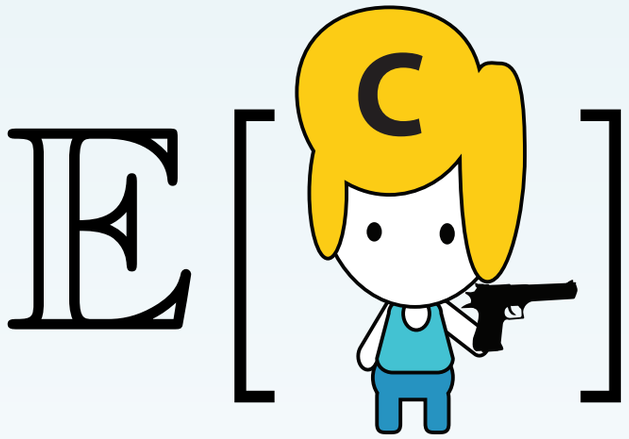
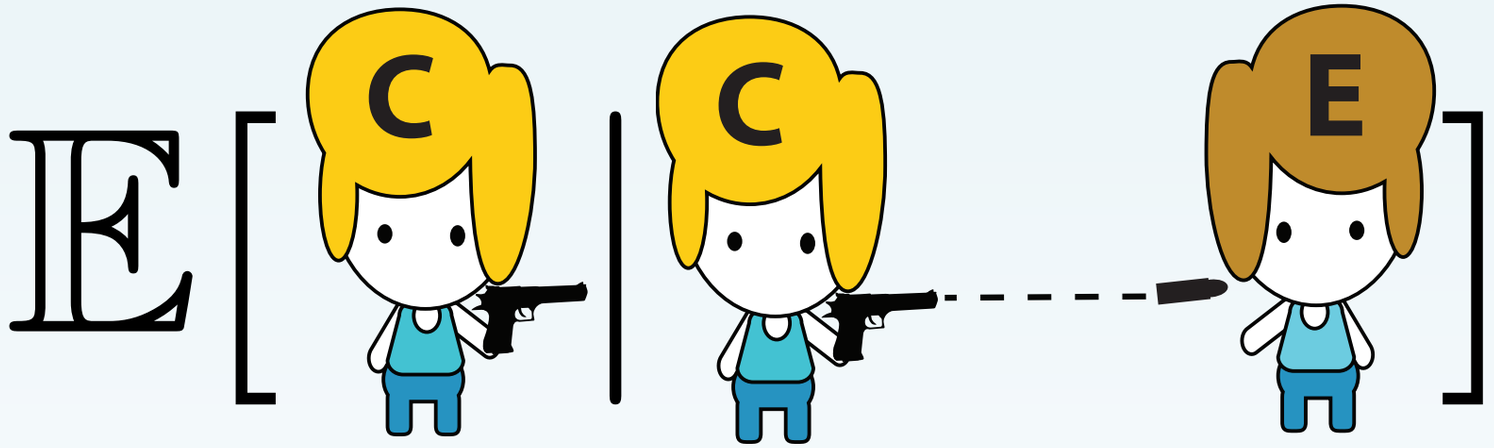
From changed expectations to attributions of responsibility

Prior expectation

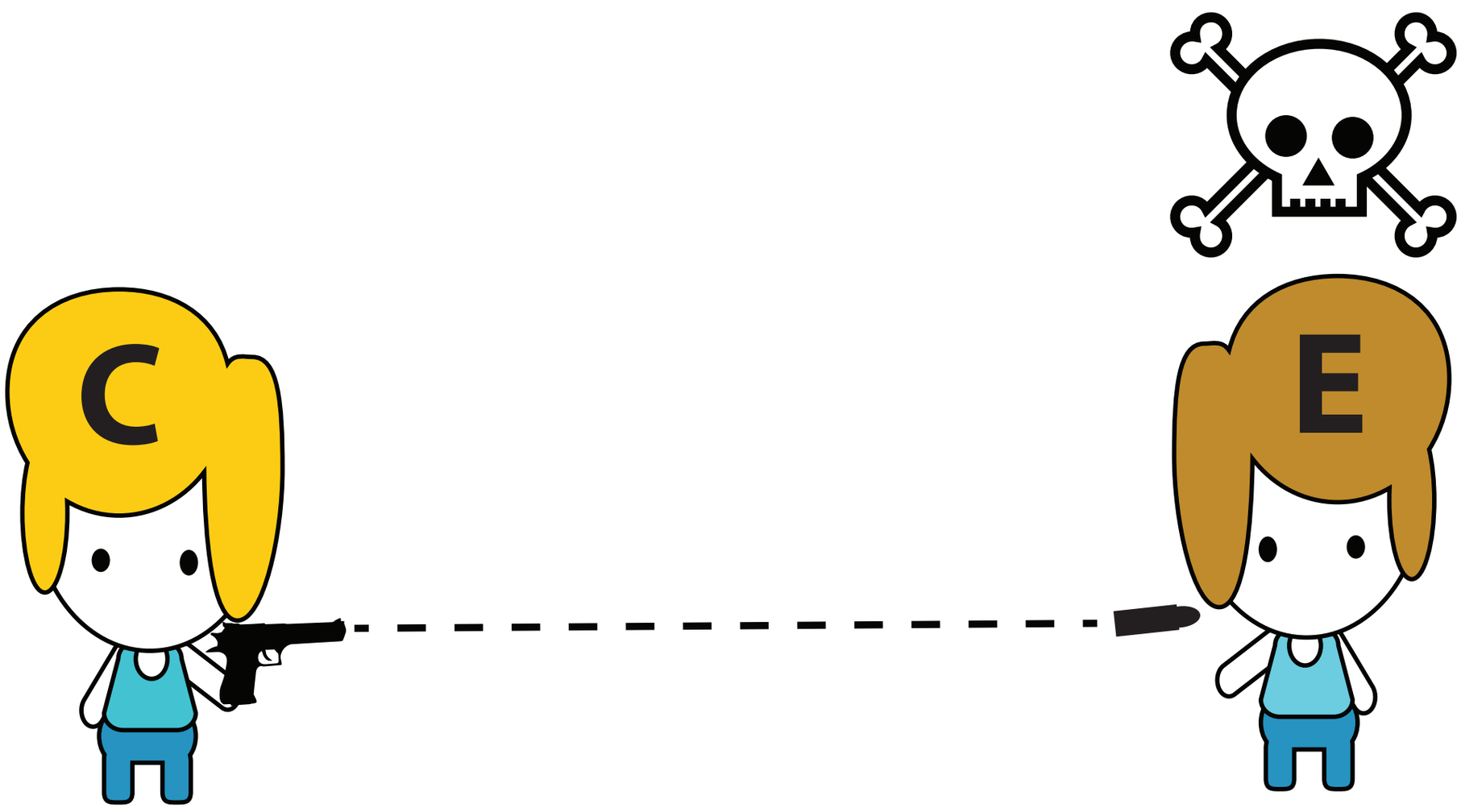


Posterior expectation

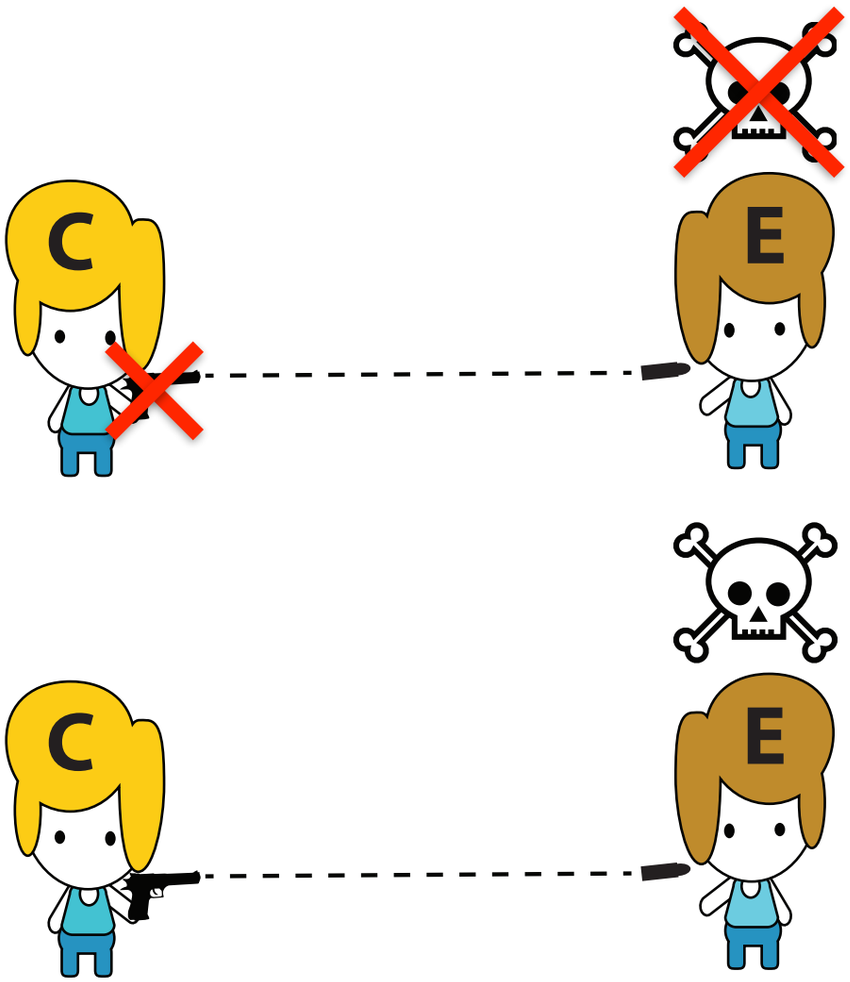


Gerstenberg, Ullman, Nagel, Kleiman-Weiner, Lagnado & Tenenbaum (under review)
Lucky or clever? From changed expectations to responsibility judgments.

How responsible was Carl for Ernie's death?

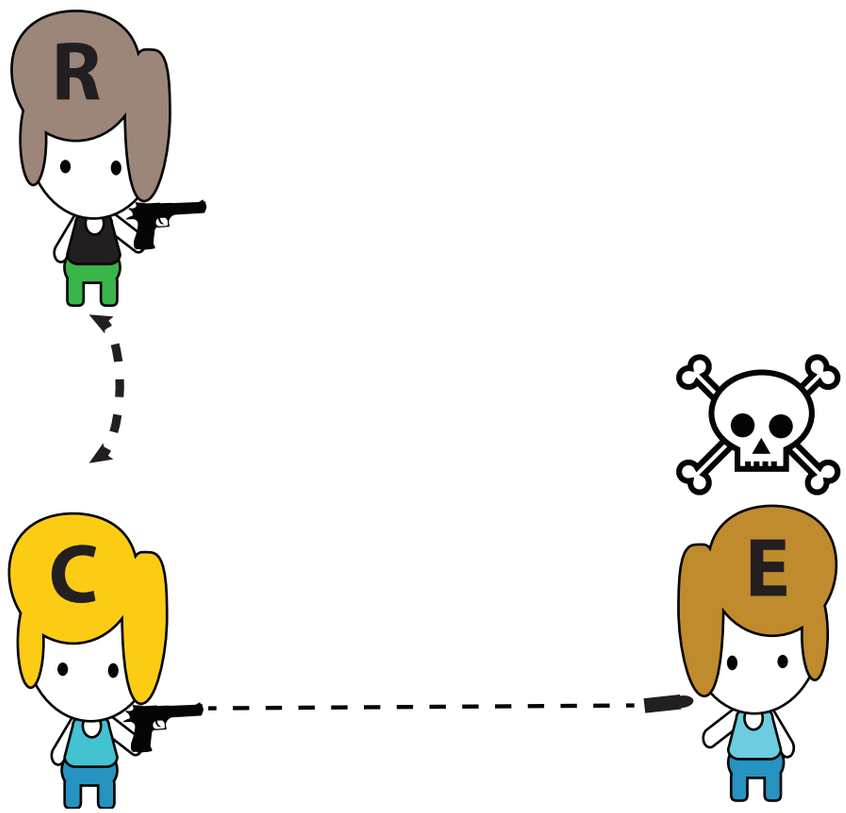


Action-centered
counterfactual



What if he
hadn't done that?

Person-centered
counterfactual



What would someone
else have done?

Action-centered counterfactual

The “but for” test

“Cause-in-fact is determined by the **but for test**:

But for the action, the result would not have happened. For example, but for running the red light, the collision would not have occurred.”

What if he hadn't done that?

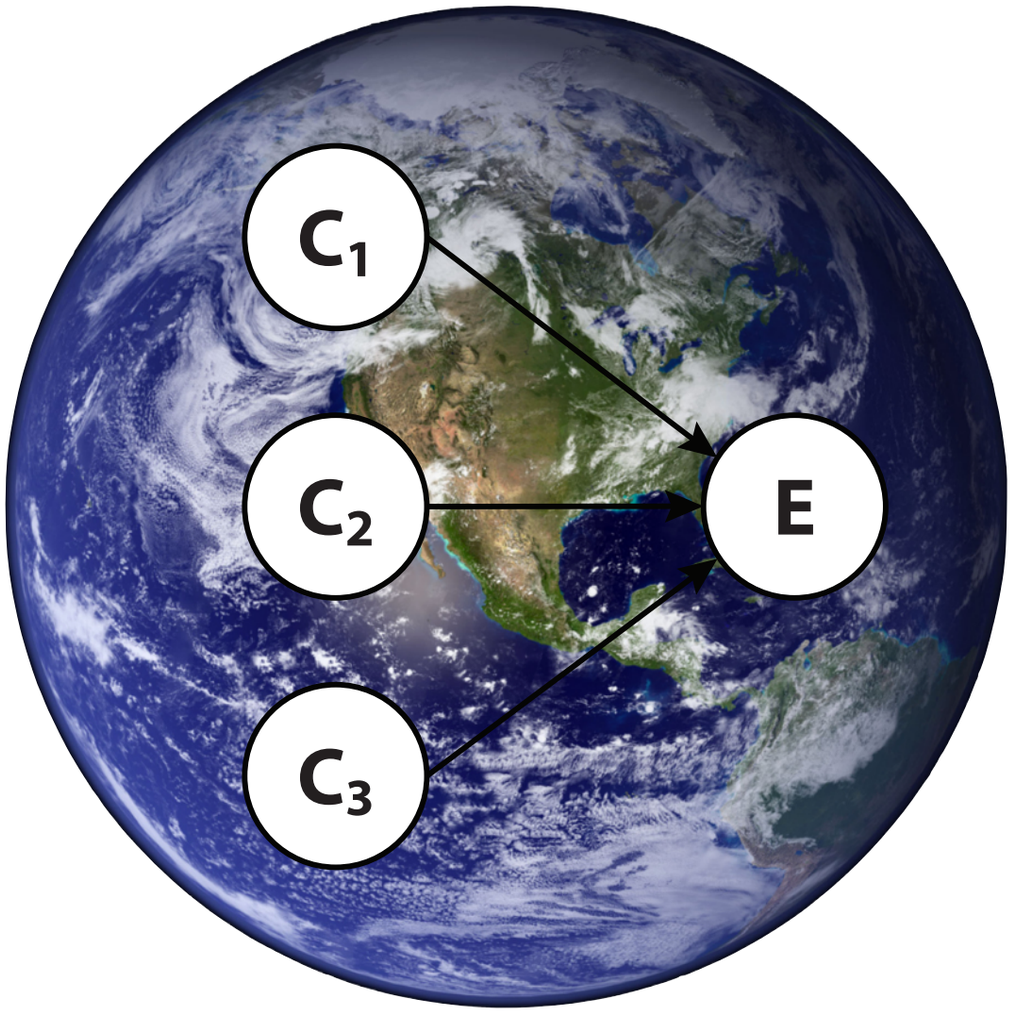
Person-centered counterfactual

The “reasonable person” test

“Each person owes a duty to **behave as a reasonable person would under the same or similar circumstances.**”

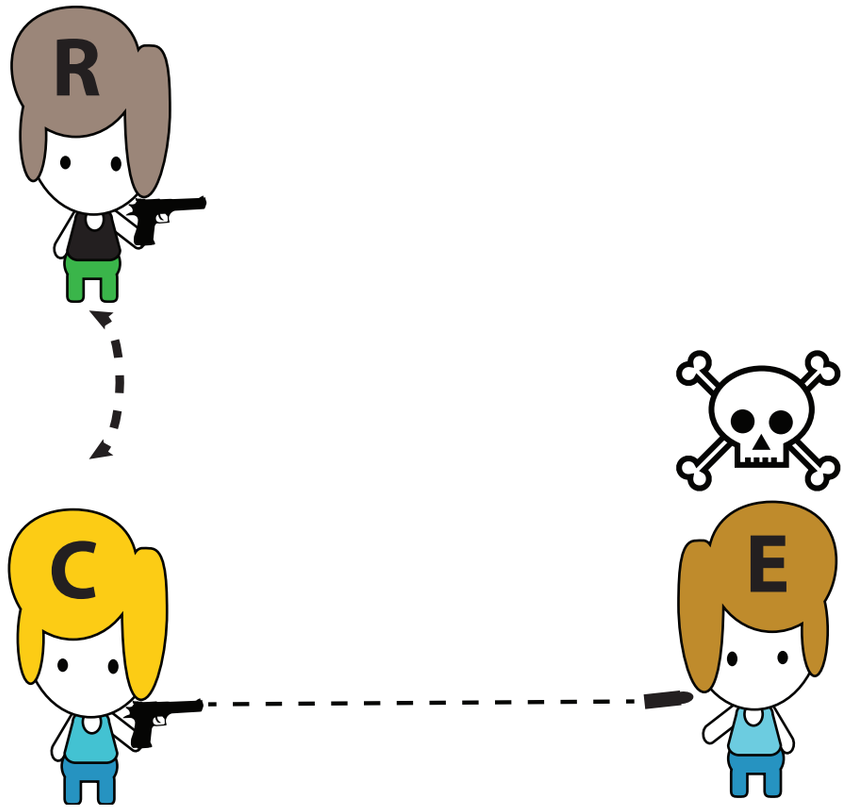
What would someone else have done?

