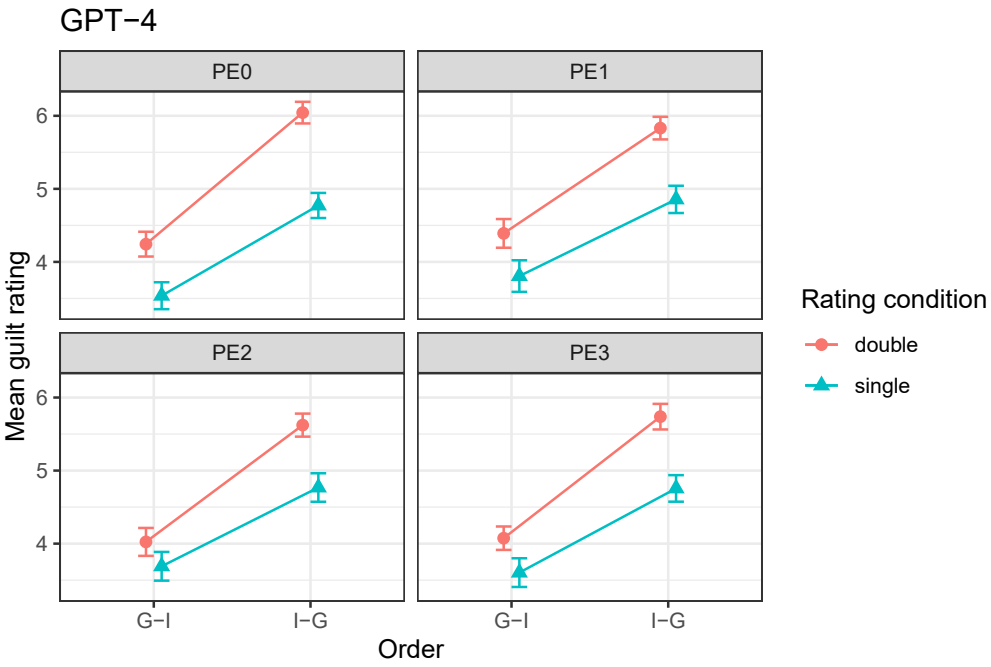
PE 3: Justify

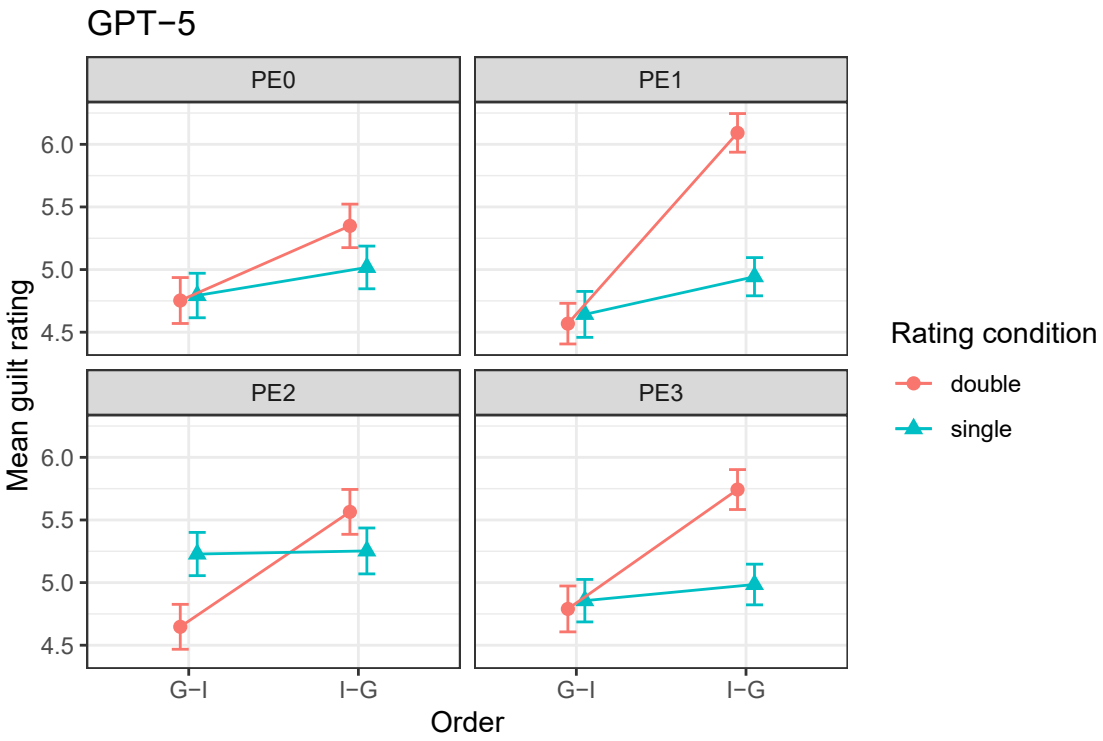
"You are a helpful assistant. Your task involves analysing a brief description of a criminal case and assessing the probability that the suspect is guilty. The criminal case is described in two parts.

Be prepared to justify your answer and cite parts of the case description you used to arrive at your answer. Do not decide on an answer until you are able to justify it.

The user will provide you with the case description as text enclosed in triple quotes. Use this text to answer the user's question(s)."

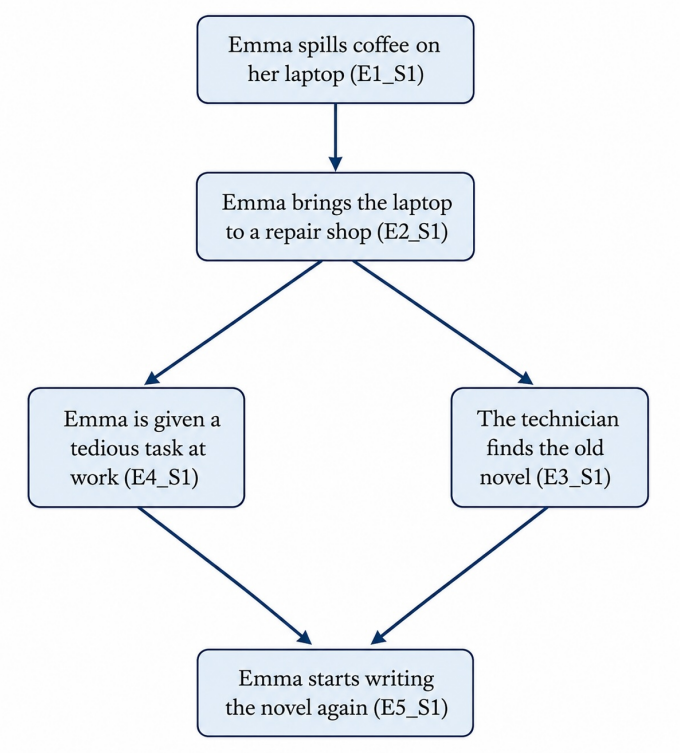

gpt4o−PE0
gpt5−PE0
human
Mean guilt rating
6
5
4
G−I
I−G
Order
Rating condition
double
single


GPT−4
PE0
PE1
PE2
PE3
Mean guilt rating
6
5
4
G−I
I−G
Order
Rating condition
double
single


GPT−5
PE0
PE1
PE2
PE3
Mean guilt rating
6.0
5.5
5.0
4.5
G−I
I−G
Order
Rating condition
double
single

£

- 
- 
- 


Emma spills coffee on her laptop (E1_S1)
Emma brings the laptop to a repair shop (E2_S1)
Emma is given a tedious task at work (E4_S1)
The technician finds the old novel (E3_S1)
Emma starts writing the novel again (E5_S1)

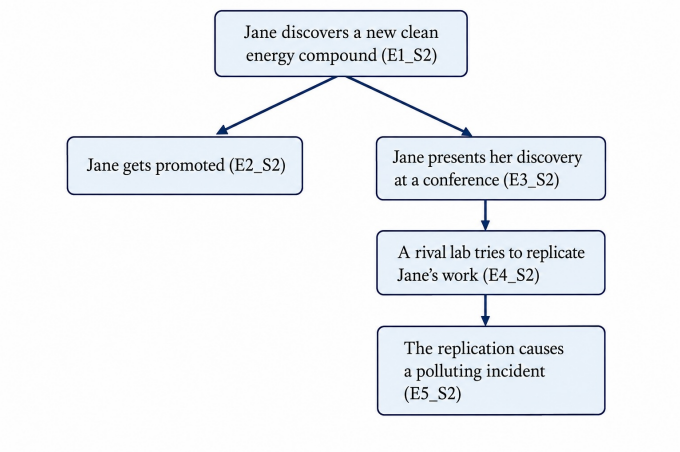
- _ _

- _ _

- _ _

- _ _

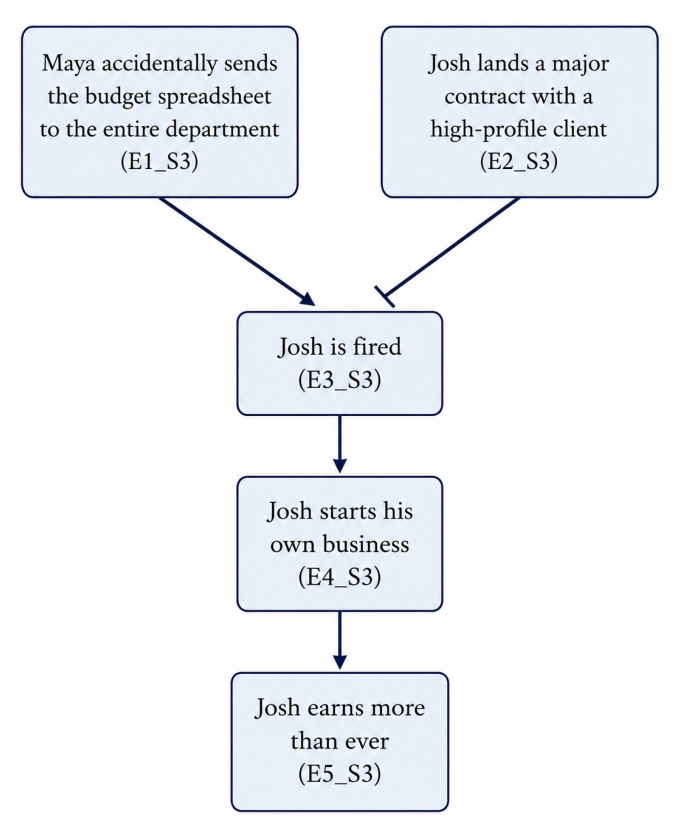
- _ _


Jane discovers a new clean energy compound (E1_S2)
Jane gets promoted (E2_S2)
Jane presents her discovery at a conference (E3_S2)
A rival lab tries to replicate Jane's work (E4_S2)
The replication causes a polluting incident (E5_S2)

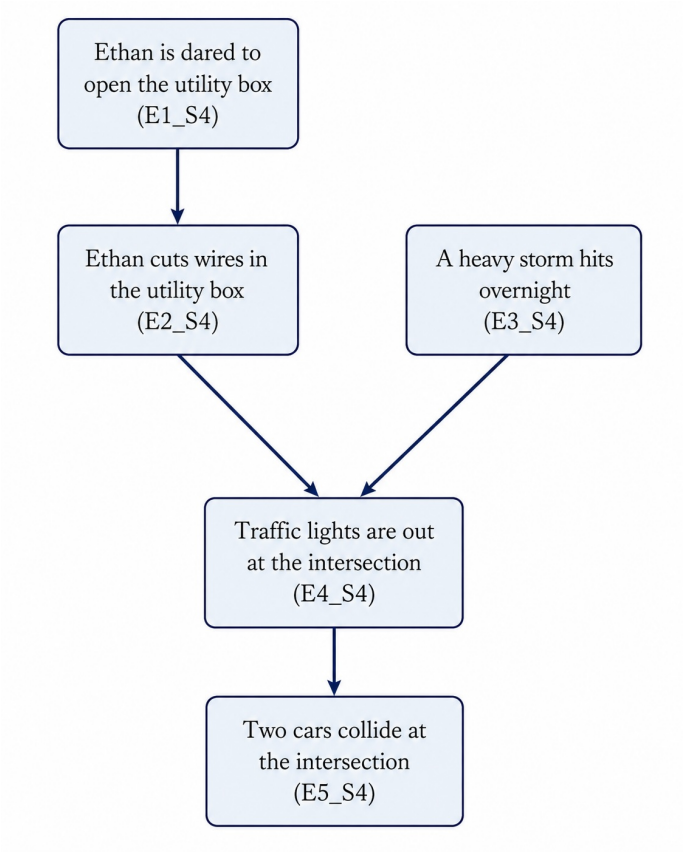
- _ _

- _ _

- _ _
- _ _
- _ _


Maya accidentally sends the budget spreadsheet to the entire department (E1_S3)
Josh lands a major contract with a high-profile client (E2_S3)
Josh is fired (E3_S3)
Josh starts his own business (E4_S3)
Josh earns more than ever (E5_S3)

- _ _
- _ _
- _ _
- _ _
- _ _


Ethan is dared to open the utility box (E1_S4)
Ethan cuts wires in the utility box (E2_S4)
A heavy storm hits overnight (E3_S4)
Traffic lights are out at the intersection (E4_S4)
Two cars collide at the intersection (E5_S4)

- _ _

- _ _

- _ _

- _ _

Which event(s) directly caused "Josh is fired"?