Action-centered
counterfactual



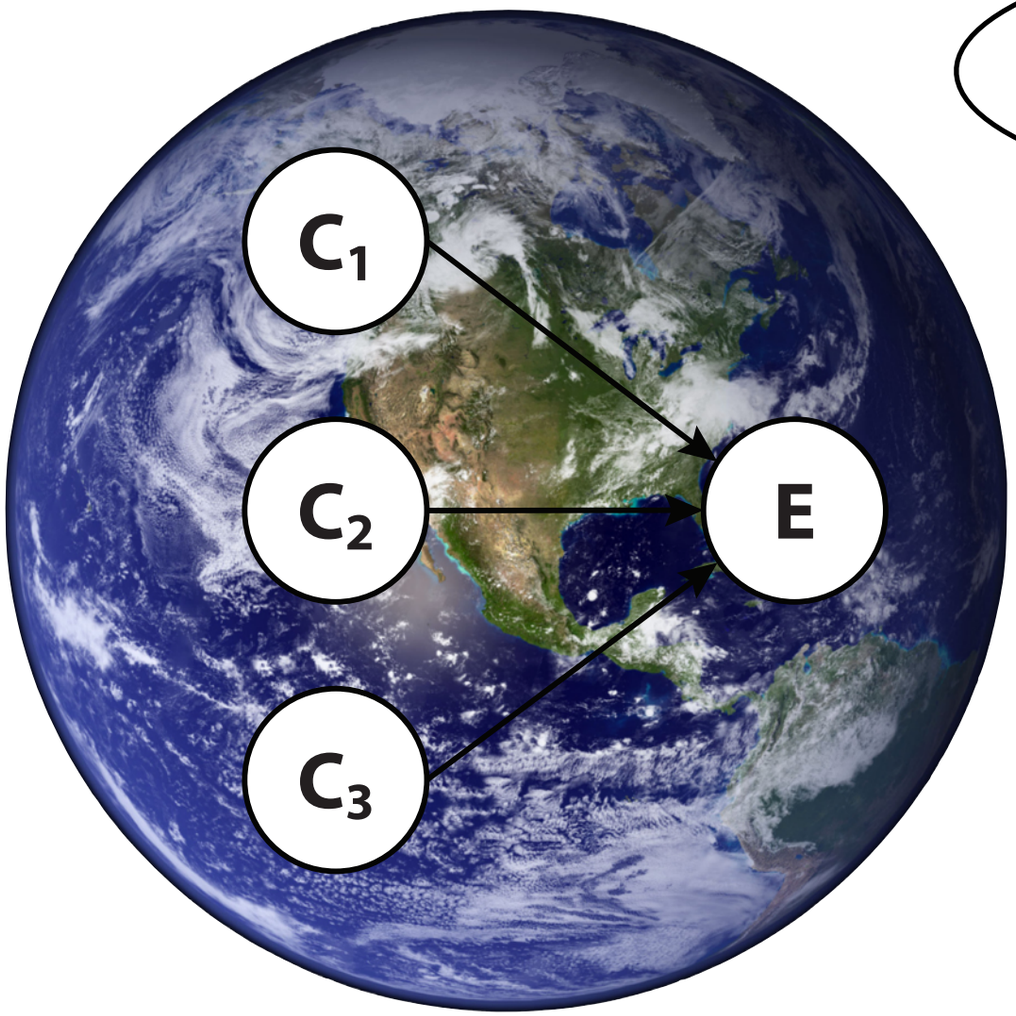
What if he
hadn't done that?

Person-centered
counterfactual



What would someone
else have done?

Action-centered
counterfactual

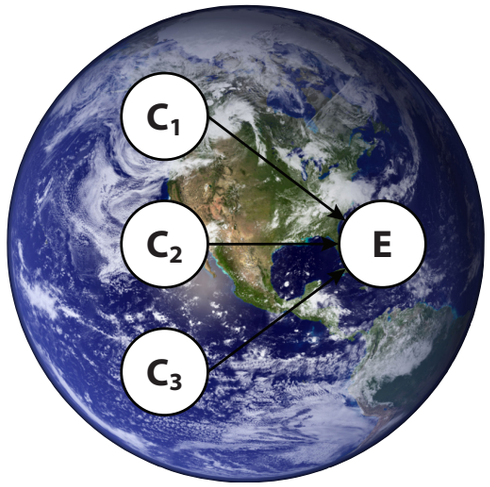


What if he
hadn't done that?

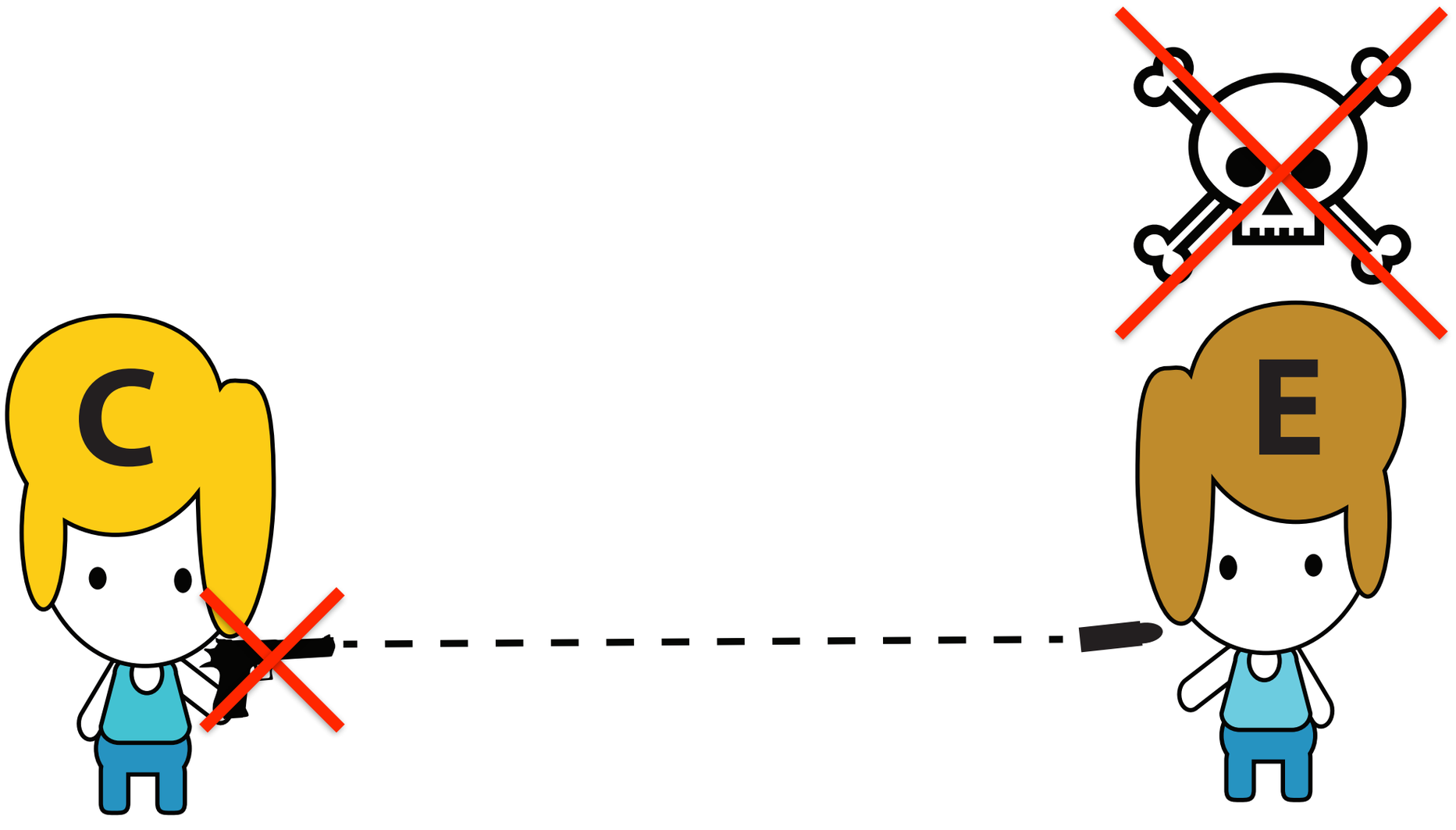
Person-centered
counterfactual

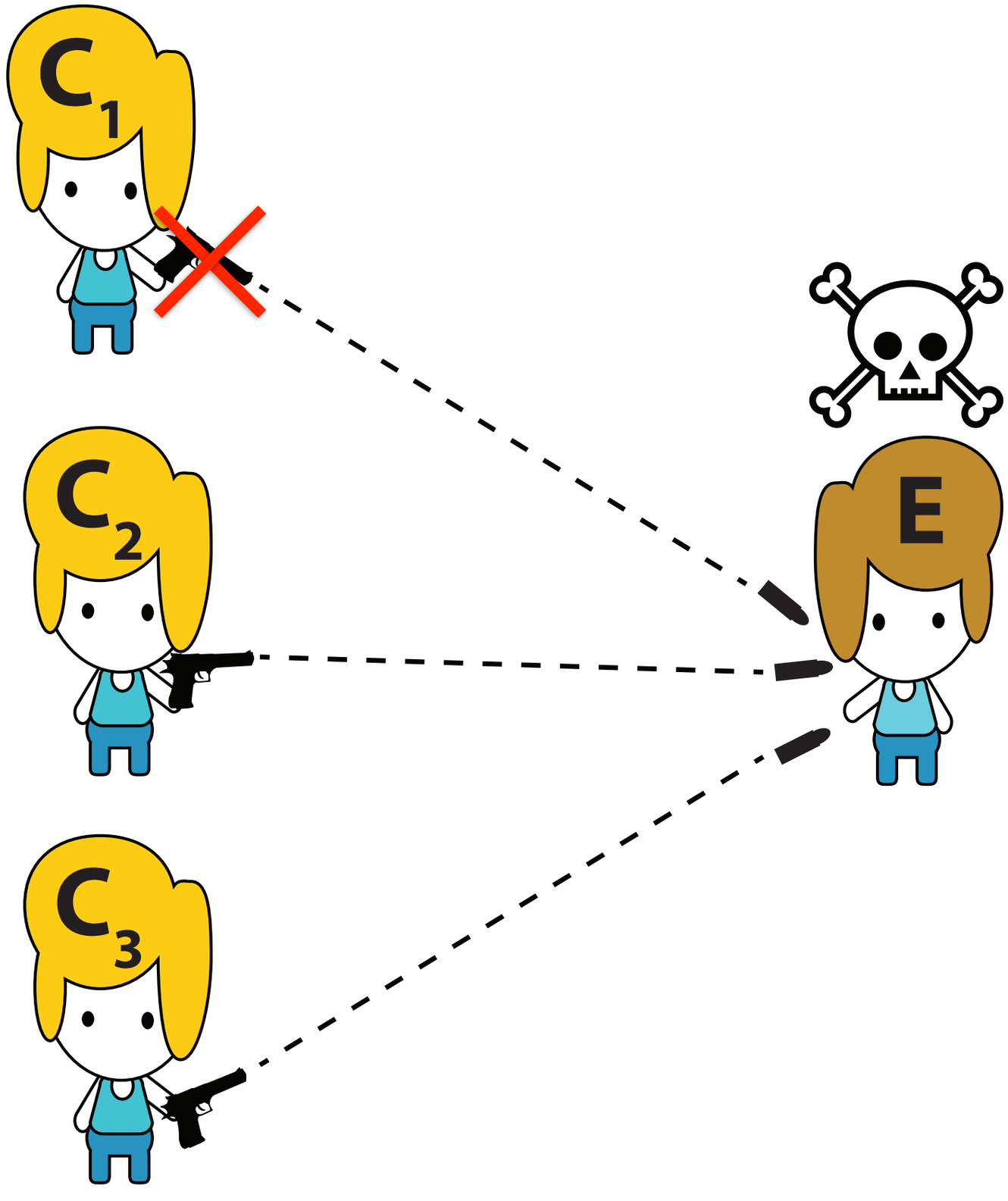
$$P(a|s_p, \mathcal{T} = \text{average}) = \frac{exp(\beta \cdot \hat{r}_a)}{\sum_{a \in \mathcal{A}} exp(\beta \cdot \hat{r}_a)}$$

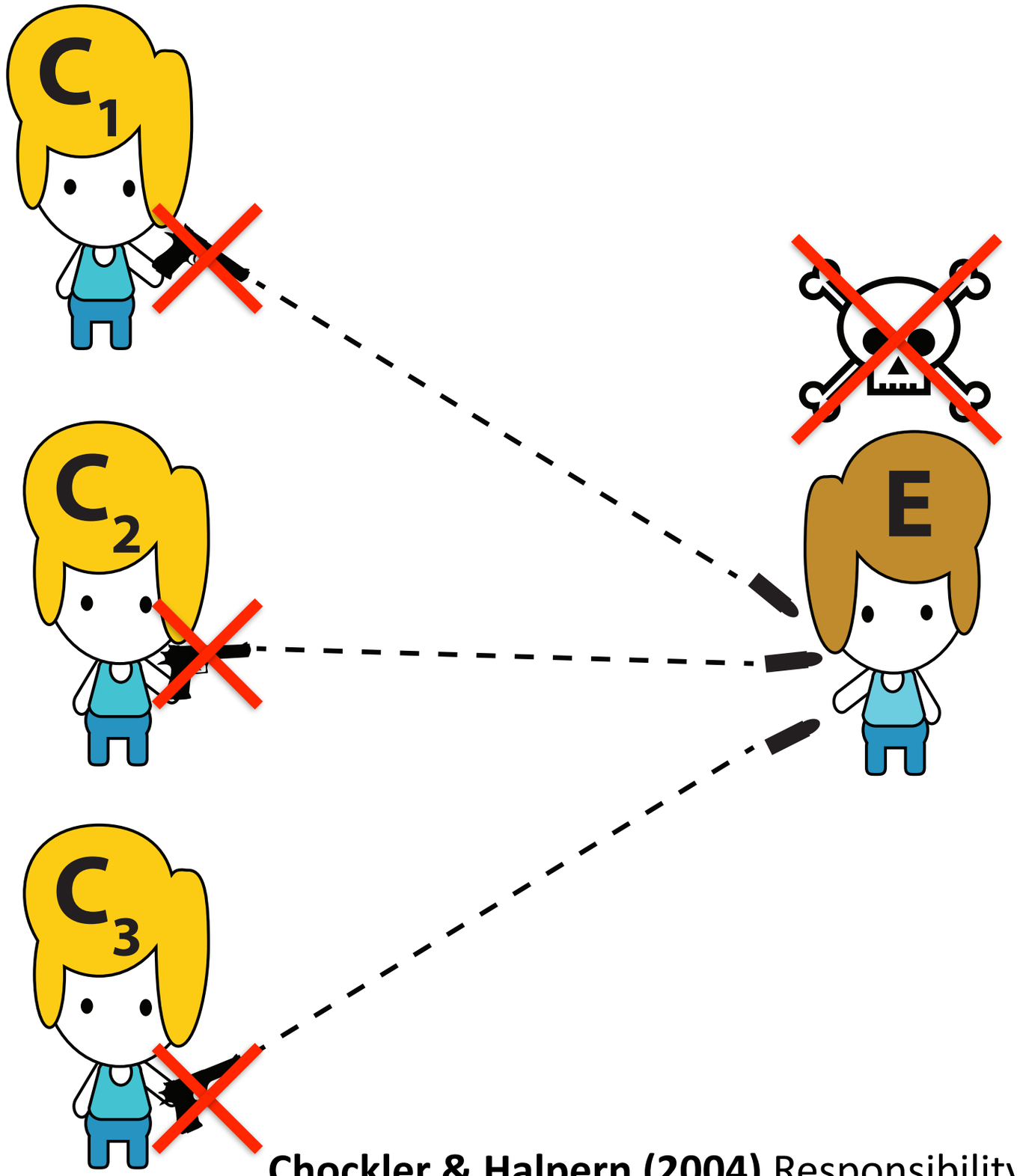
What would someone
else have done?



Action-centered
counterfactual

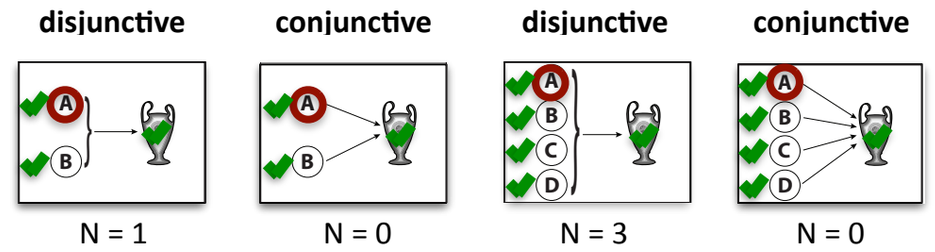


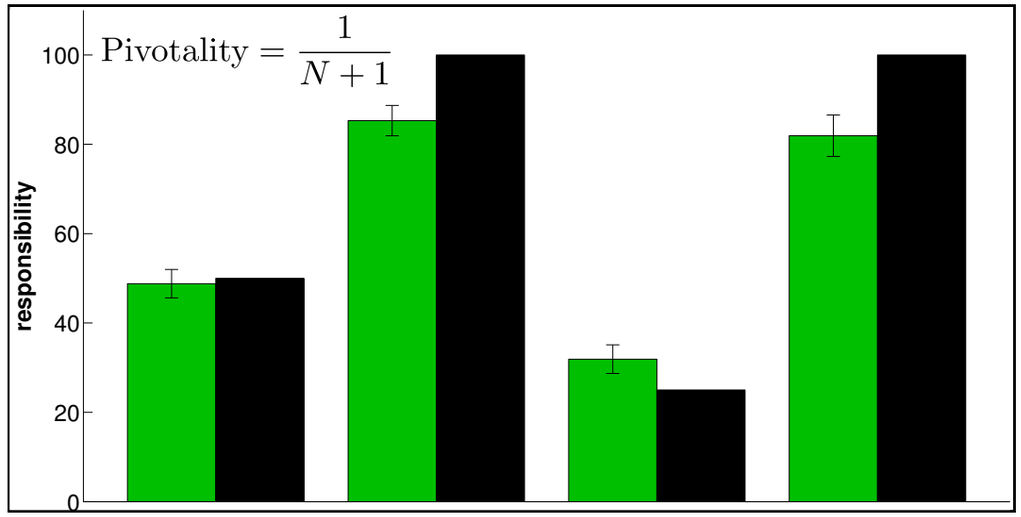

C1
C2
C3
E



$$\text{Pivotality} = \frac{1}{N+1}$$

N **= minimal number of interventions on the causal model to make the causal event under consideration pivotal**

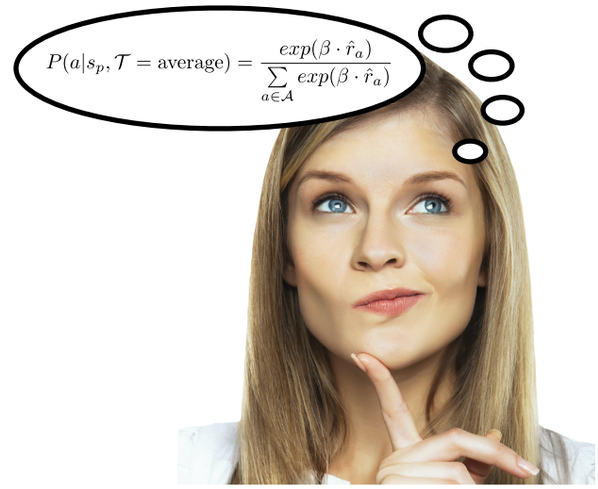


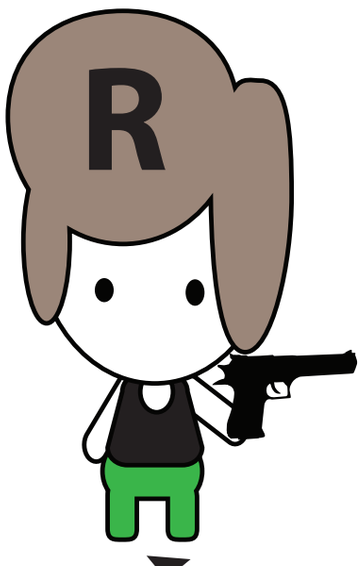


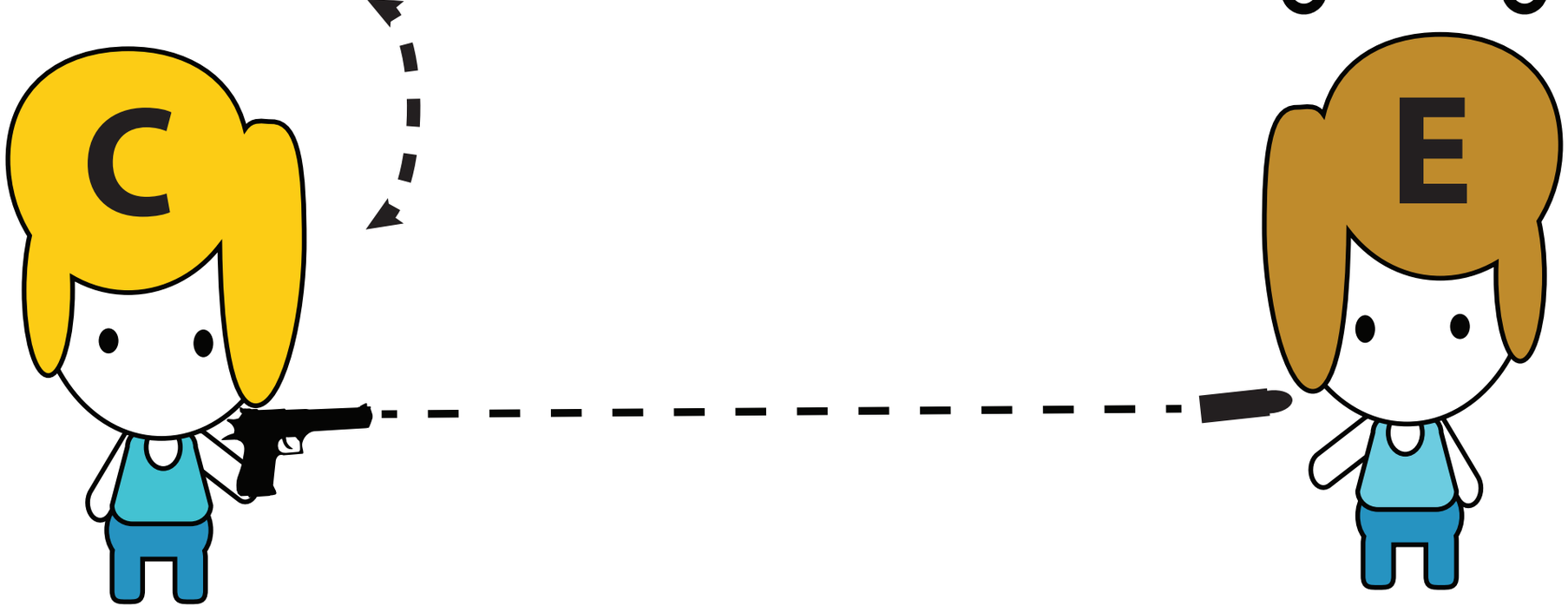
Chockler & Halpern (2004) Responsibility and Blame: A Structural-Model Approach

Gerstenberg & Lagnado (2012) Spreading the blame; **Zultan, Gerstenberg & Lagnado (2012)** Finding fault; **Lagnado, Gerstenberg & Zultan (2013)** Causal responsibility and counterfactuals

Person-centered
counterfactual







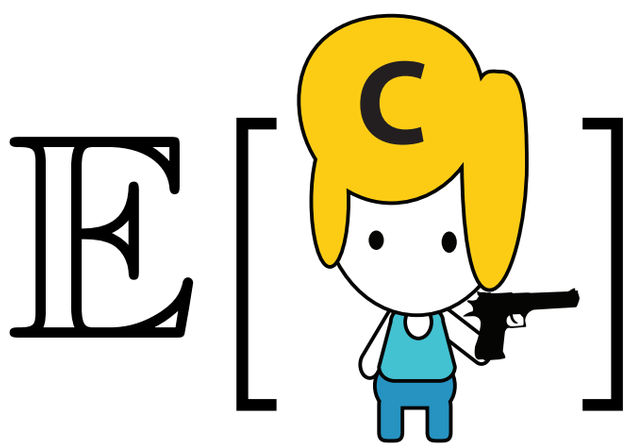
Wins above replacement

Lou Gehrig
(Gehrig smacked 493 homers)

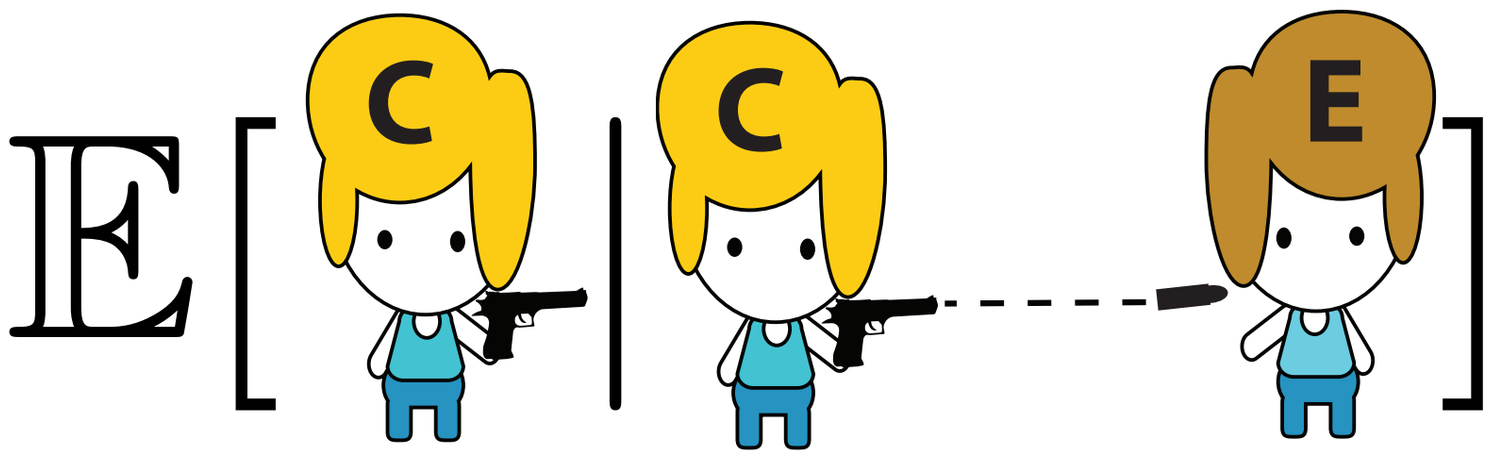
Babe Ruth
(Ruth hit 714 homers)

714 (Babe Ruth) —
493 (Lou Gehrig) = 221

Prior expectation
about future behavior



Posterior expectation
about future behavior



Change in expectation
about future behavior

Everything else

Expectations

Responsibility

Who was responsible?

"The answer we choose will depend on what we take to be the normal, proper, or expected course of events; the person that we hold responsible is the one **who steps outside this expected pattern**."