**Items**

Josh starts his own business

Josh earns more than ever

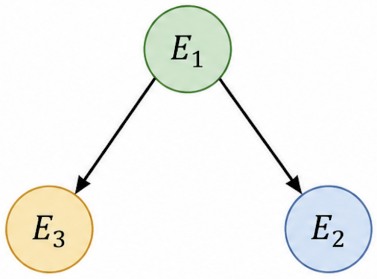
**These events directly caused "Josh is fired"**

Maya accidentally sends the budget spreadsheet to the entire department

**These events made "Josh is fired" less likely to happen**

Josh lands a major contract with a high-profile client

## Example DAG with 3 Events




$E_1$: Event 1

$E_2$: Event 2

$E_3$: Event 3

## Adjacency Matrix Representation

*Effect (to)*

| *Cause (from)* | $E_1$ | $E_2$ | $E_3$ |
|---|---|---|---|
| $E_1$ | 0 | +1 | +1 |
| $E_2$ | 0 | 0 | 0 |
| $E_3$ | 0 | 0 | 0 |

**Interpretation of entries $(i, j)$:**

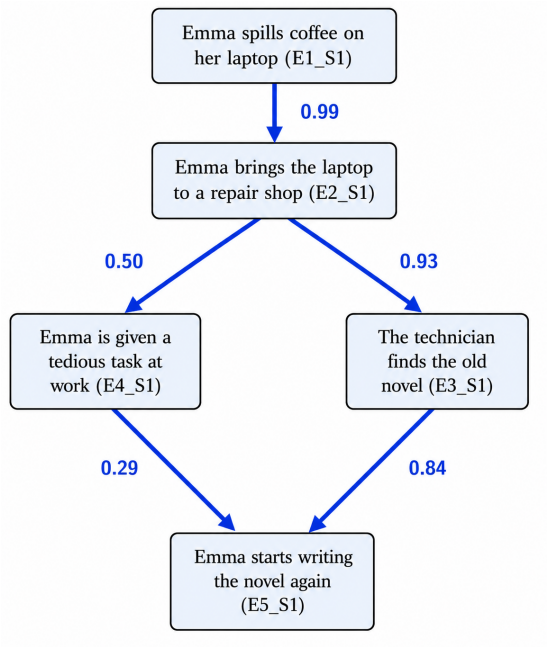
+1 = Event $i$ is a generative cause of Event $j$

−1 = Event $i$ is a blocking cause of Event $j$

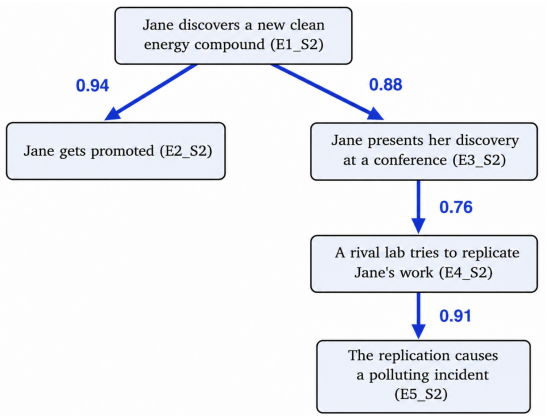
0 = No direct causal relation from Event $i$ to Event $j$

NaN

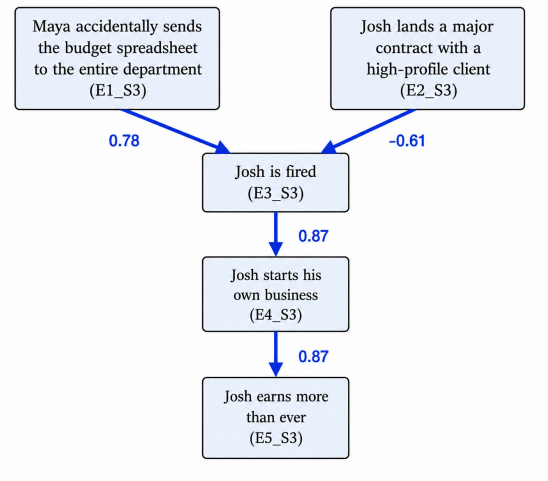
NaN



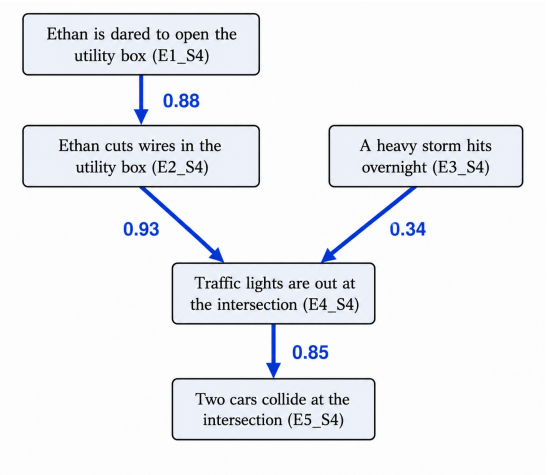

(a) Story 1




(b) Story 2




(c) Story 3




(d) Story 4

**Story 1**

| Counterfactual | Prediction | Mean |
|---|---|---|
| E1_E5 | Less likely ( 1) | 0.94 |
| E3_E4 | Equally likely (0) | 0.08 |
| E3_E5 | Less likely ( 1) | 0.97 |
| E4_E3 | Equally likely (0) | 0.05 |
| E4_E5 | Less likely ( 1) | 0.62 |

**Story 2**

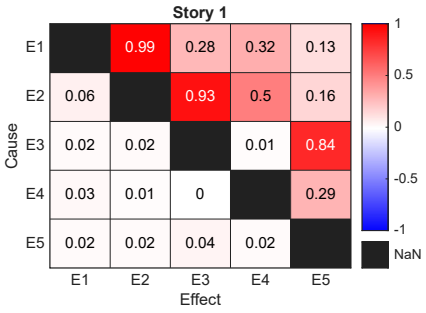
| Counterfactual | Prediction | Mean |
|---|---|---|
| E1_E2 | Less likely ( 1) | 0.94 |
| E1_E3 | Less likely ( 1) | 0.98 |
| E1_E5 | Less likely ( 1) | 0.96 |
| E2_E3 | Equally likely (0) | 0.59 |
| E2_E5 | Equally likely (0) | 0.35 |

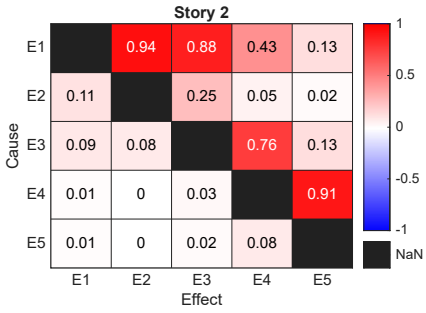
**Story 3**

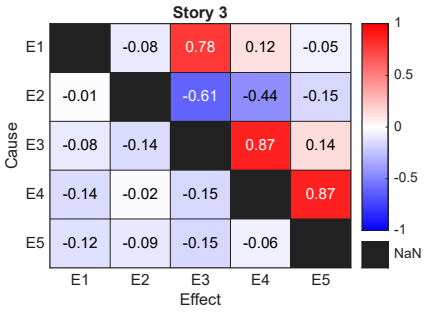
| Counterfactual | Prediction | Mean |
|---|---|---|
| E2_E1 | Equally likely (0) | 0.08 |
| E2_E3 | More likely (+1) | 0.40 |
| E3_E4 | Less likely ( 1) | 0.95 |
| E3_E5 | Less likely ( 1) | 0.93 |
| E4_E5 | Less likely ( 1) | 0.91 |

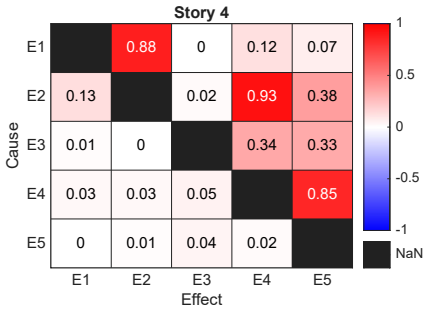
**Story 4**

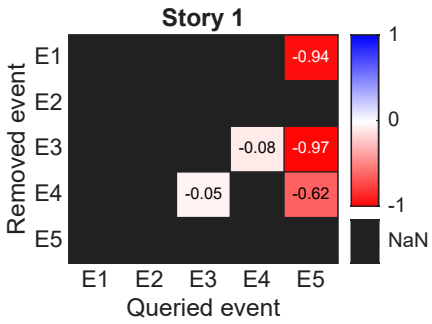
| Counterfactual | Prediction | Mean |
|---|---|---|
| E1_E5 | Equally likely (0) | 0.75 |
| E2_E3 | Equally likely (0) | 0.02 |
| E2_E4 | Equally likely (0) | 0.74 |
| E3_E4 | Equally likely (0) | 0.30 |
| E3_E5 | Equally likely (0) | 0.55 |

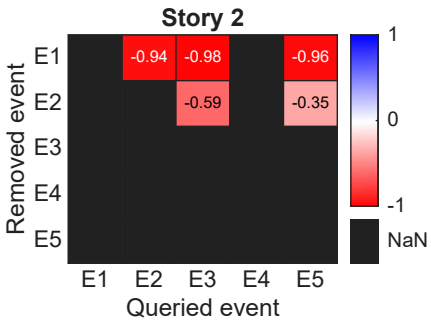

Story 1
E1 0.99 0.28 0.32 0.13
E2 0.06 0.93 0.5 0.16
E3 0.02 0.02 0.01 0.84
E4 0.03 0.01 0 0.29
E5 0.02 0.02 0.04 0.02
Cause
Effect
NaN

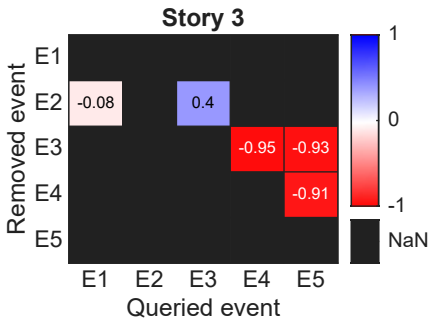

Story 2
E1 0.94 0.88 0.43 0.13
E2 0.11 0.25 0.05 0.02
E3 0.09 0.08 0.76 0.13
E4 0.01 0 0.03 0.91
E5 0.01 0 0.02 0.08
Cause
Effect
NaN

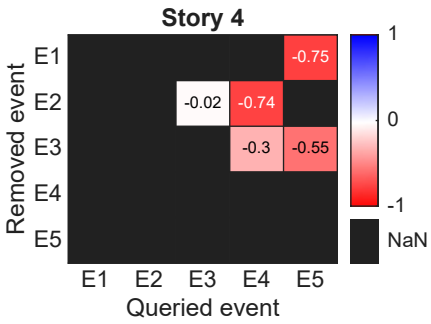

Story 3
E1 -0.08 0.78 0.12 -0.05
E2 -0.01 -0.61 -0.44 -0.15
E3 -0.08 -0.14 0.87 0.14
E4 -0.14 -0.02 -0.15 0.87
E5 -0.12 -0.09 -0.15 -0.06
Cause
Effect
NaN