Mackie (1955) Responsibility and Language

of research

Door A

50%

Door B

10%

Johnson & Rips (2015) Do the right thing

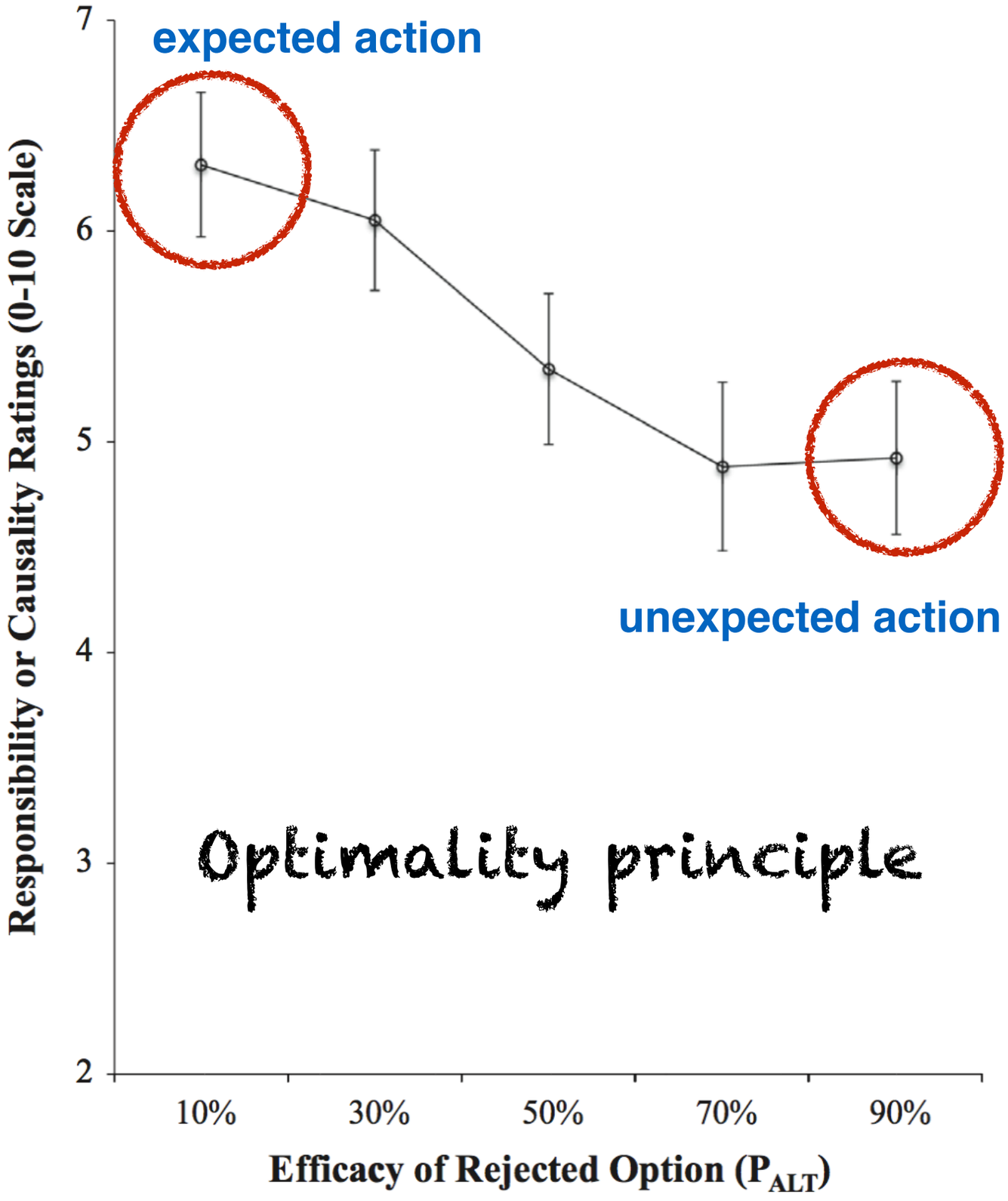
Door A

50%

Door B

90%

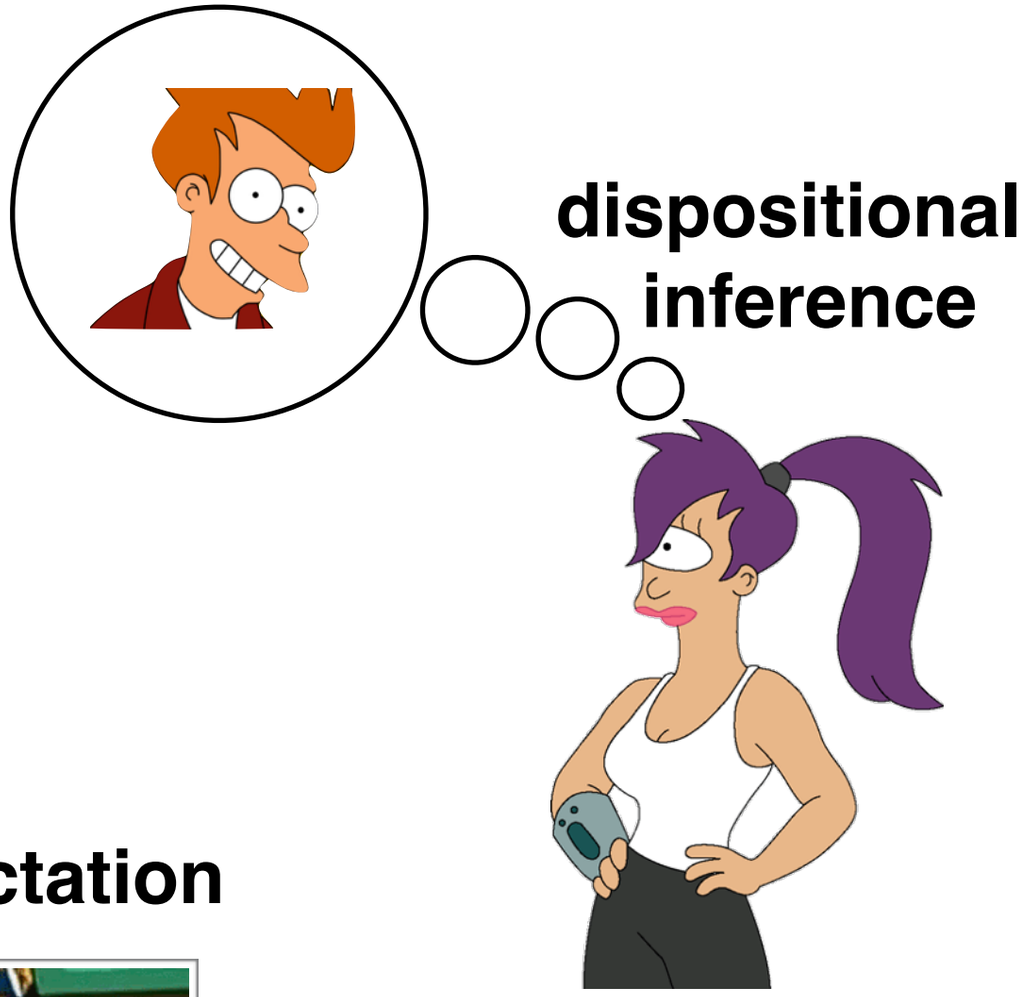
Johnson & Rips (2015) Do the right thing



Johnson & Rips (2015) Do the right thing

action expectation

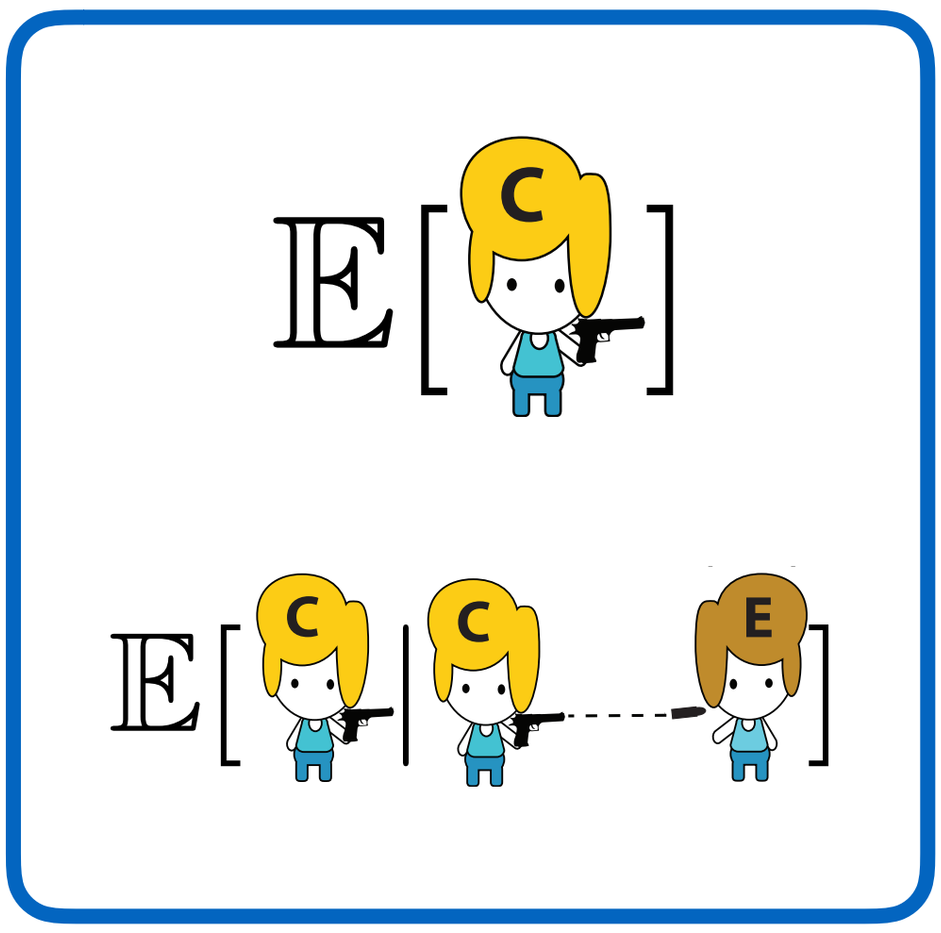
responsibility



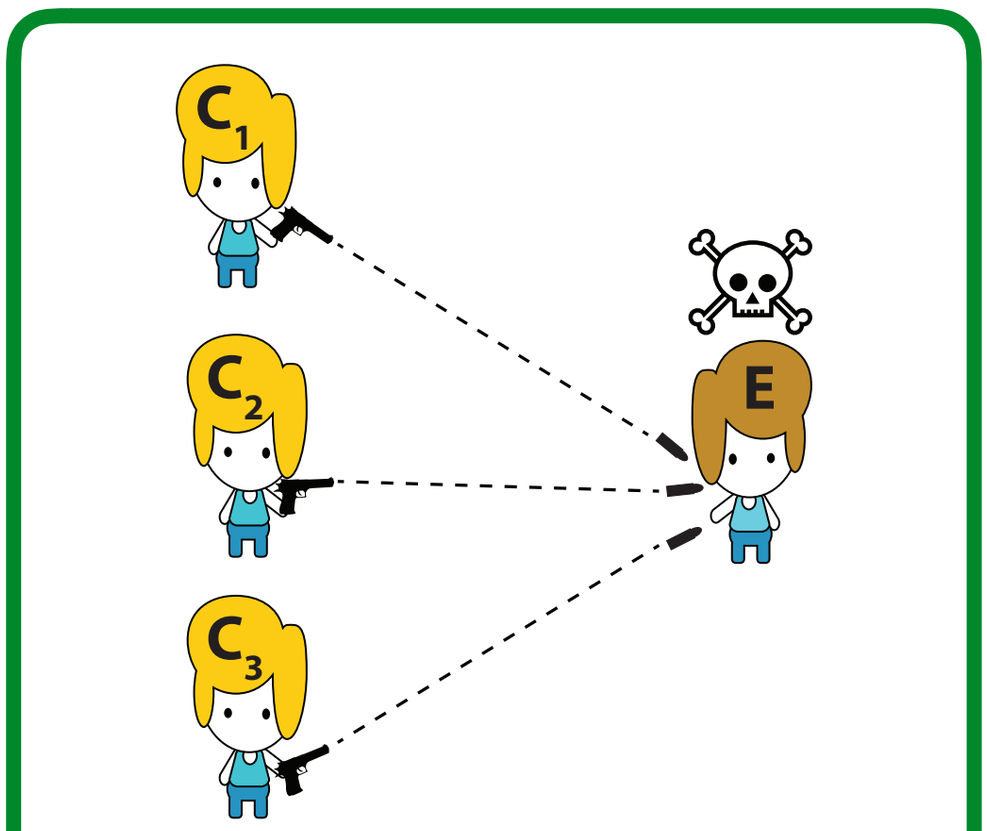
action expectation

responsibility

Responsibility = f(Difference in Expectation , Pivotality)

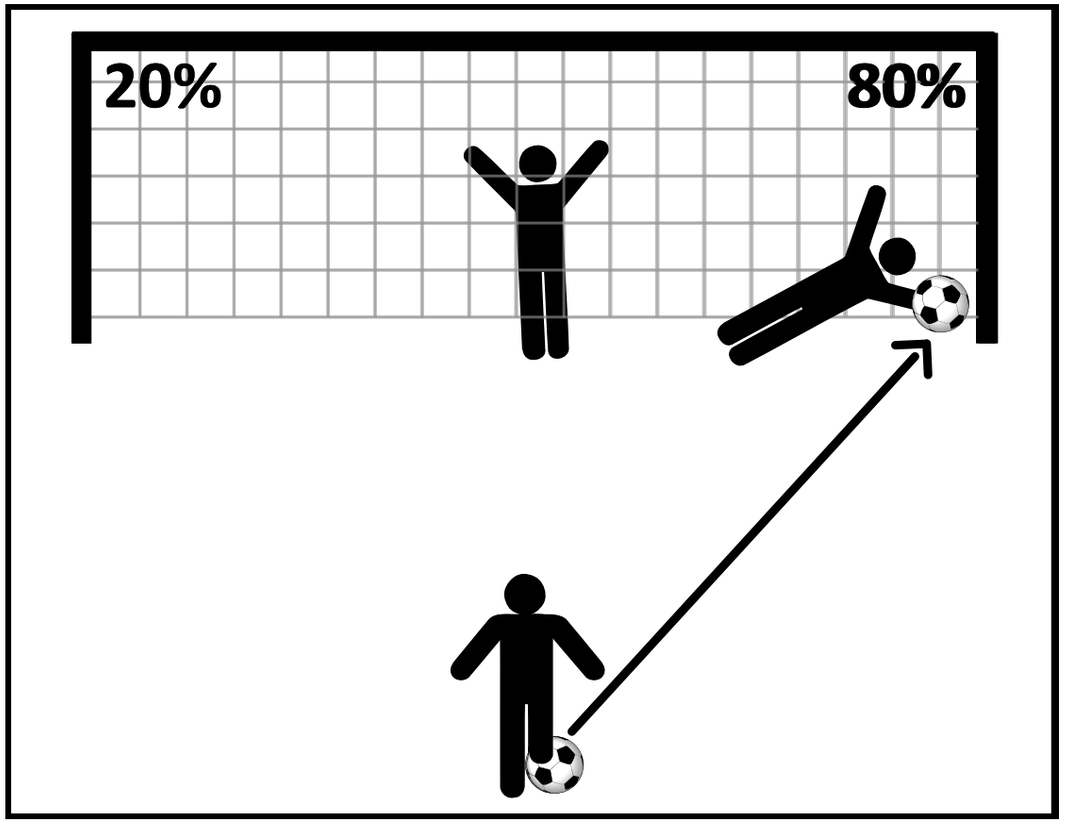


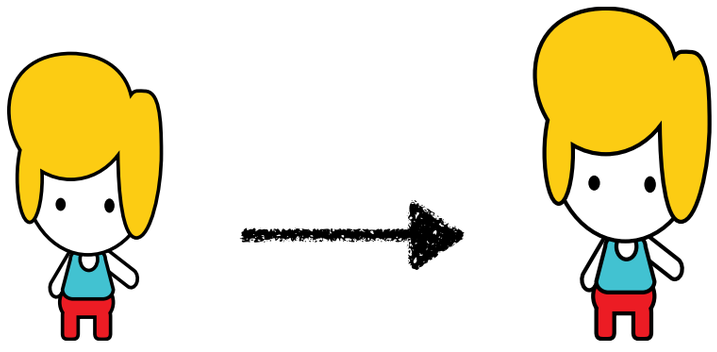
Person-centered
counterfactual



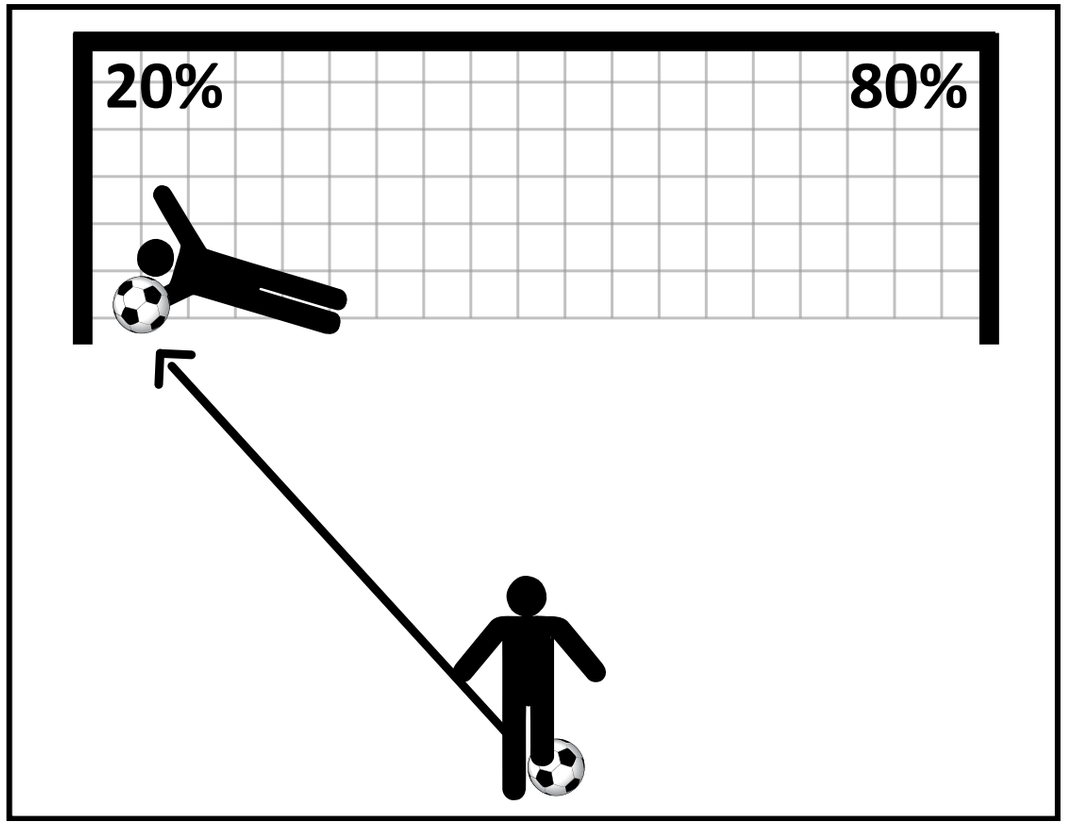
Action-centered
counterfactual

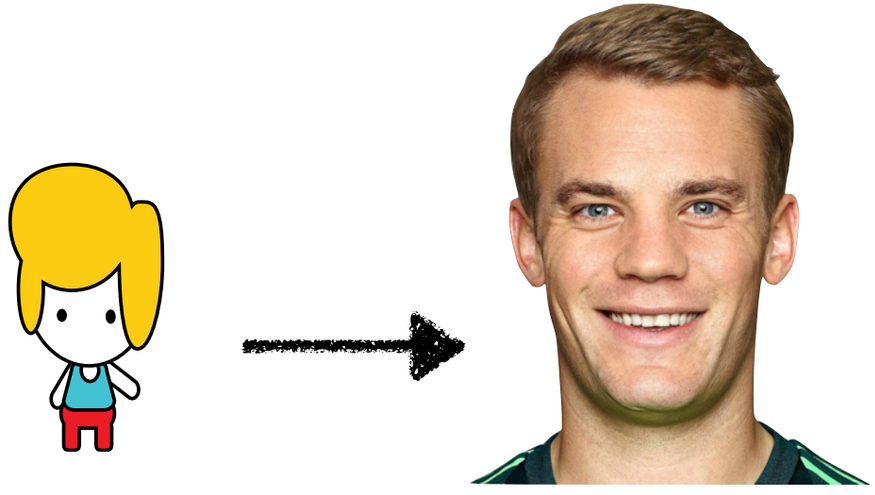
expected



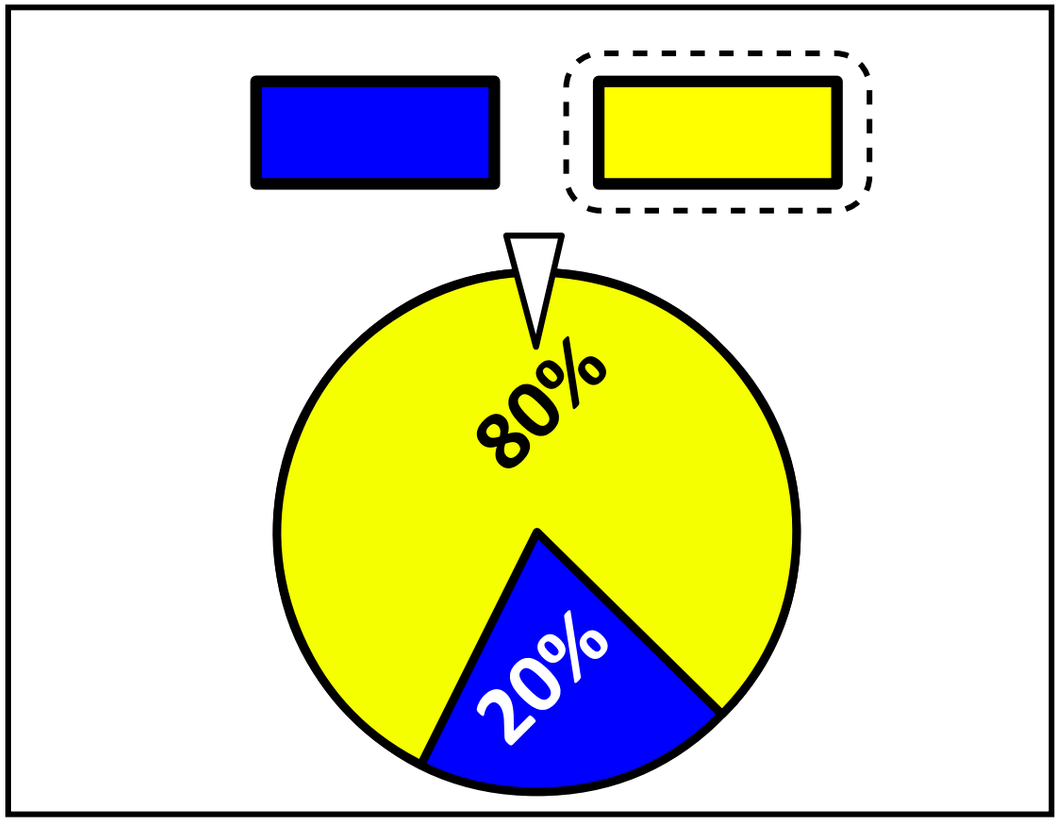


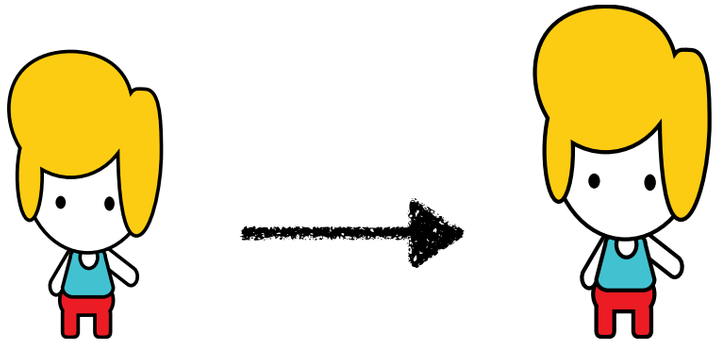
unexpected



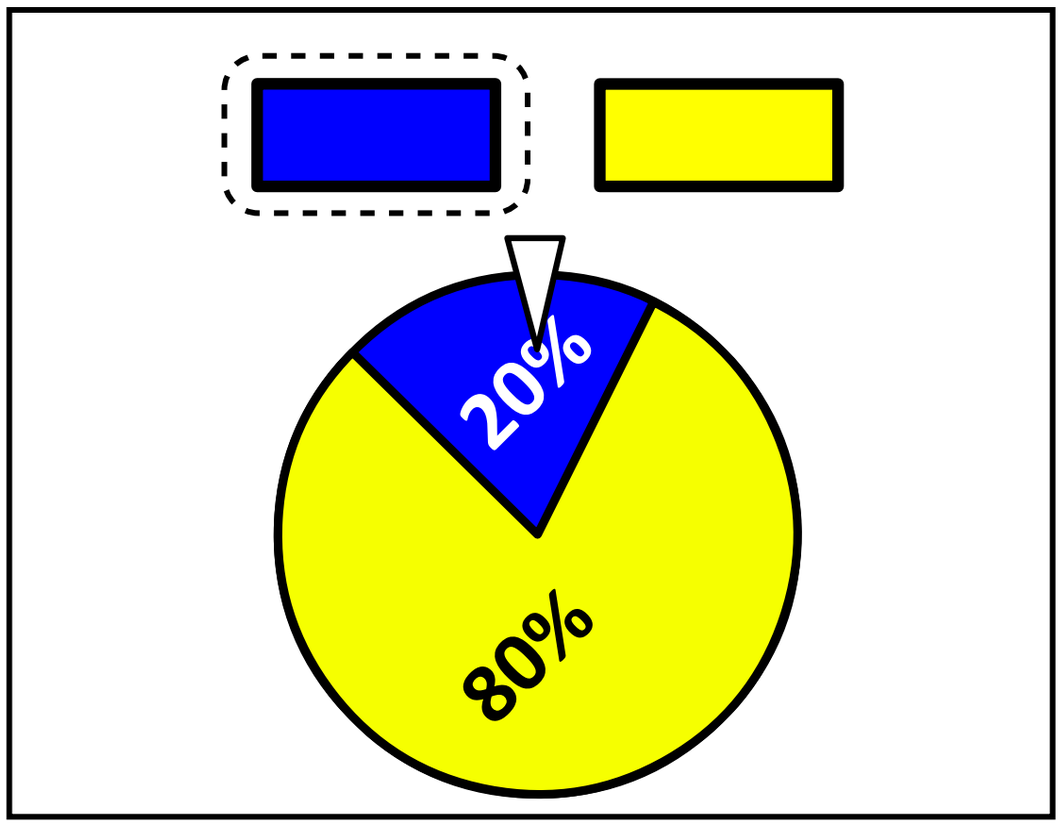


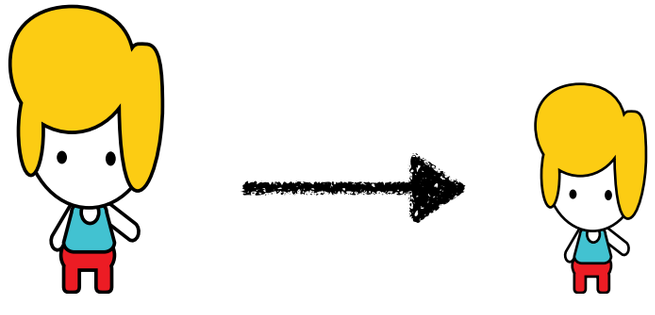
expected



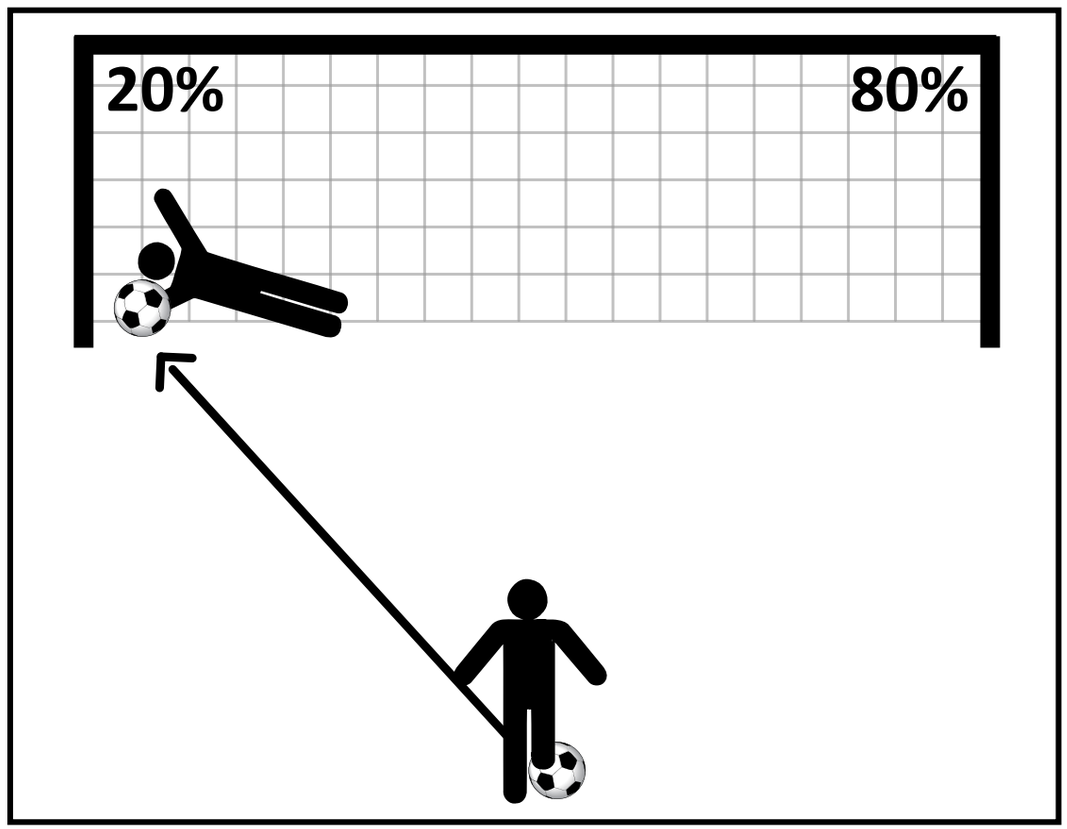


unexpected