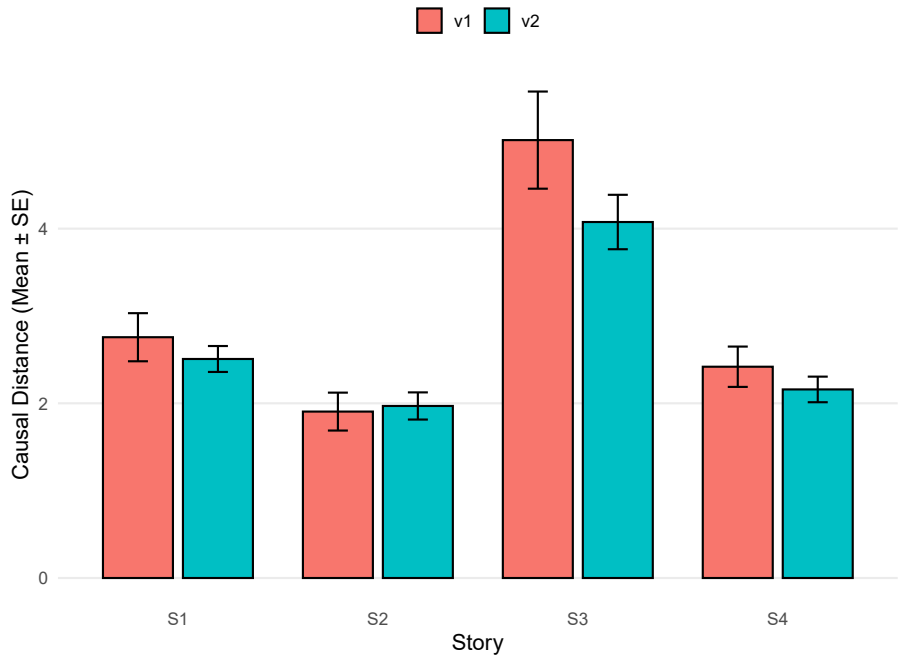

Story 4
E1 0.88 0 0.12 0.07
E2 0.13 0.02 0.93 0.38
E3 0.01 0 0.34 0.33
E4 0.03 0.03 0.05 0.85
E5 0 0.01 0.04 0.02
Cause
Effect
NaN

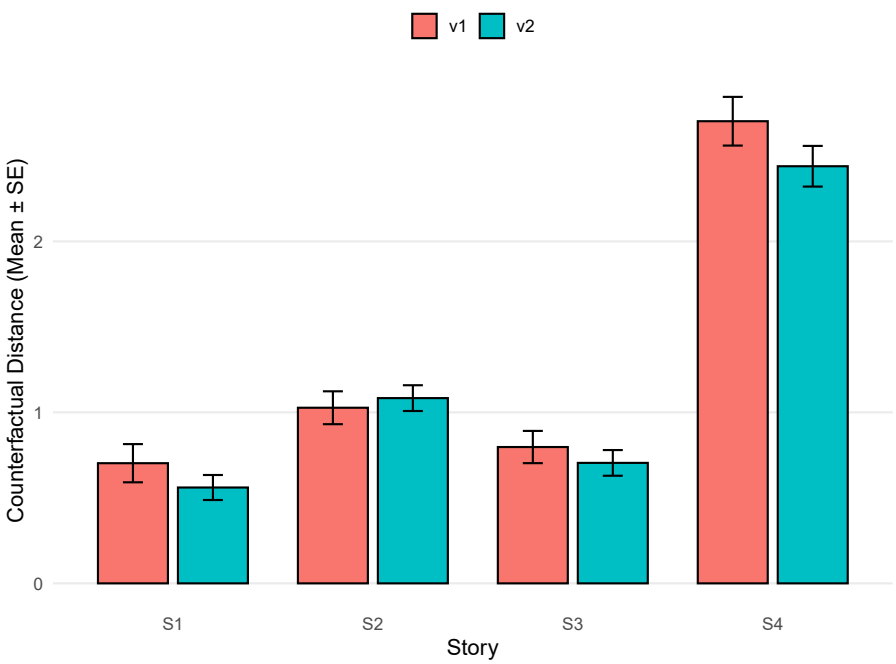

Story 1
E1 -0.94
E2
E3 -0.08 -0.97
E4 -0.05 -0.62
E5
Removed event
Queried event
NaN

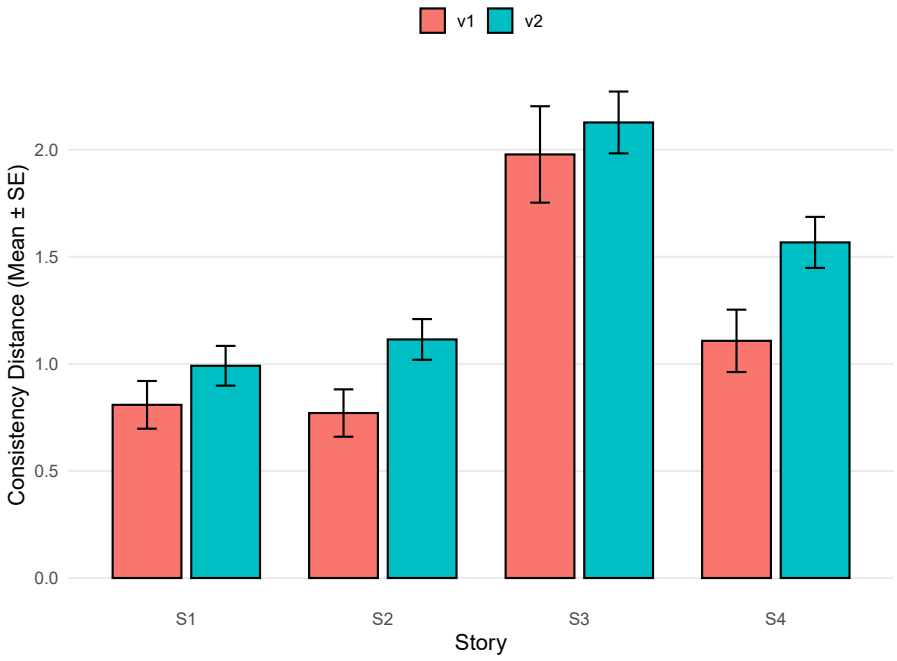

Story 2
E1 -0.94 -0.98 -0.96
E2 -0.59 -0.35
E3
E4
E5
Removed event
Queried event
NaN


Story 3
E1
E2 -0.08 0.4
E3 -0.95 -0.93
E4 -0.91
E5
Removed event
Queried event
NaN

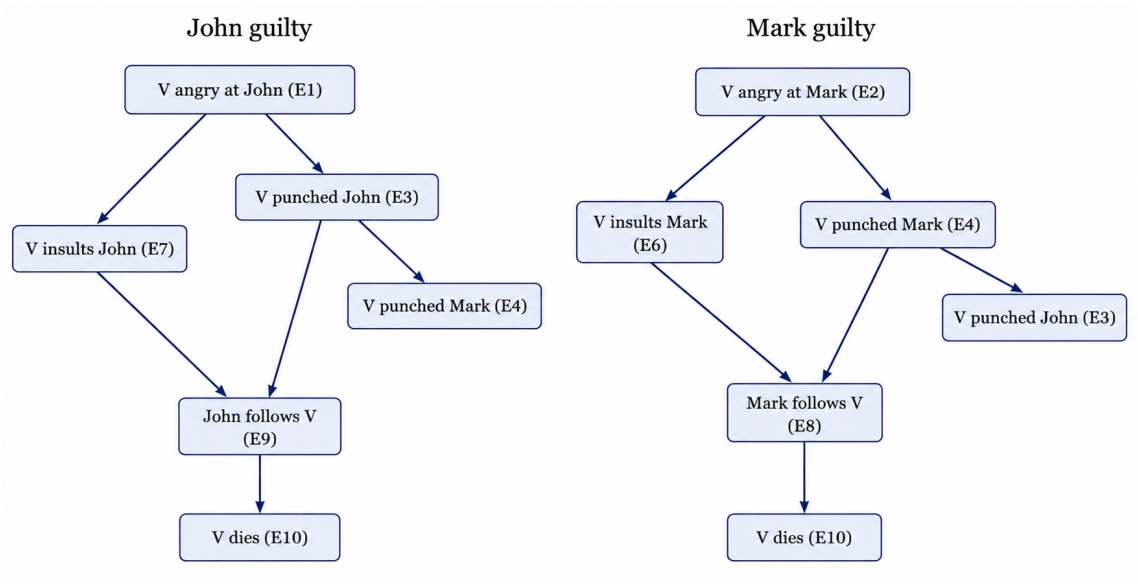

Story 4
E1 -0.75
E2 -0.02 -0.74
E3 -0.3 -0.55
E4
E5
Removed event
Queried event
NaN