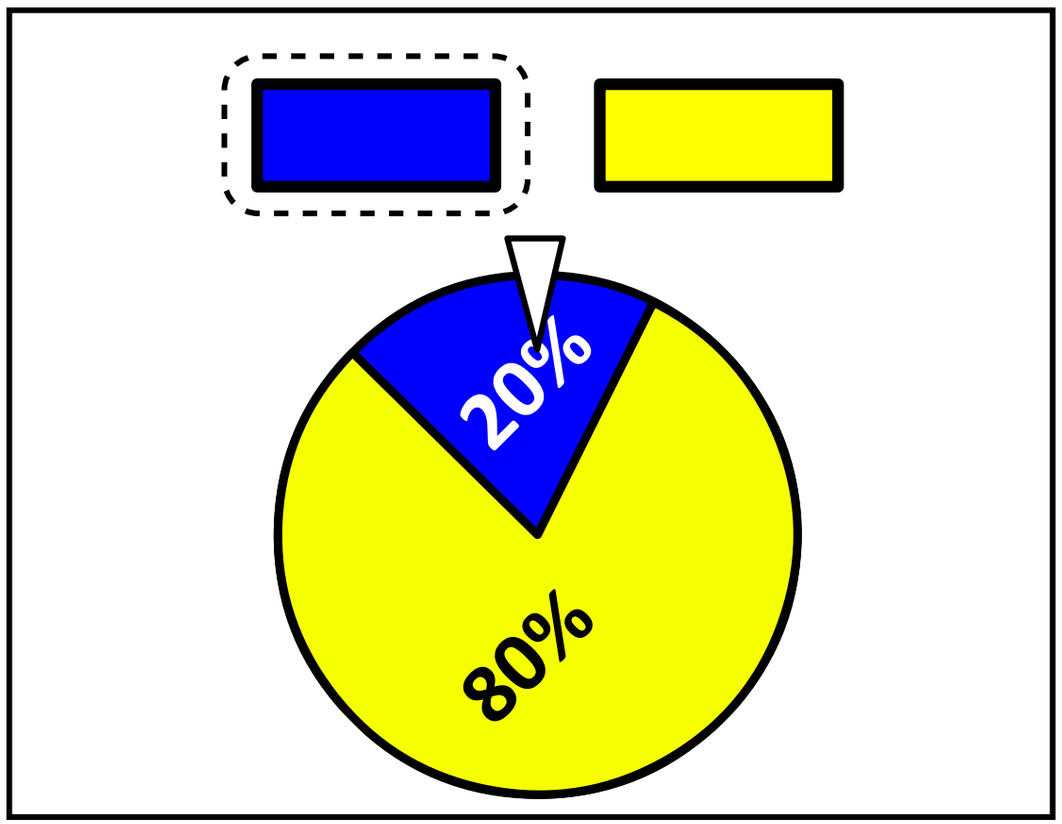


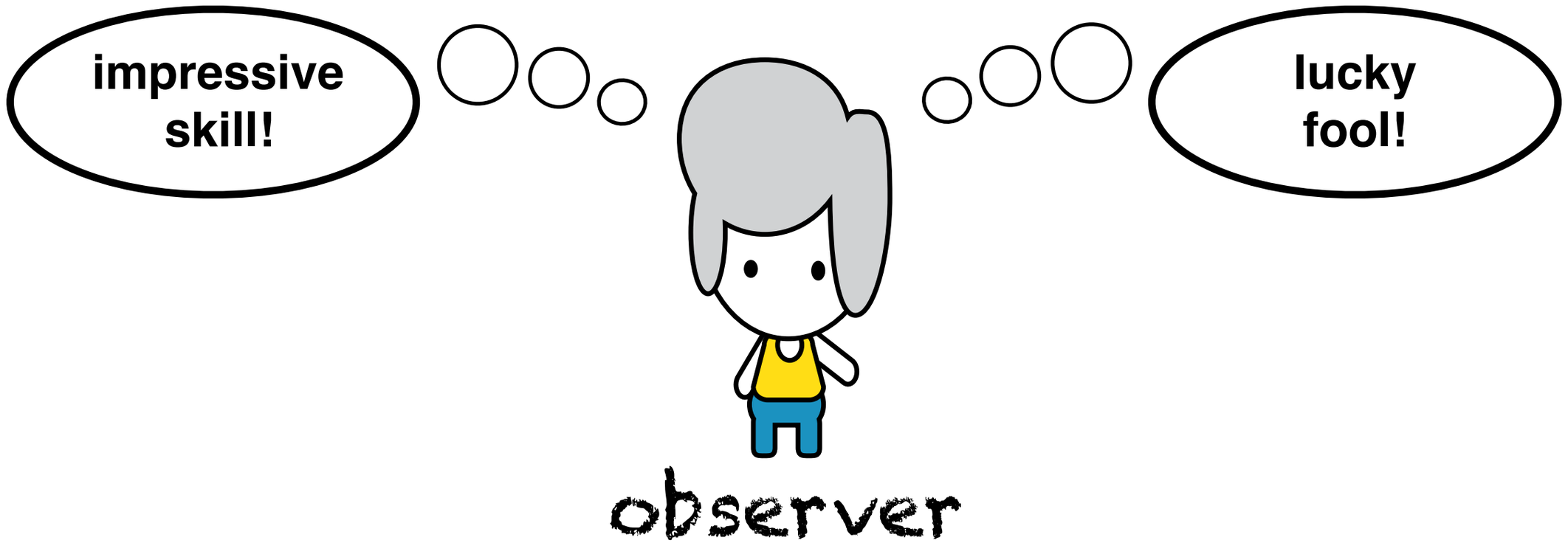


unexpected

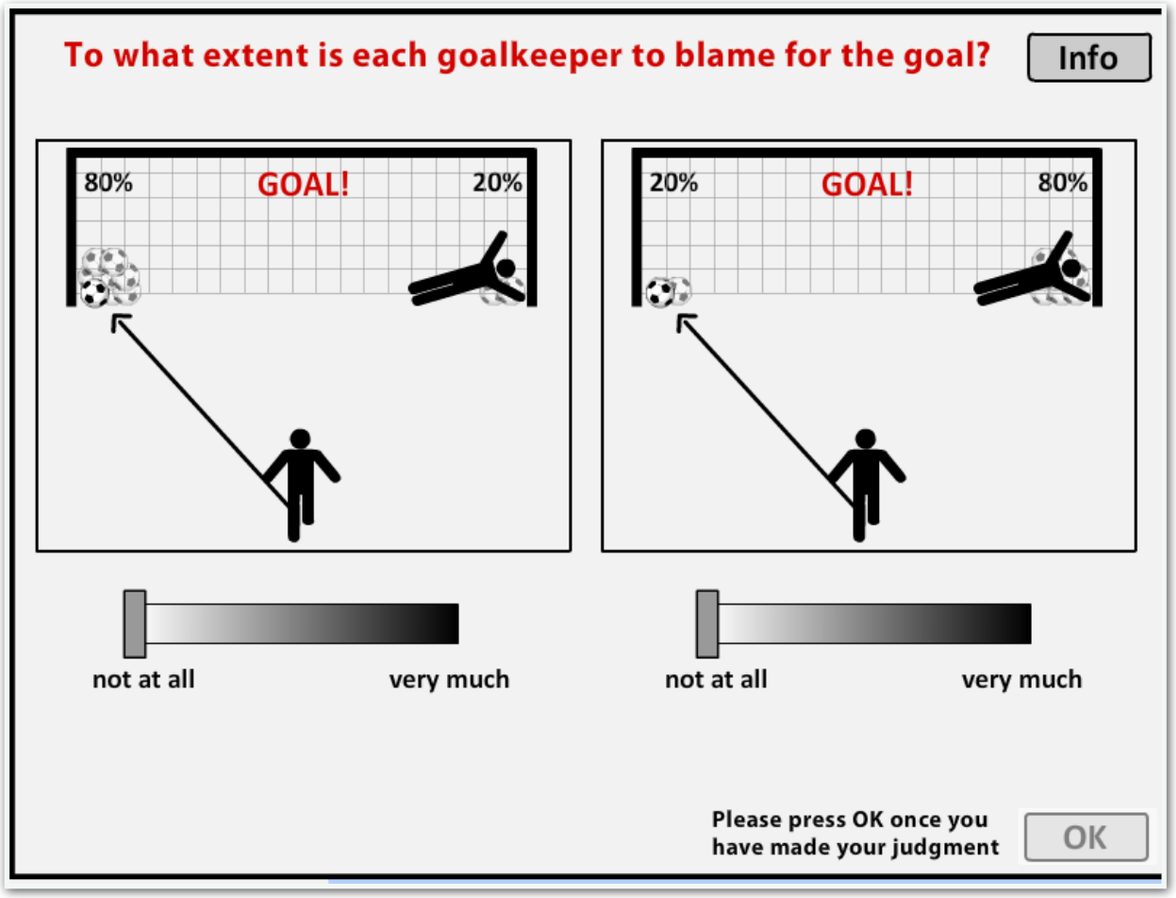


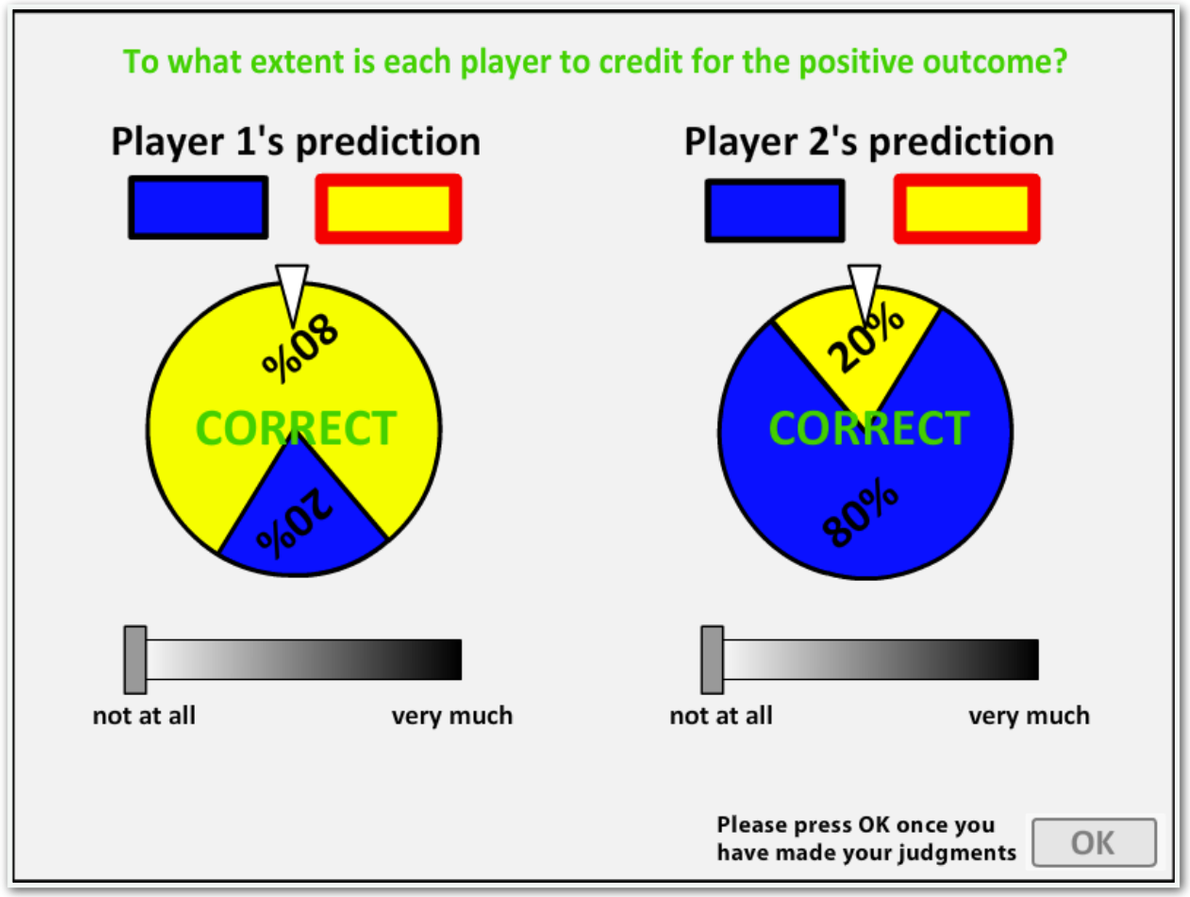
unexpected

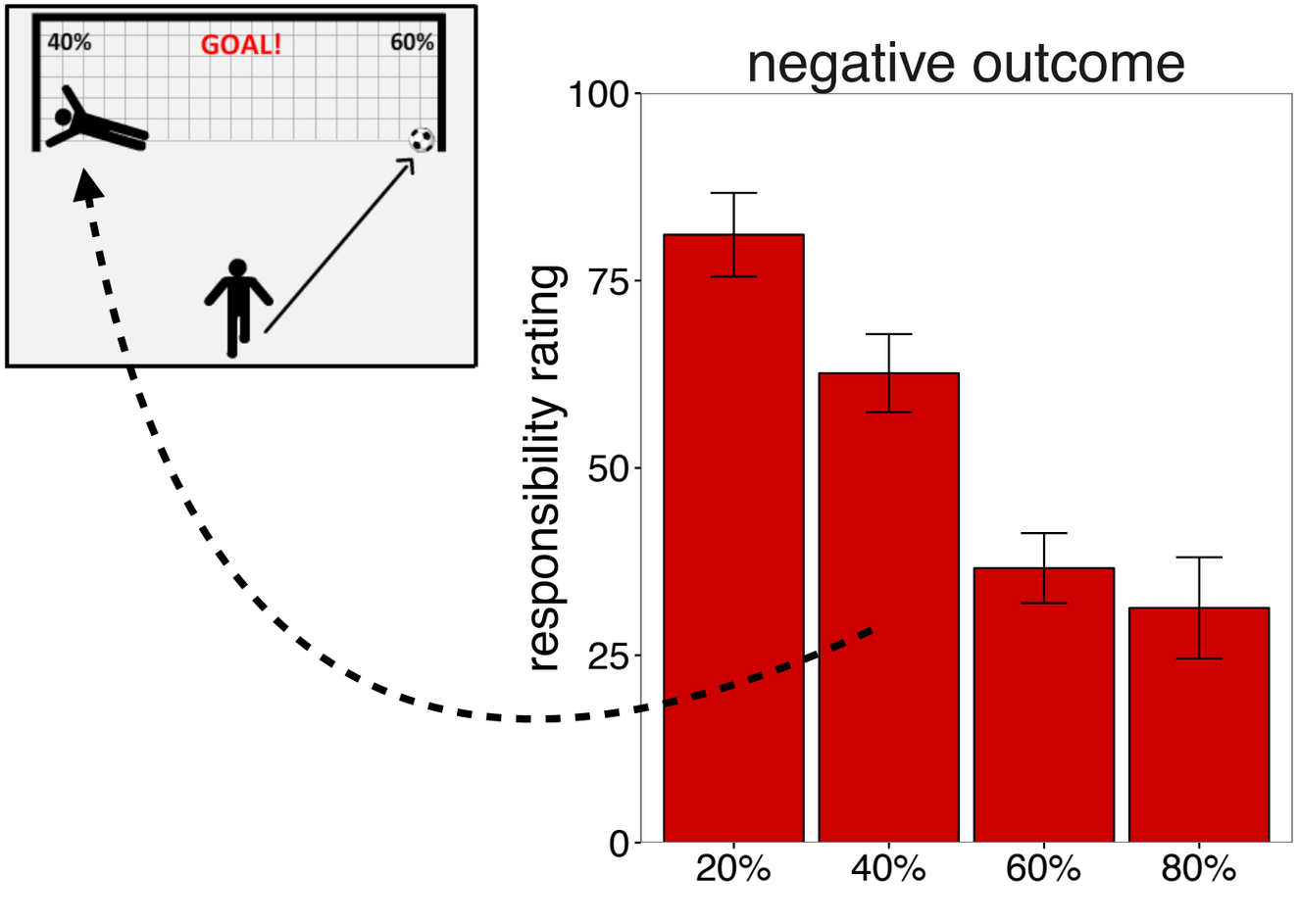


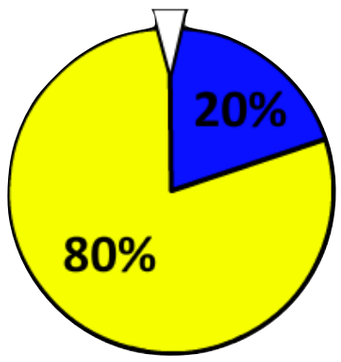


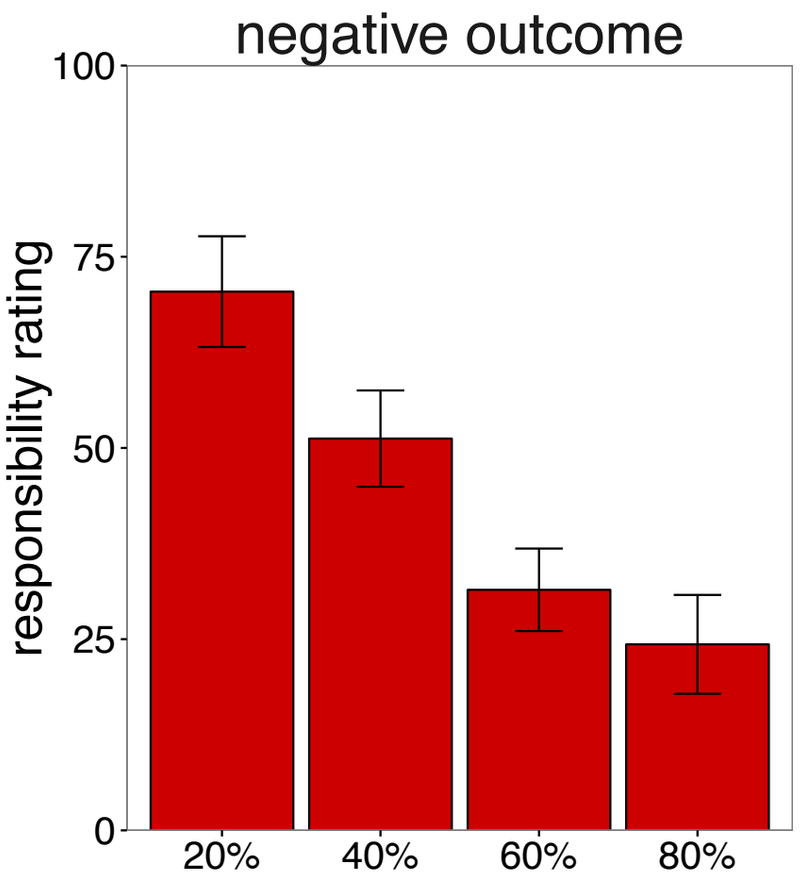
Experiment 1

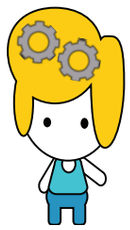




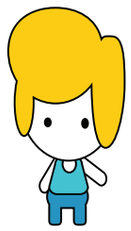

40%
GOAL!
60%
negative outcome
responsibility rating
100
75
50
25
0
20%
40%
60%
80%