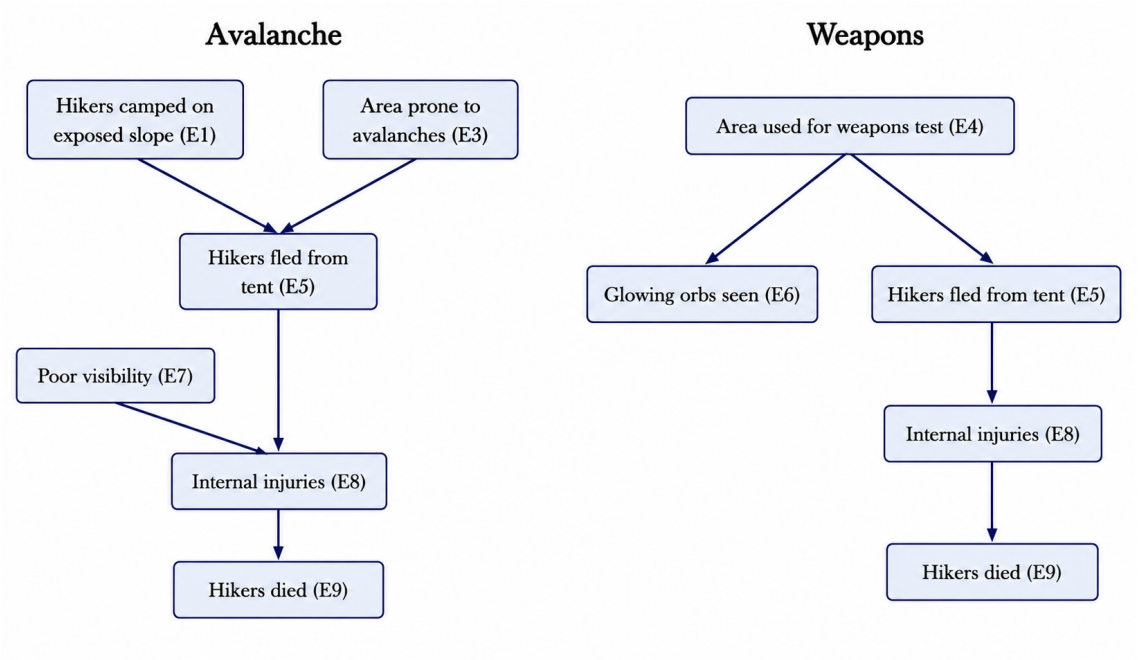

v1
v2
Causal Distance (Mean ± SE)
0
2
4
S1
S2
S3
S4
Story


v1
v2
Counterfactual Distance (Mean ± SE)
0
1
2
S1
S2
S3
S4
Story


v1
v2
Consistency Distance (Mean ± SE)
2.0
1.5
1.0
0.5
0.0
S1
S2
S3
S4
Story

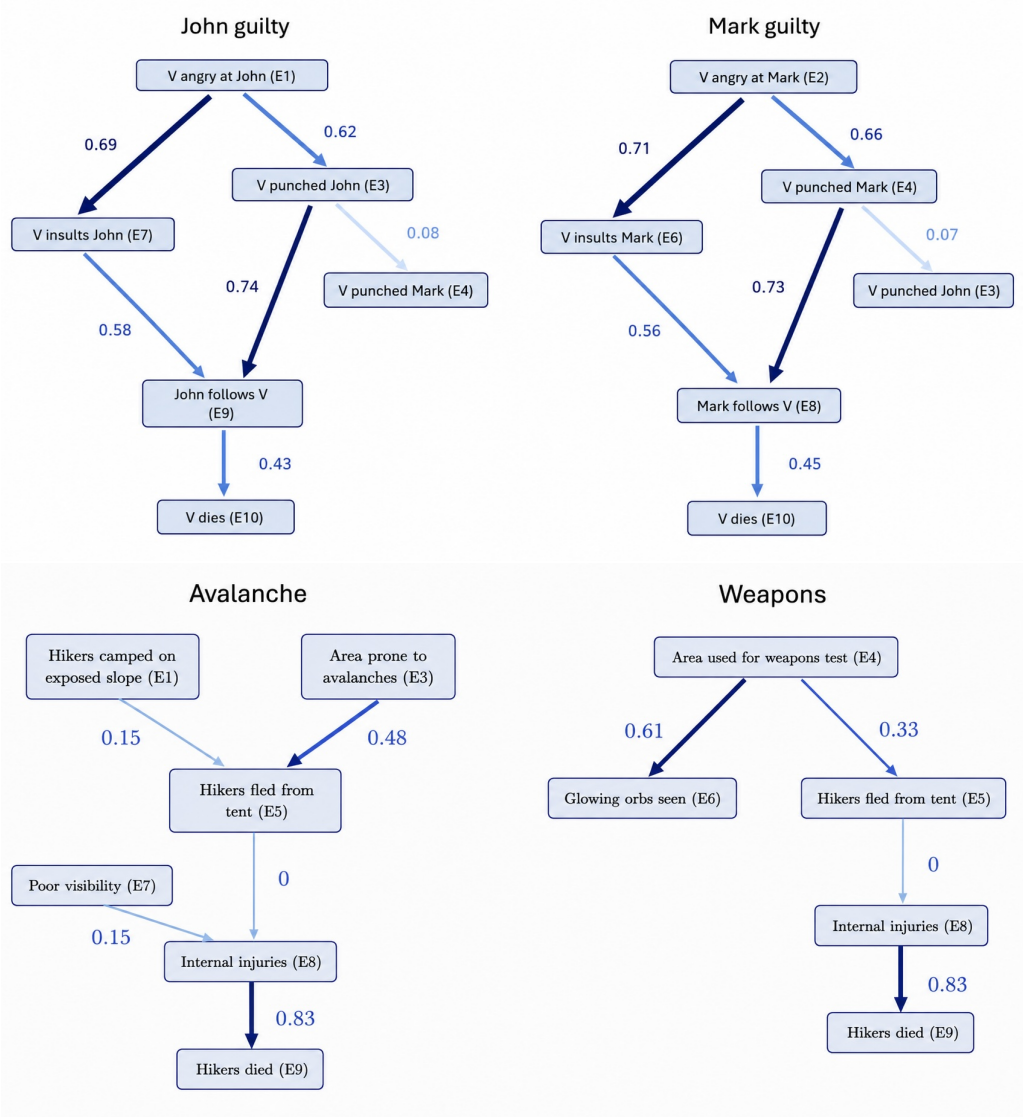
£


John guilty
V angry at John (E1)
V punched John (E3)
V insults John (E7)
V punched Mark (E4)
John follows V (E9)
V dies (E10)
Mark guilty
V angry at Mark (E2)
V insults Mark (E6)
V punched Mark (E4)
V punched John (E3)
Mark follows V (E8)
V dies (E10)


Avalanche
Hikers camped on exposed slope (E1)
Area prone to avalanches (E3)
Hikers fled from tent (E5)
Poor visibility (E7)
Internal injuries (E8)
Hikers died (E9)
Weapons
Area used for weapons test (E4)
Glowing orbs seen (E6)
Hikers fled from tent (E5)
Internal injuries (E8)
Hikers died (E9)

- 
- 
- 
- 

NaN

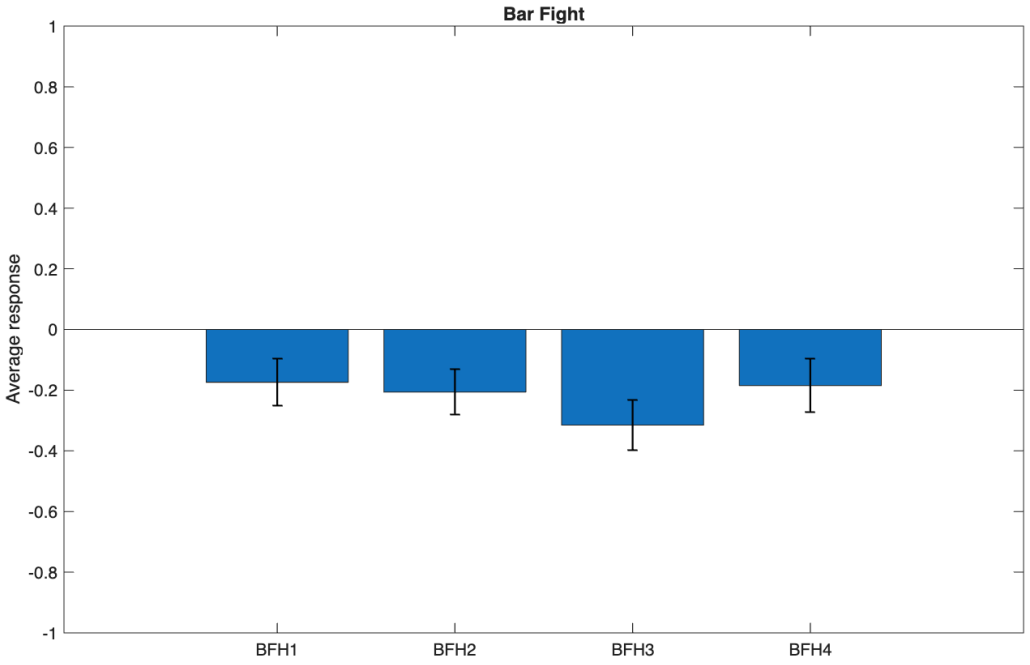

John guilty
V angry at John (E1)
0.69
0.62
V insults John (E7)
V punched John (E3)
0.08
0.74
V punched Mark (E4)
0.58
John follows V (E9)
0.43
V dies (E10)
Mark guilty
V angry at Mark (E2)
0.71
0.66
V insults Mark (E6)
V punched Mark (E4)
0.07
0.73
V punched John (E3)
0.56
Mark follows V (E8)
0.45
V dies (E10)
Avalanche
Hikers camped on exposed slope (E1)
Area prone to avalanches (E3)
0.15
0.48
Hikers fled from tent (E5)
0
Poor visibility (E7)
0.15
Internal injuries (E8)
0.83
Hikers died (E9)
Weapons
Area used for weapons test (E4)
0.61
0.33
Glowing orbs seen (E6)
Hikers fled from tent (E5)
0
Internal injuries (E8)
0.83
Hikers died (E9)

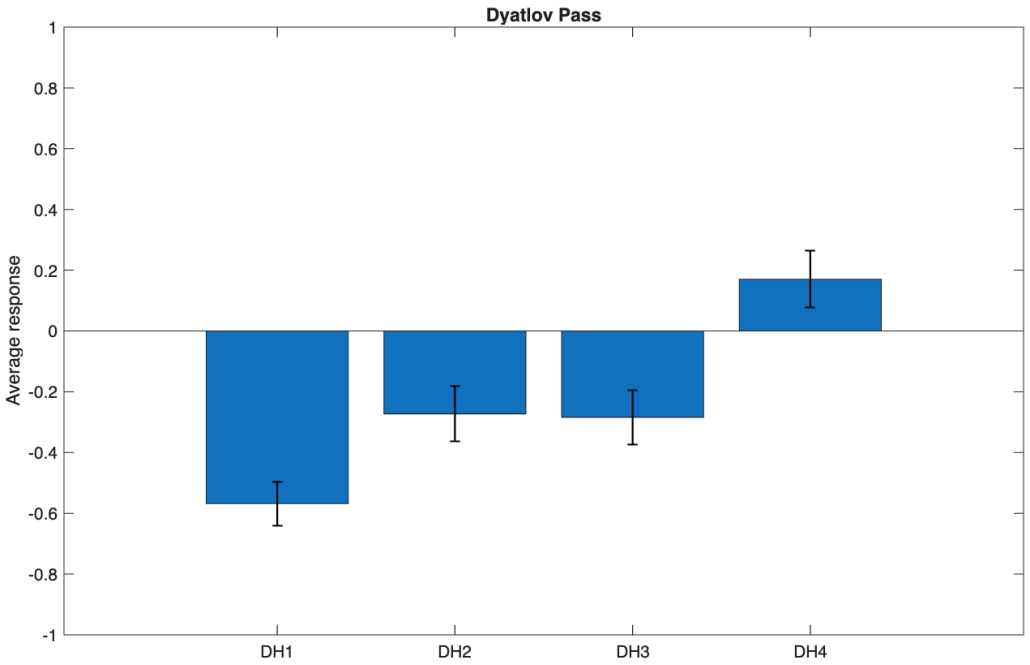
**Bar Fight**

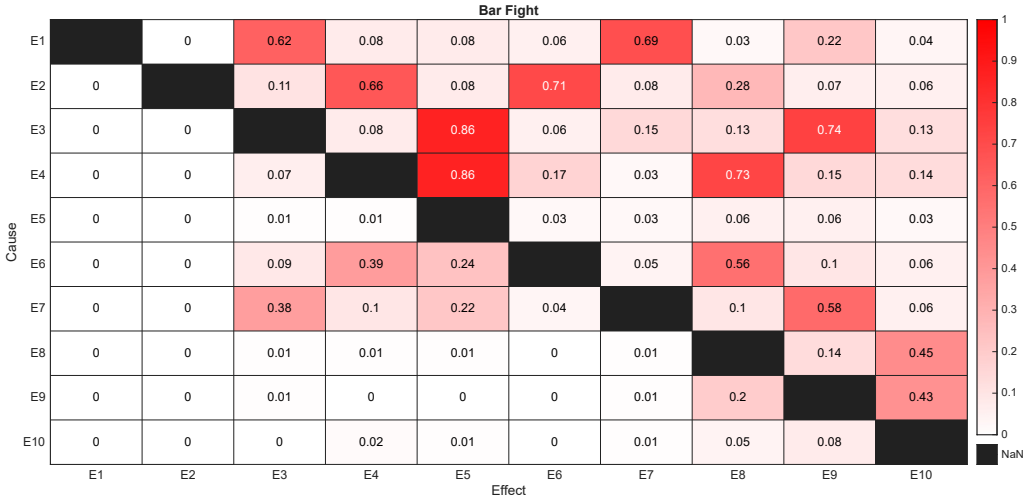
| Question | John | Mark | Mean |
|---|---|---|---|
| BFH1 | 1 | +1 | 0.17 |
| BFH2 | +1 | 1 | 0.21 |
| BFH3 | +1 | 1 | 0.32 |
| BFH4 | 1 | +1 | 0.18 |

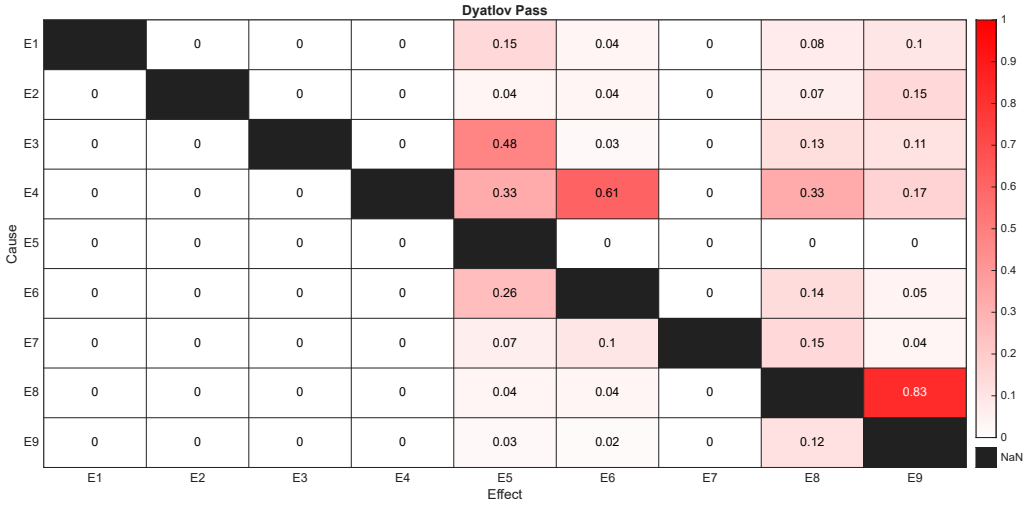
**Dyatlov Pass**

| Question | Avalanche | Weapons | Mean |
|---|---|---|---|
| DH1 | 1 | +1 | 0.57 |
| DH2 | 1 | +1 | 0.27 |
| DH3 | 1 | +1 | 0.28 |
| DH4 | +1 | +1 | 0.17 |