20%
80%

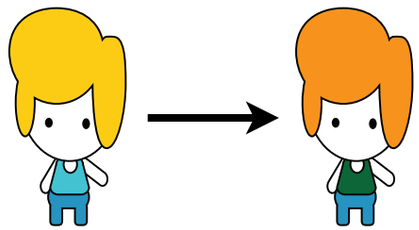

negative outcome
responsibility rating
100
75
50
25
0
20%
40%
60%
80%



- intuitive theory of how people work

E()

- expectations



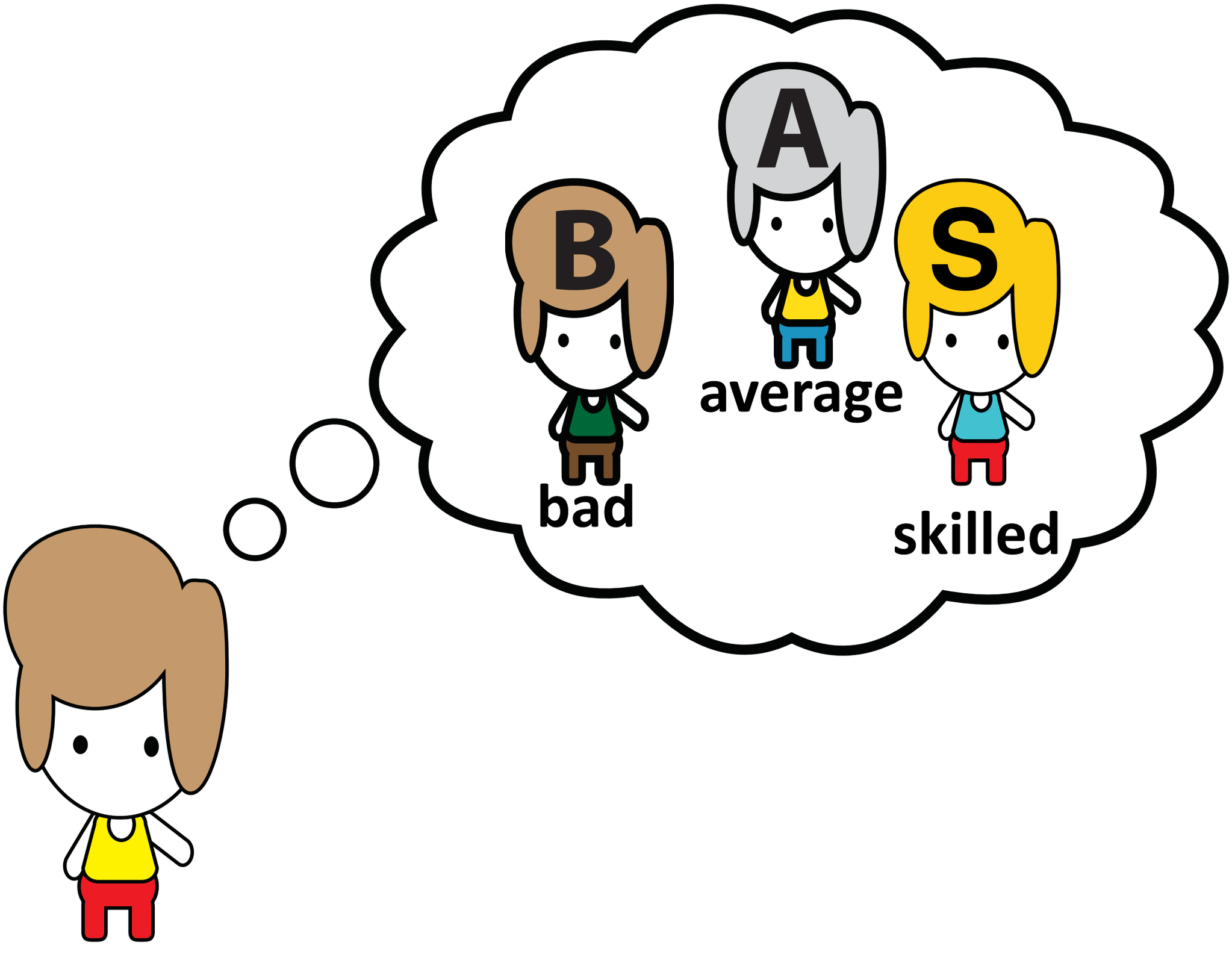
- dispositional inference

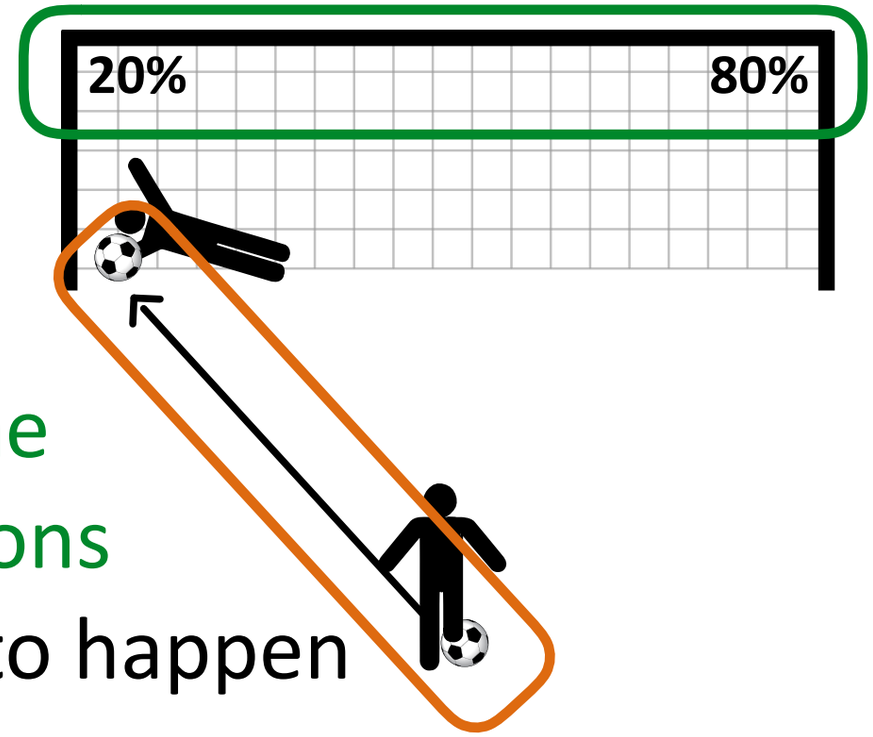
E()>E()

- blame for negative outcomes if expectations decreased

E()<E()

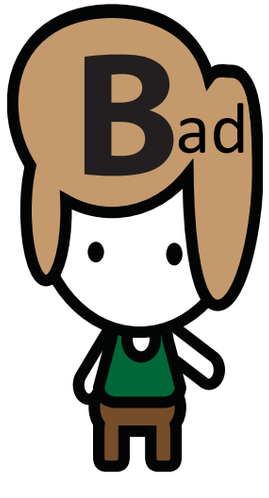
- credit for positive outcomes if expectations increased


B
bad
A
average
S
skilled

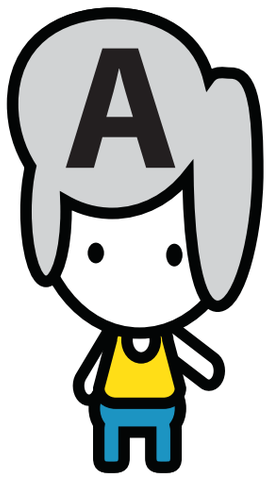


- makes his decisions based on the probability of the different options
- cannot anticipate what's going to happen

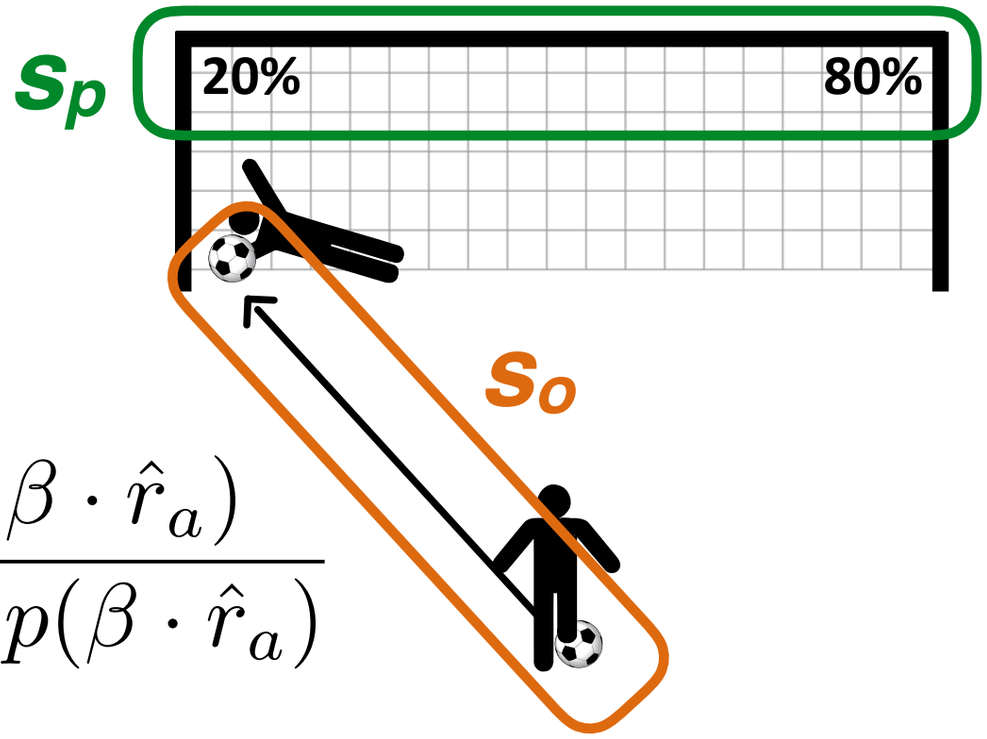
- can anticipate what's going to happen
- makes his decisions based on that knowledge

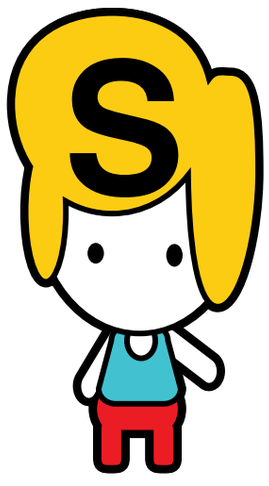


- falsely believes that he can anticipate what's going to happen
- is more likely to do the wrong thing rather than the right thing

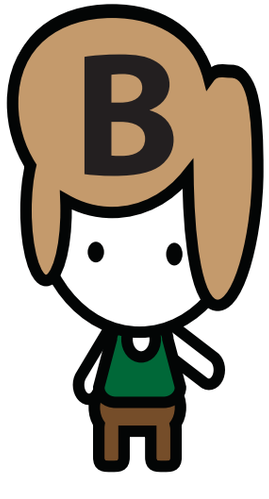


$$P(a|s_p, \mathcal{T} = \text{average}) = \frac{exp(\beta \cdot \hat{r}_a)}{\sum_{a \in \mathcal{A}} exp(\beta \cdot \hat{r}_a)}$$

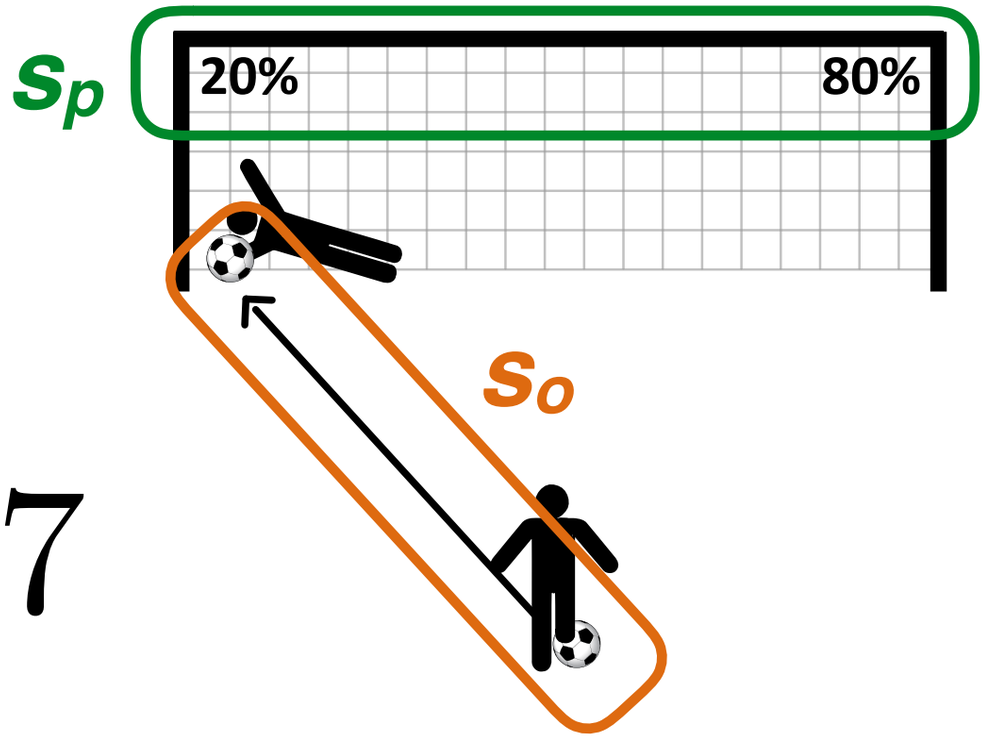


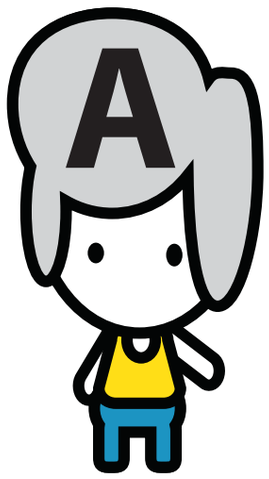


$$P(a|s_o, \mathcal{T} = \text{skilled}) = \frac{exp(\beta \cdot r_a^{\text{true}})}{\sum_{a \in \mathcal{A}} exp(\beta \cdot r_a^{\text{true}})}$$



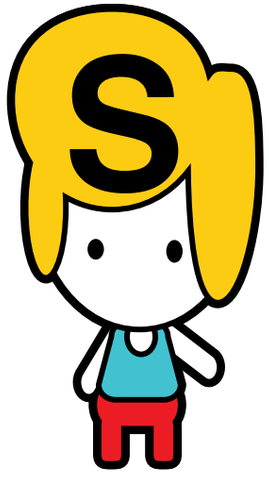
$$P(a|s_o, \mathcal{T} = \text{bad}) = \frac{exp(\beta \cdot (-r_a^{\text{true}}))}{\sum_{a \in \mathcal{A}} exp(\beta \cdot (-r_a^{\text{true}}))}$$