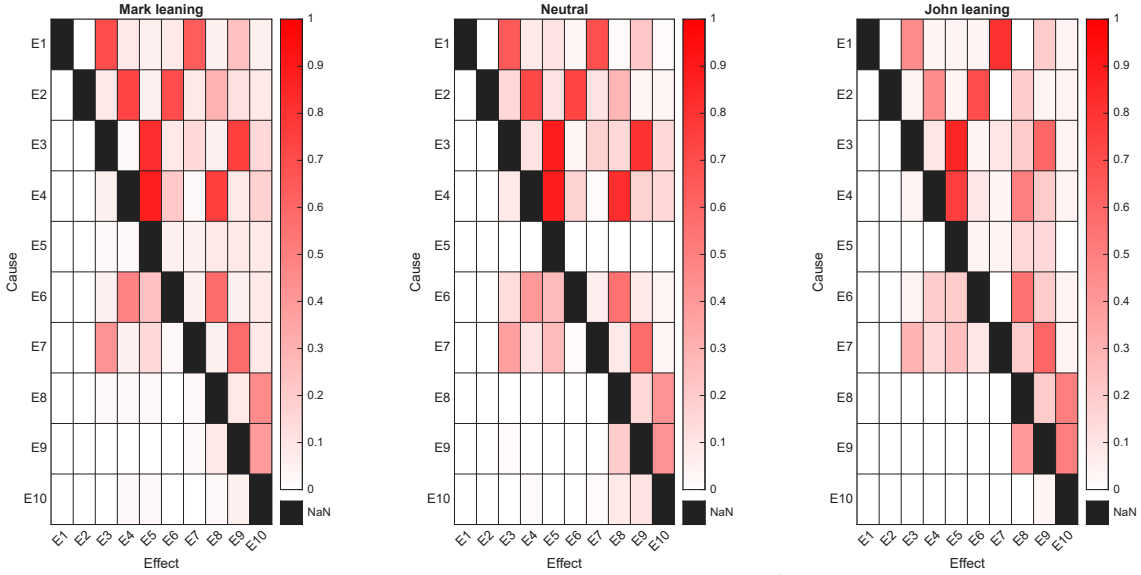

Bar Fight
Average response
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
BFH1
BFH2
BFH3
BFH4


Dyatlov Pass
Average response
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
DH1
DH2
DH3
DH4

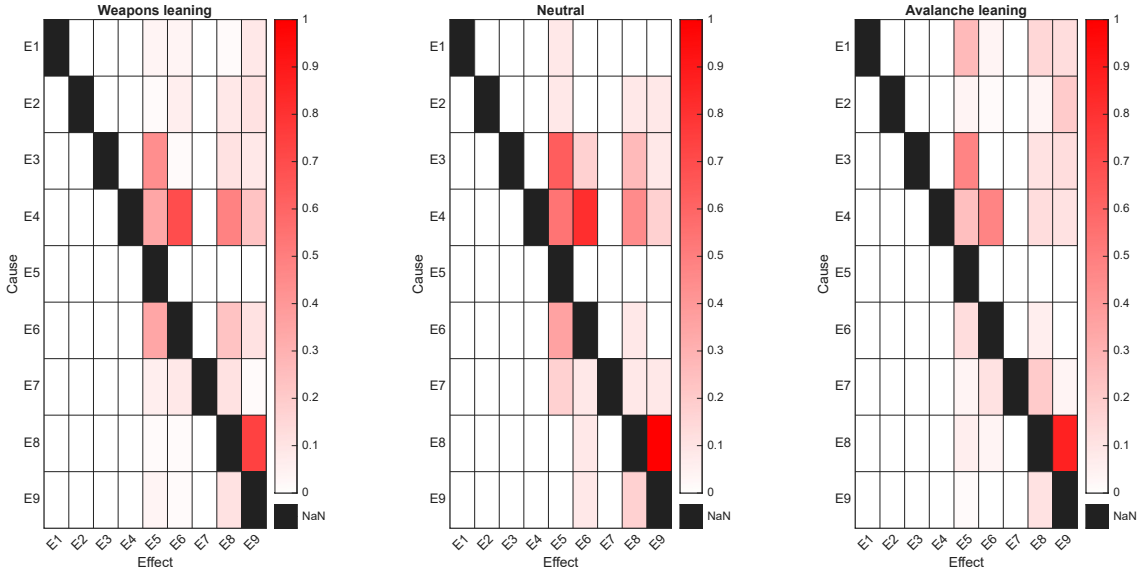

Bar Fight
Cause
Effect
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10
E1 0 0.62 0.08 0.08 0.06 0.69 0.03 0.22 0.04
E2 0 0.11 0.66 0.08 0.71 0.08 0.28 0.07 0.06
E3 0 0 0.08 0.86 0.06 0.15 0.13 0.74 0.13
E4 0 0 0.07 0.86 0.17 0.03 0.73 0.15 0.14
E5 0 0 0.01 0.01 0.03 0.03 0.06 0.06 0.03
E6 0 0 0.09 0.39 0.24 0.05 0.56 0.1 0.06
E7 0 0 0.38 0.1 0.22 0.04 0.1 0.58 0.06
E8 0 0 0.01 0.01 0.01 0 0.01 0.14 0.45
E9 0 0 0.01 0 0 0 0.01 0.2 0.43
E10 0 0 0 0.02 0.01 0 0.01 0.05 0.08
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
NaN

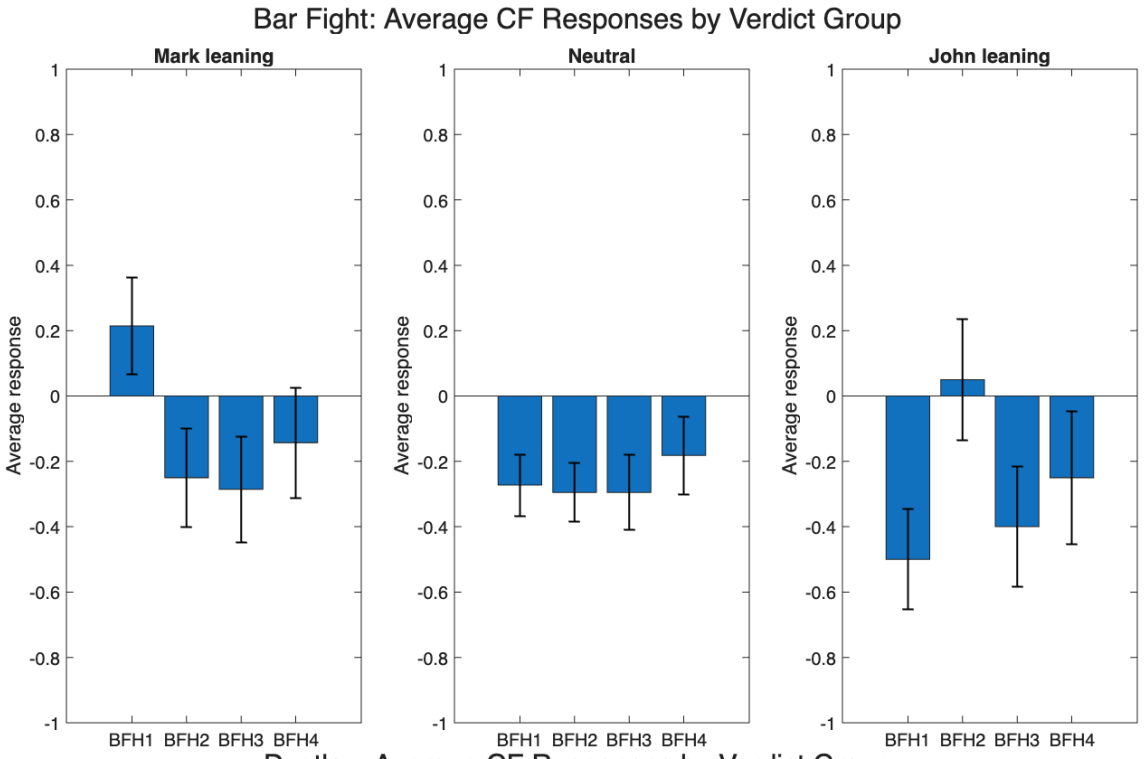

Dyatlov Pass
Cause
Effect
E1 E2 E3 E4 E5 E6 E7 E8 E9
E1 0 0 0 0.15 0.04 0 0.08 0.1
E2 0 0 0 0 0.04 0.04 0 0.07 0.15
E3 0 0 0 0 0.48 0.03 0 0.13 0.11
E4 0 0 0 0 0.33 0.61 0 0.33 0.17
E5 0 0 0 0 0 0 0 0 0
E6 0 0 0 0 0.26 0 0 0.14 0.05
E7 0 0 0 0 0.07 0.1 0 0.15 0.04
E8 0 0 0 0 0.04 0.04 0 0 0.83
E9 0 0 0 0 0.03 0.02 0 0.12 0
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
NaN

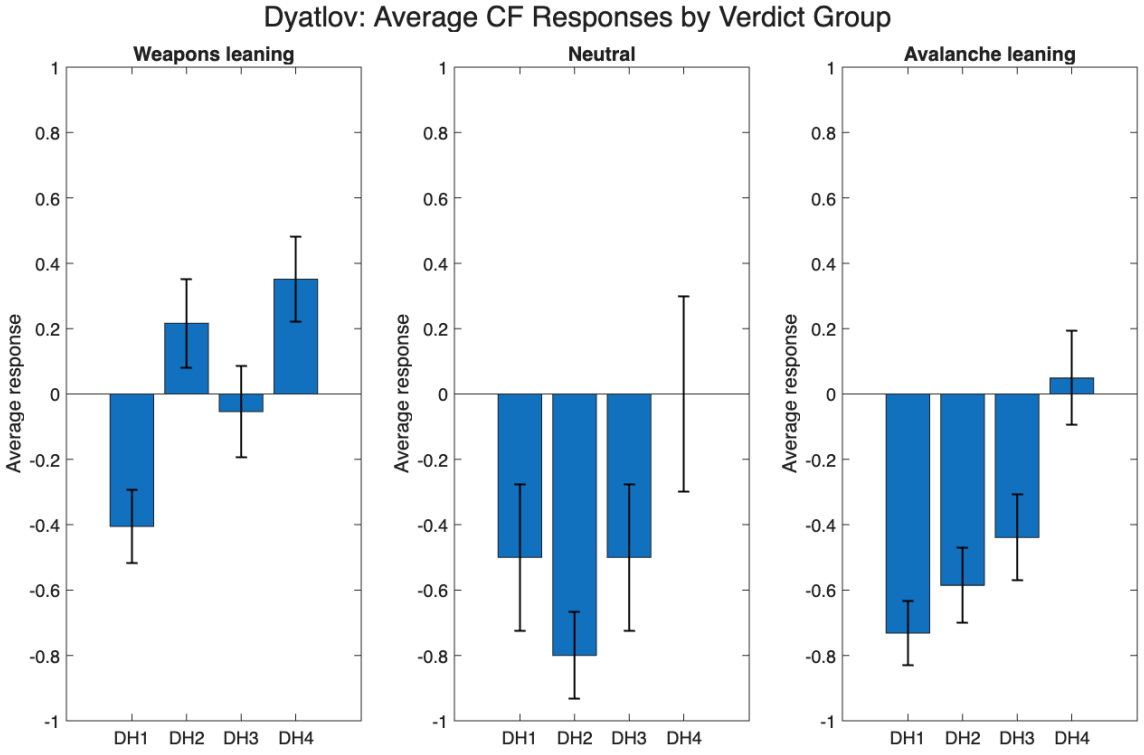
Bar Fight: Average Adjacency Matrix by Verdict Group




Dyatlov: Average Adjacency Matrix by Verdict Group




Bar Fight: Average CF Responses by Verdict Group
Mark leaning
Neutral
John leaning
Average response
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
BFH1 BFH2 BFH3 BFH4


Dyatlov: Average CF Responses by Verdict Group
Weapons leaning
Neutral
Avalanche leaning
Average response
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
DH1 DH2 DH3 DH4

| | £ | £ |
| --- | --- | --- |

$\mathfrak{F}$ $\mathfrak{F}$

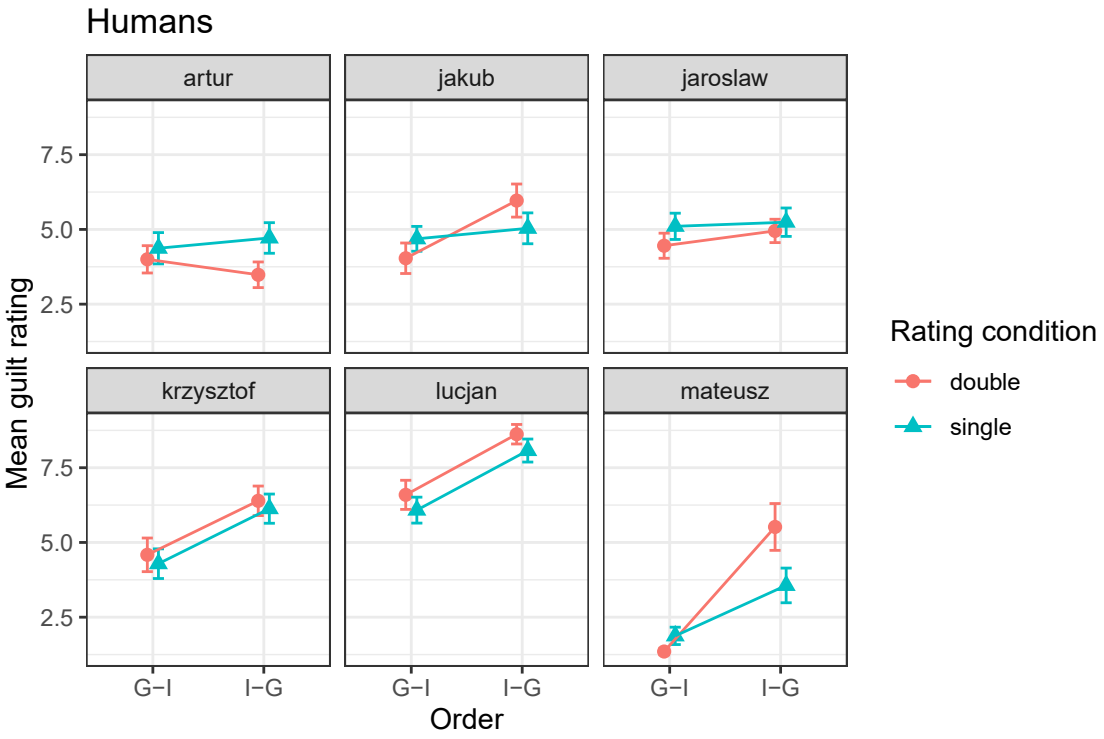
$\mathfrak{F}$ $\mathfrak{F}$

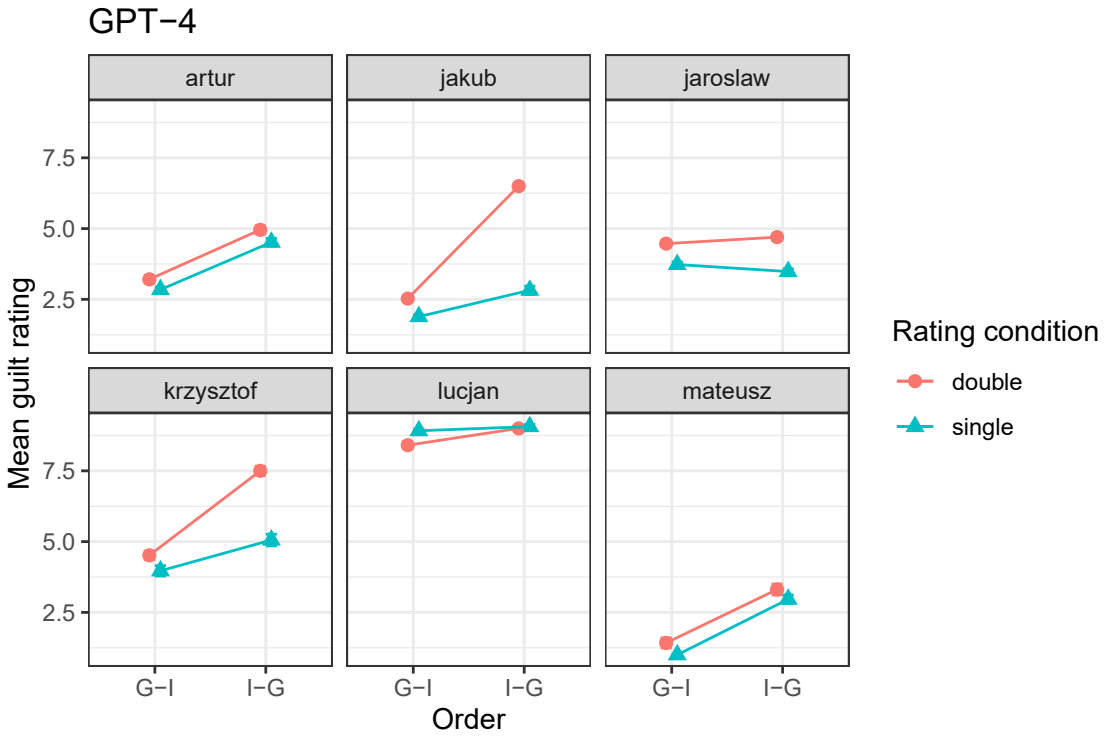
$\mathfrak{F}$ $\mathfrak{F}$

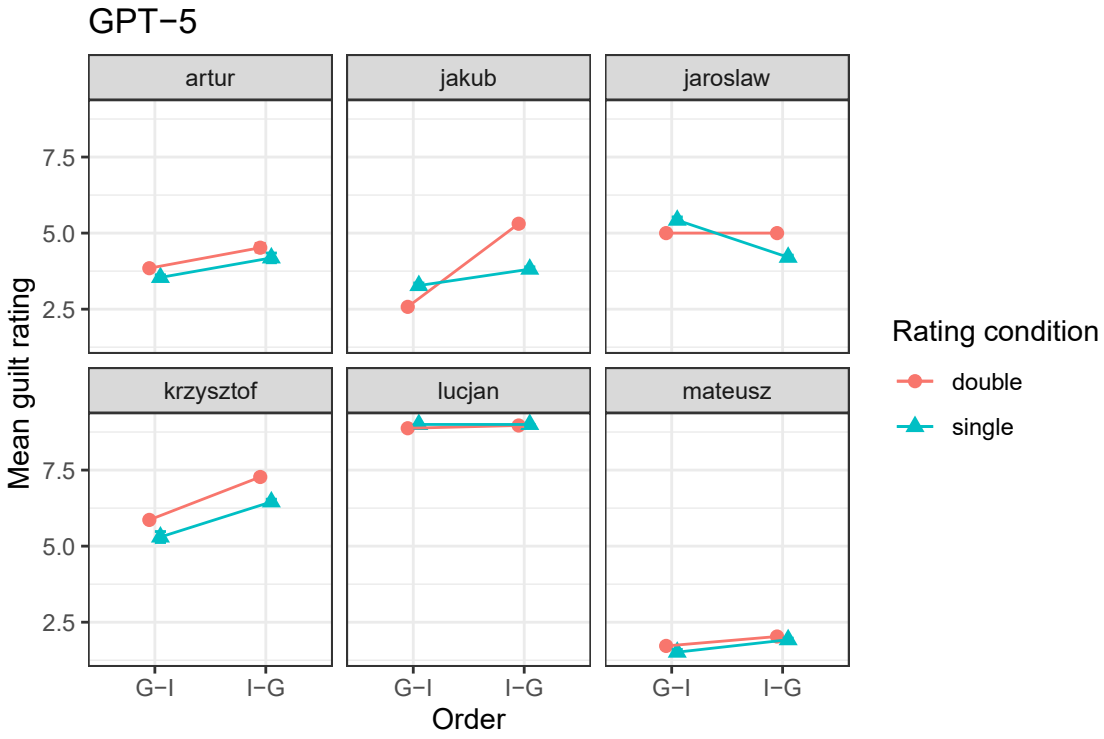
| | £ | £ |
| --- | --- | --- |
| | | |

{ }

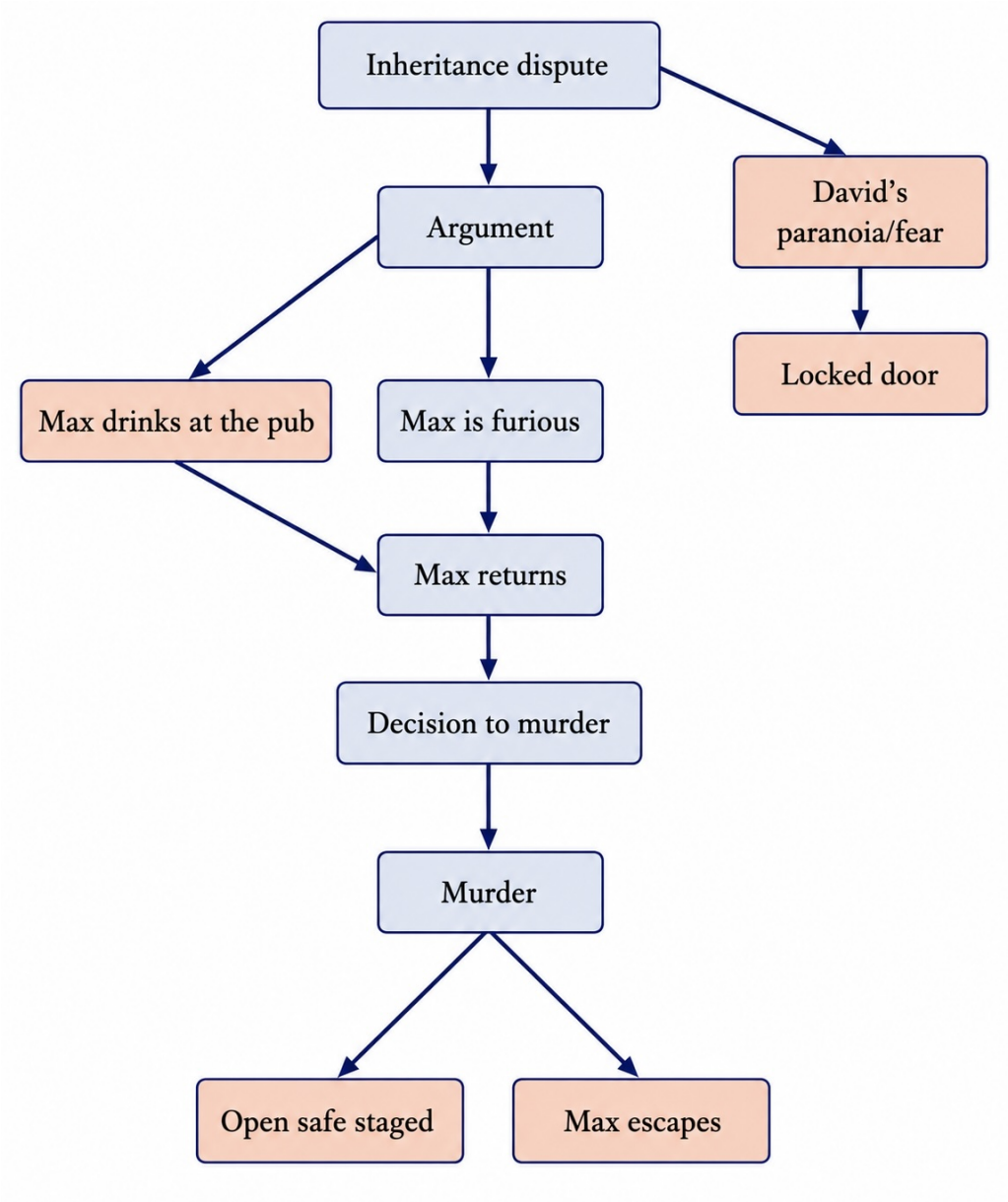
{ }

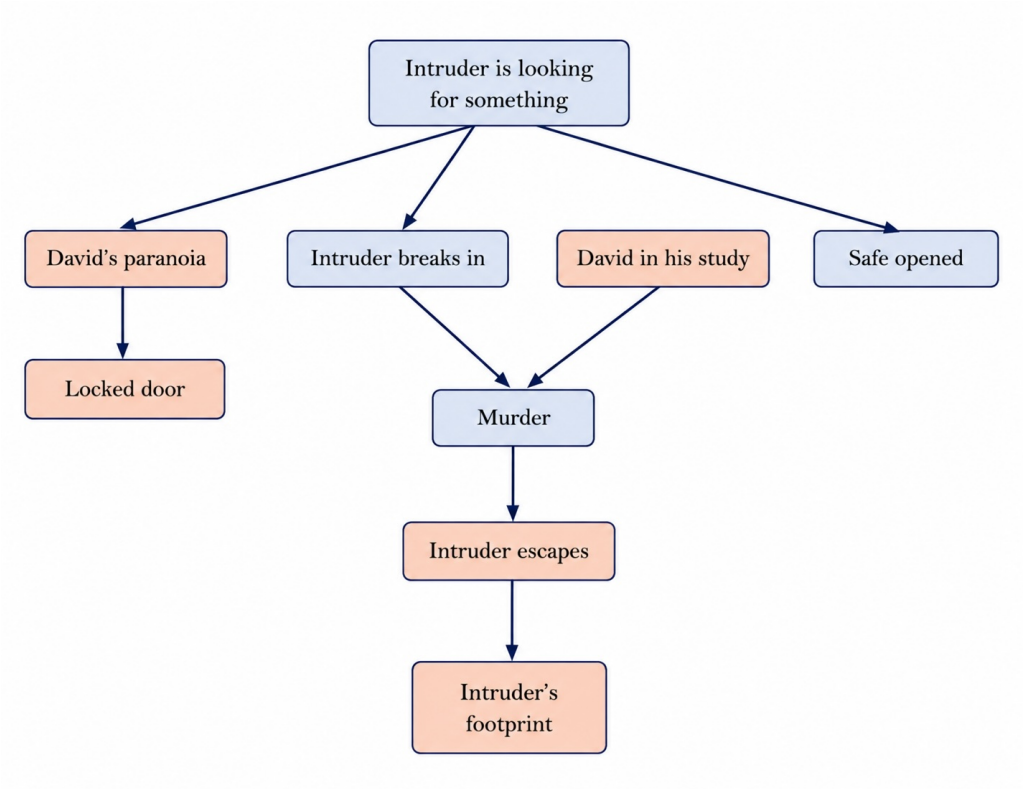

Humans
artur
jakub
jaroslaw
krzysztof
lucjan
mateusz
Mean guilt rating
7.5
5.0
2.5
G–I
I–G
Order
Rating condition
double
single


GPT−4
artur
jakub
jaroslaw
krzysztof
lucjan
mateusz
Mean guilt rating
7.5
5.0
2.5
G−I
I−G
Order
Rating condition
double
single


GPT−5
artur
jakub
jaroslaw
krzysztof
lucjan
mateusz
Mean guilt rating
7.5
5.0
2.5
G−I
I−G
Order
Rating condition
double
single

£

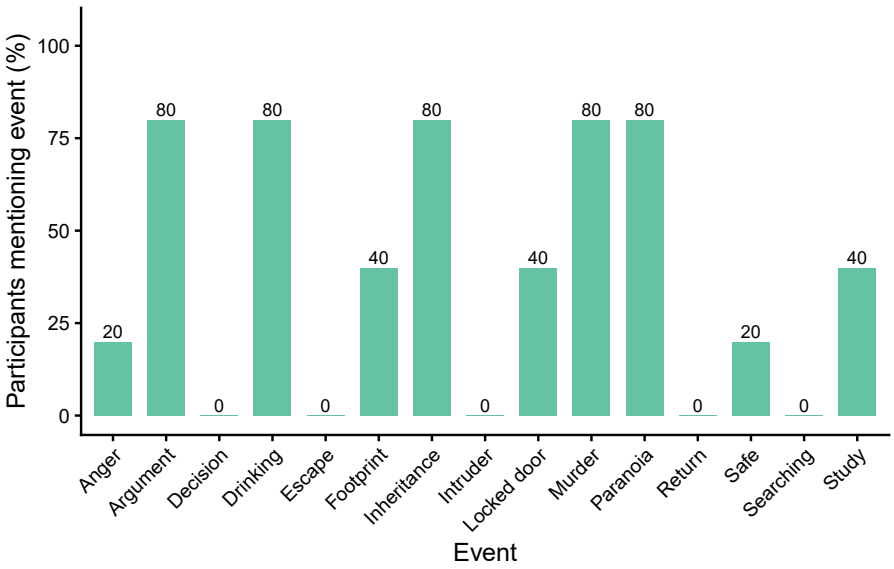

Inheritance dispute
Argument
David's paranoia/fear
Locked door
Max drinks at the pub
Max is furious
Max returns
Decision to murder
Murder
Open safe staged
Max escapes

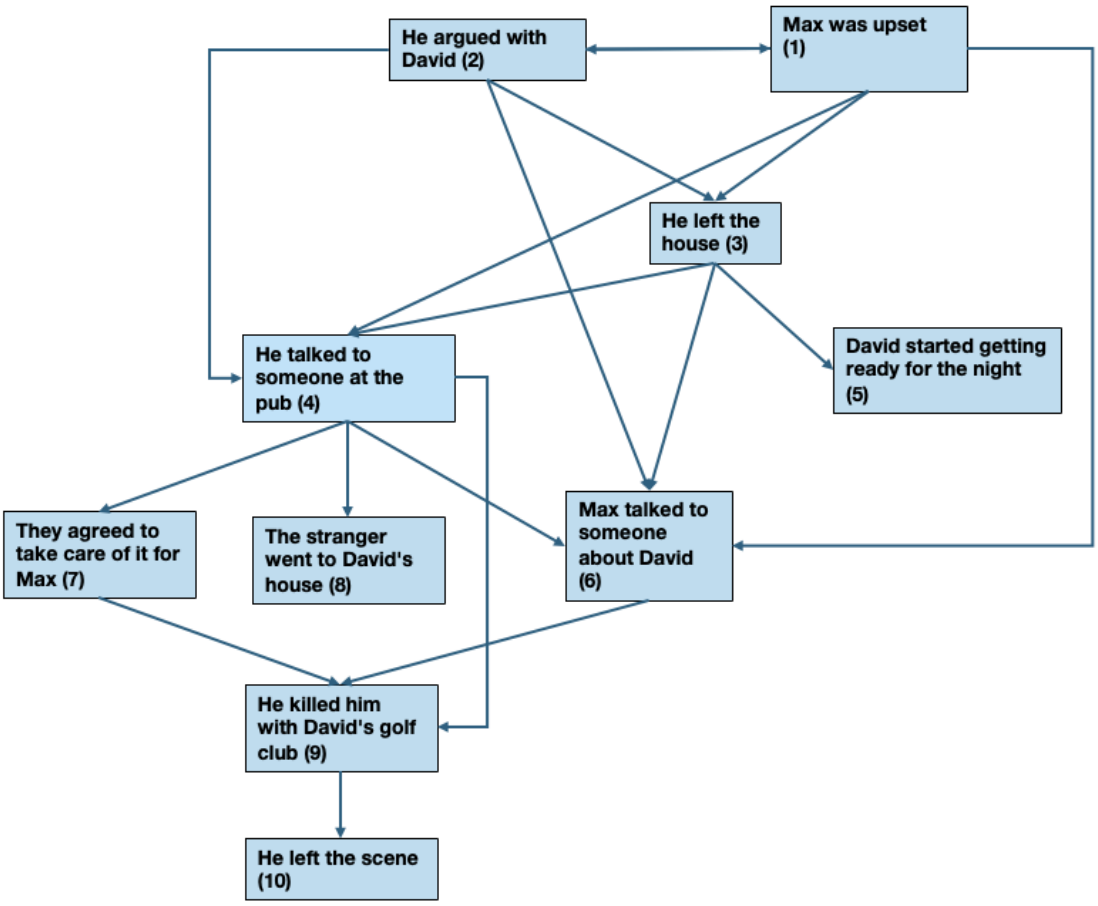

Intruder is looking for something
David's paranoia
Intruder breaks in
David in his study
Safe opened
Locked door
Murder
Intruder escapes
Intruder's footprint

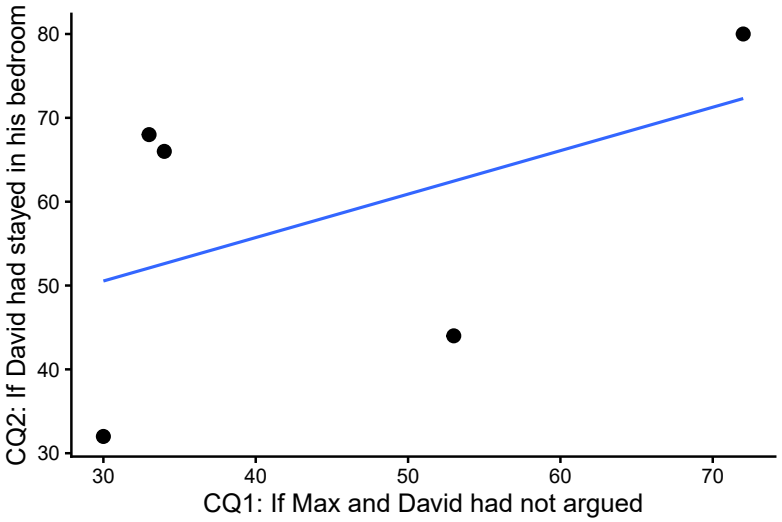
Suppose Max and David hadn't argued that day. Given this change, do you think David would still have been murdered that night?

Definitely not murdered | Probably not murdered | Probably still murdered | Definitely still murdered

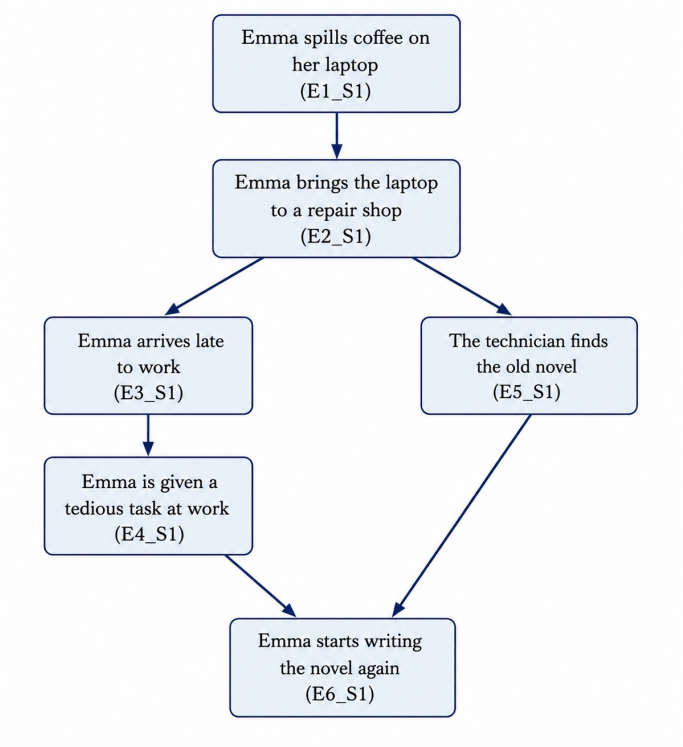



Participants mentioning event (%)
100
75
50
25
0
20
80
0
80
0
40
80
0
40
80
80
0
20
0
40
Anger
Argument
Decision
Drinking
Escape
Footprint
Inheritance
Intruder
Locked door
Murder
Paranoia
Return
Safe
Searching
Study
Event


He argued with David (2)
Max was upset (1)
He left the house (3)
He talked to someone at the pub (4)
David started getting ready for the night (5)
They agreed to take care of it for Max (7)
The stranger went to David's house (8)
Max talked to someone about David (6)
He killed him with David's golf club (9)
He left the scene (10)


CQ2: If David had stayed in his bedroom
80
70
60
50
40
30
30
40
50
60
70
CQ1: If Max and David had not argued

£


Emma spills coffee on her laptop (E1_S1)
Emma brings the laptop to a repair shop (E2_S1)
Emma arrives late to work (E3_S1)
The technician finds the old novel (E5_S1)
Emma is given a tedious task at work (E4_S1)
Emma starts writing the novel again (E6_S1)

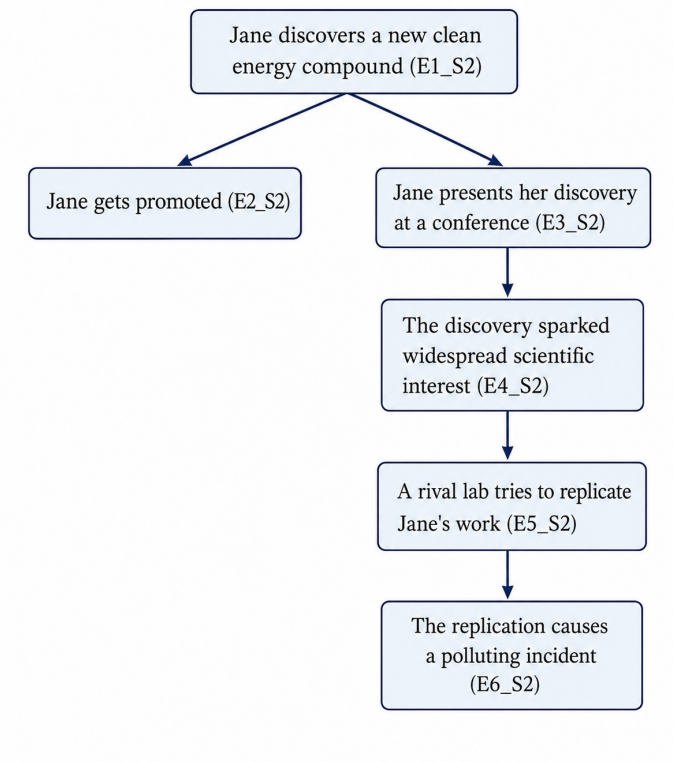
- _ _

- _ _

- _ _

- _ _

- _ _

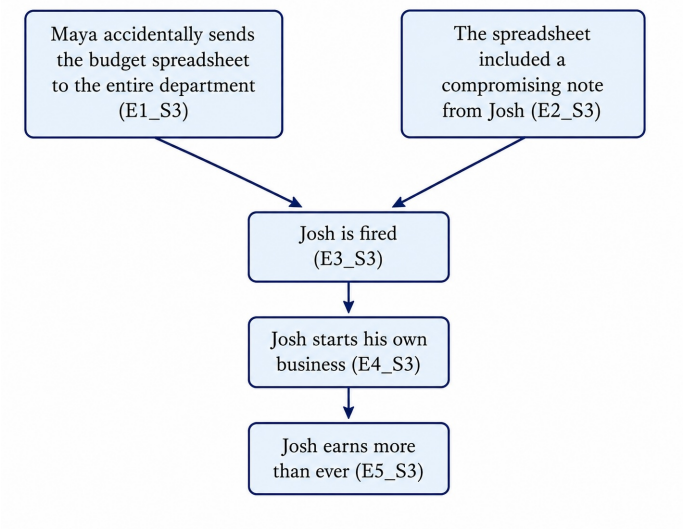

Jane discovers a new clean energy compound (E1_S2)
Jane gets promoted (E2_S2)
Jane presents her discovery at a conference (E3_S2)
The discovery sparked widespread scientific interest (E4_S2)
A rival lab tries to replicate Jane's work (E5_S2)
The replication causes a polluting incident (E6_S2)

- _ _

- _ _

- _ _

- _ _

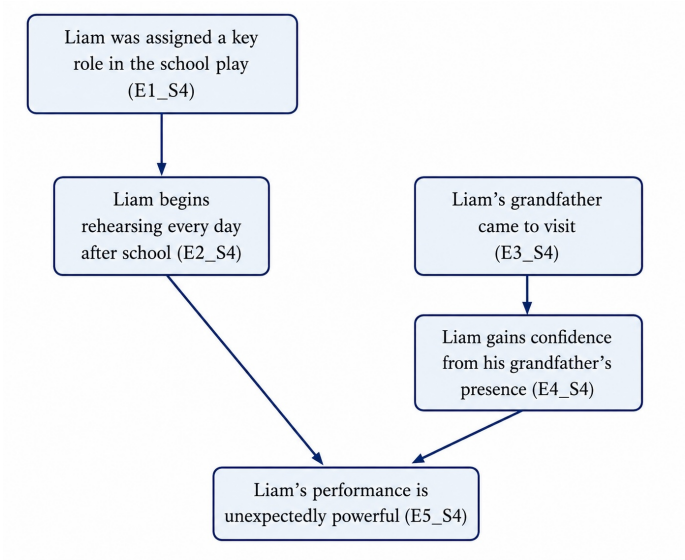

Maya accidentally sends the budget spreadsheet to the entire department (E1_S3)
The spreadsheet included a compromising note from Josh (E2_S3)
Josh is fired (E3_S3)
Josh starts his own business (E4_S3)
Josh earns more than ever (E5_S3)

- _ _

- _ _

- _ _

- _ _


Liam was assigned a key role in the school play (E1_S4)
Liam begins rehearsing every day after school (E2_S4)
Liam's grandfather came to visit (E3_S4)
Liam gains confidence from his grandfather's presence (E4_S4)
Liam's performance is unexpectedly powerful (E5_S4)

- _ _

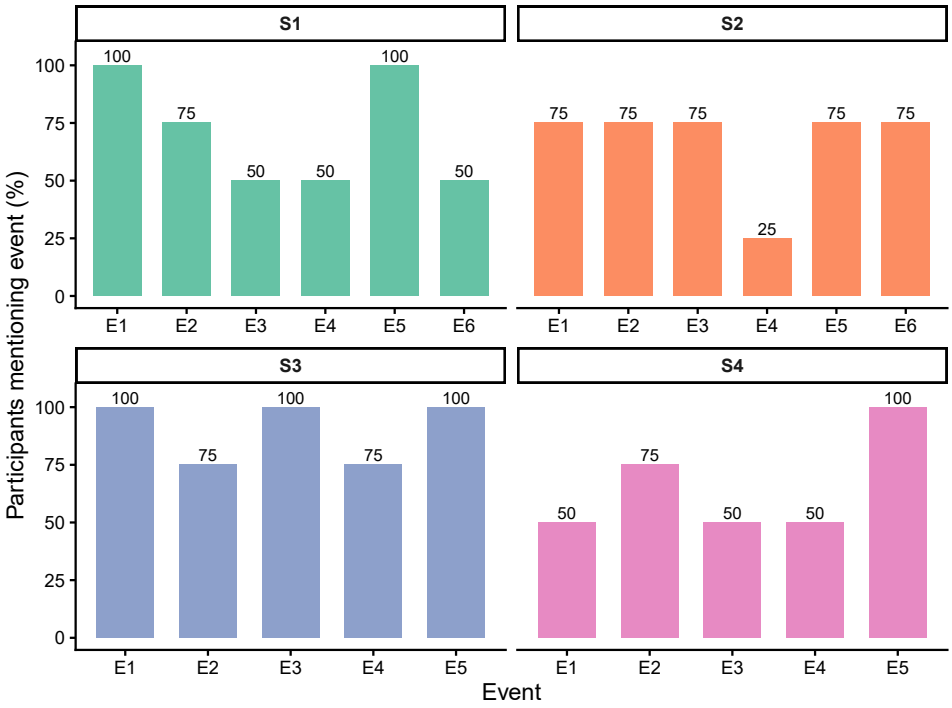
- _ _
- _ _

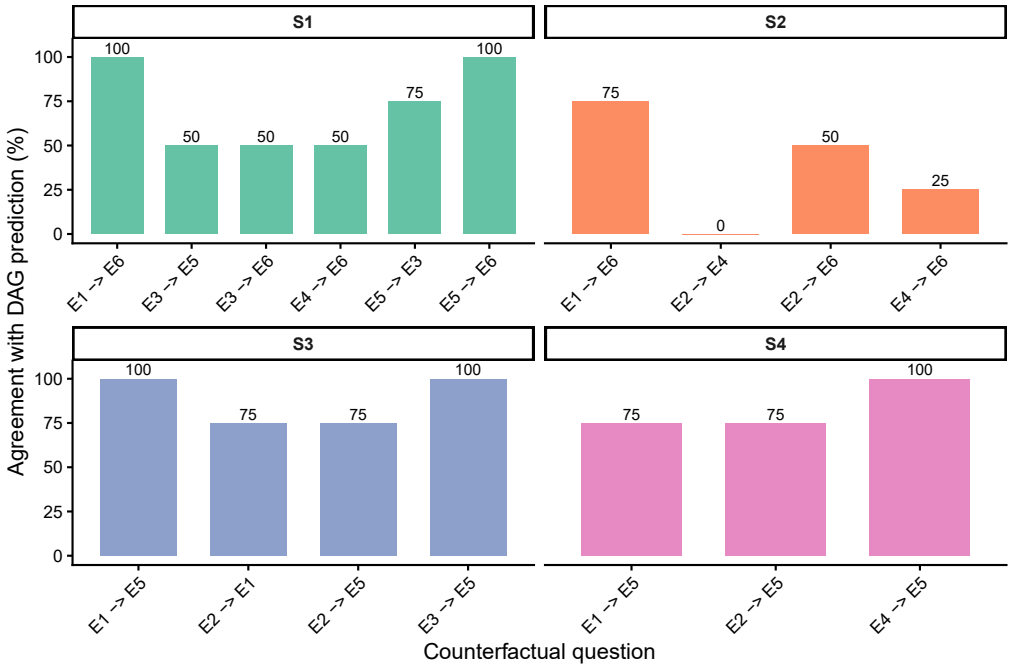
Suppose Emma hadn't spilled her coffee that morning. Do you think she would have been more, less, or equally likely to start writing again?

Much less likely | Less likely | Equally likely | More likely | Much more likely



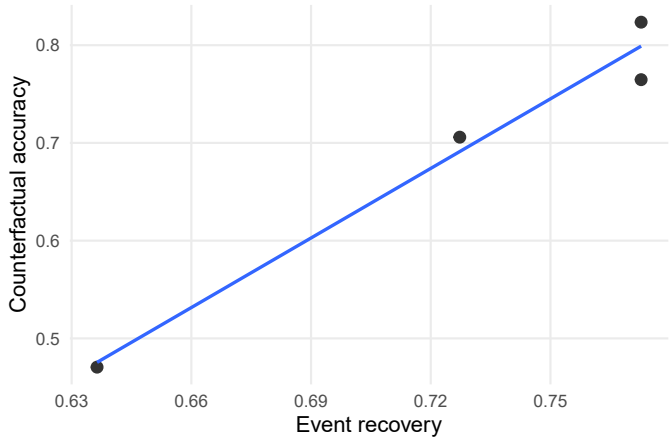
→


S1
S2
S3
S4
Participants mentioning event (%)
Event
0
25
50
75
100
E1
E2
E3
E4
E5
E6
100
75
50
50
100
50
75
75
75
25
75
75
100
75
100
75
100
50
75
50
50
100


S1
S2
S3
S4
Agreement with DAG prediction (%)
0
25
50
75
100
100
50
50
50
75
100
75
0
50
25
100
75
75
100
75
75
100
E1 -> E6
E3 -> E5
E3 -> E6
E4 -> E6
E5 -> E3
E5 -> E6
E1 -> E6
E2 -> E4
E2 -> E6
E4 -> E6
E1 -> E5
E2 -> E1
E2 -> E5
E3 -> E5
E1 -> E5
E2 -> E5
E4 -> E5
Counterfactual question

→ → → → →

→


0.8
0.7
0.6
0.5
Counterfactual accuracy
0.63
0.66
0.69
0.72
0.75
Event recovery