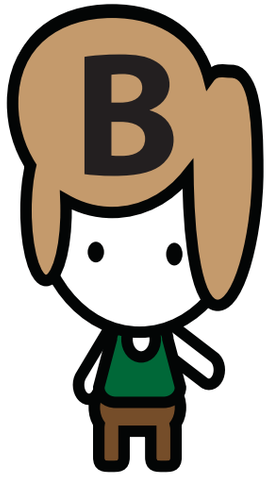


goalie world or **spinner** world

$$\mathbb{E}[r|w, \text{A}] = 0.7$$



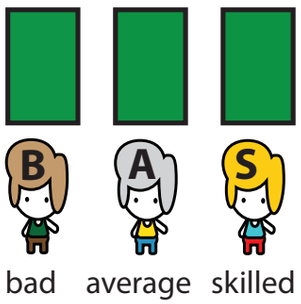
$$\mathbb{E}[r|w, \text{S}] = 0.9$$



$$\mathbb{E}[r|w, \text{B}] = 0.3$$

Dispositional inference

Prior



$$\mathbb{E}[r|w, \text{B}] = 0.3$$

$$\mathbb{E}[r|w, \text{A}] = 0.7$$

$$\mathbb{E}[r|w, \text{S}] = 0.9$$

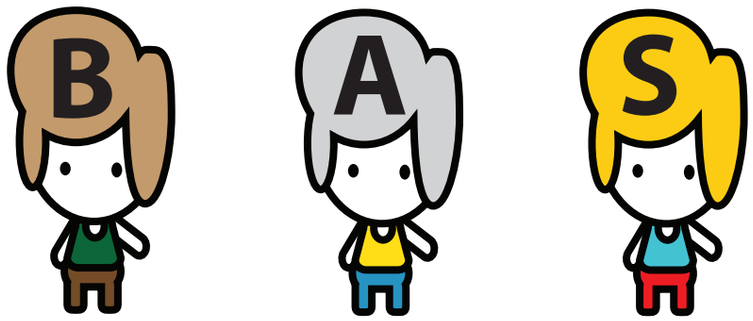
Difference in expectation

$$\mathbb{E}[r|\mathcal{W} = w] = \frac{1}{3} \times 0.3 + \frac{1}{3} \times 0.7 + \frac{1}{3} \times 0.9 \approx 0.63$$

$$\mathbb{E}[r|\mathcal{W} = w, s_{\text{obs}}, a_{\text{obs}}] = \frac{1}{9} \times 0.3 + \frac{2}{9} \times 0.7 + \frac{6}{9} \times 0.9 \approx 0.79$$

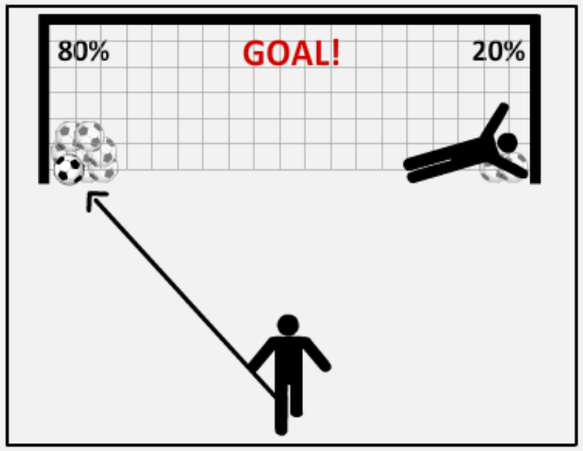
$$\text{Difference in expected reward} = \mathbb{E}[r|\mathcal{W} = w, s_{\text{obs}}, a_{\text{obs}}] - \mathbb{E}[r|\mathcal{W} = w]$$

Experiment 2



To what extent is each goalkeeper to blame for the goal?

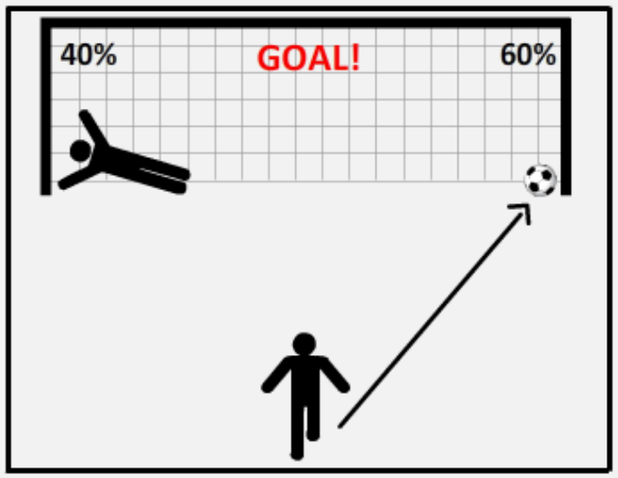
Info



not at all very much

What type of goalkeeper do you think this goalkeeper is?

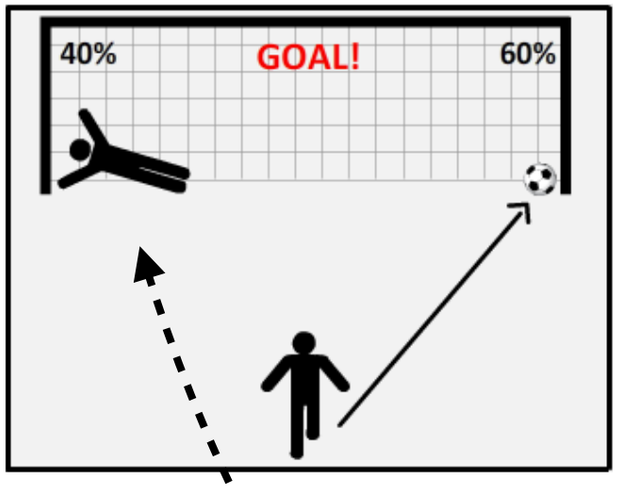
(please indicate your judgment by typing in the text boxes below, make sure that your three judgments sum up to 100%)

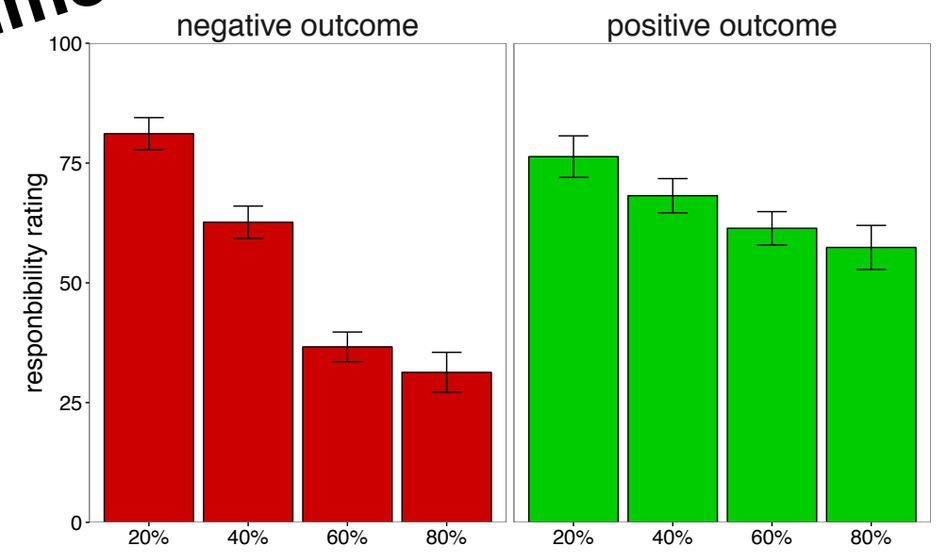


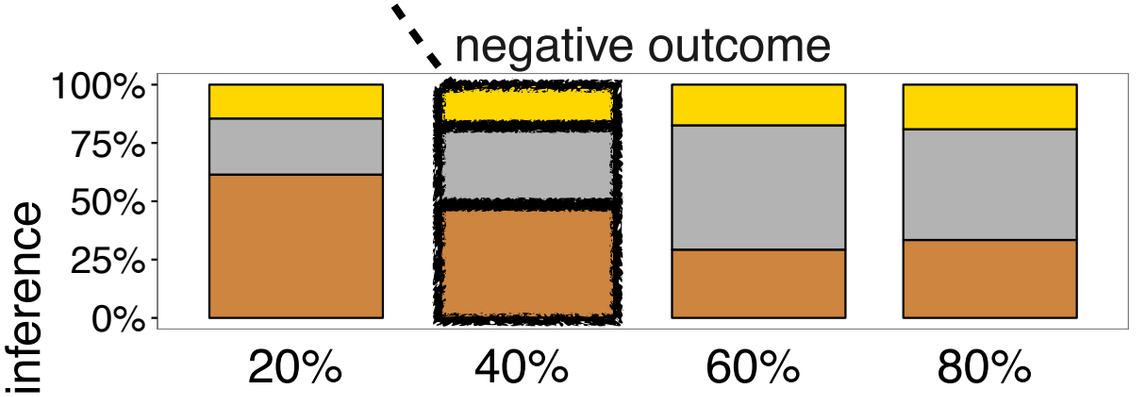
bad goalkeeper	33	%
average goalkeeper	33	%
skilled goalkeeper	34	%

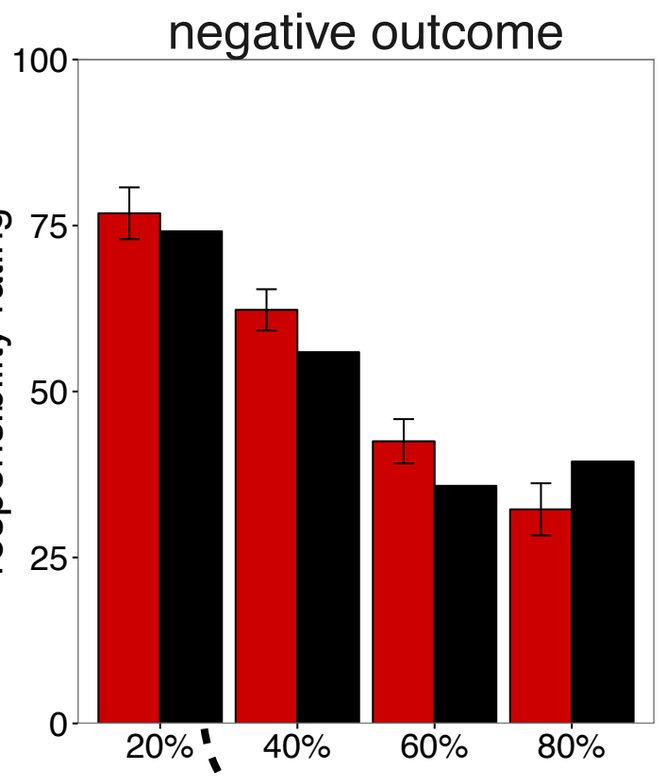
Please press OK after you have made your judgments. OK

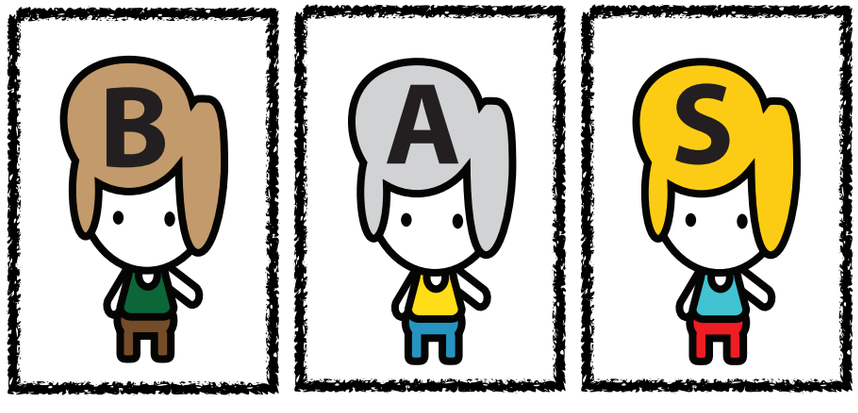
Experiment 1



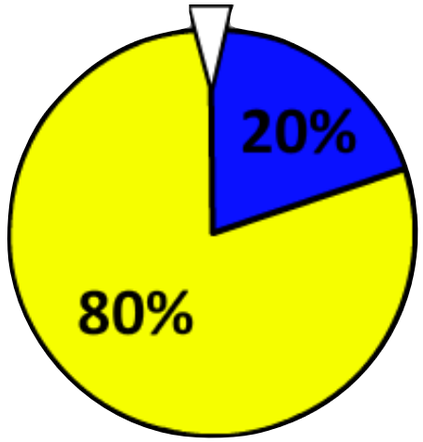




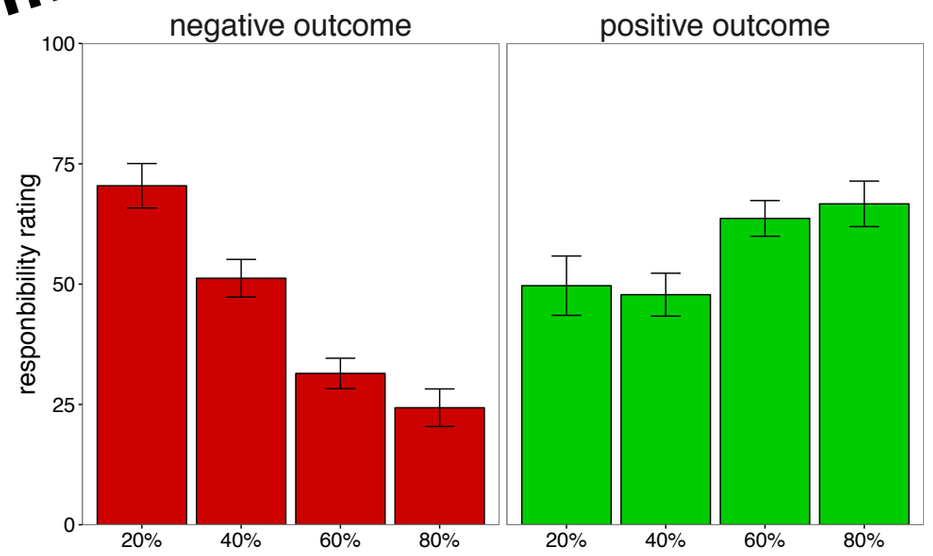


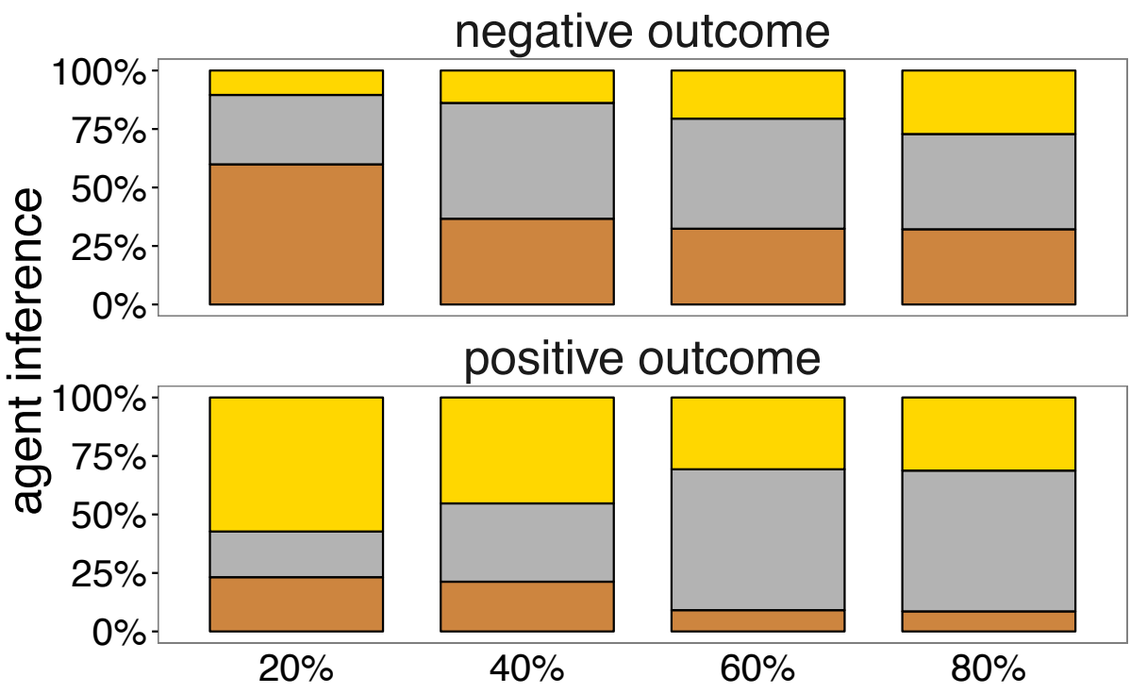


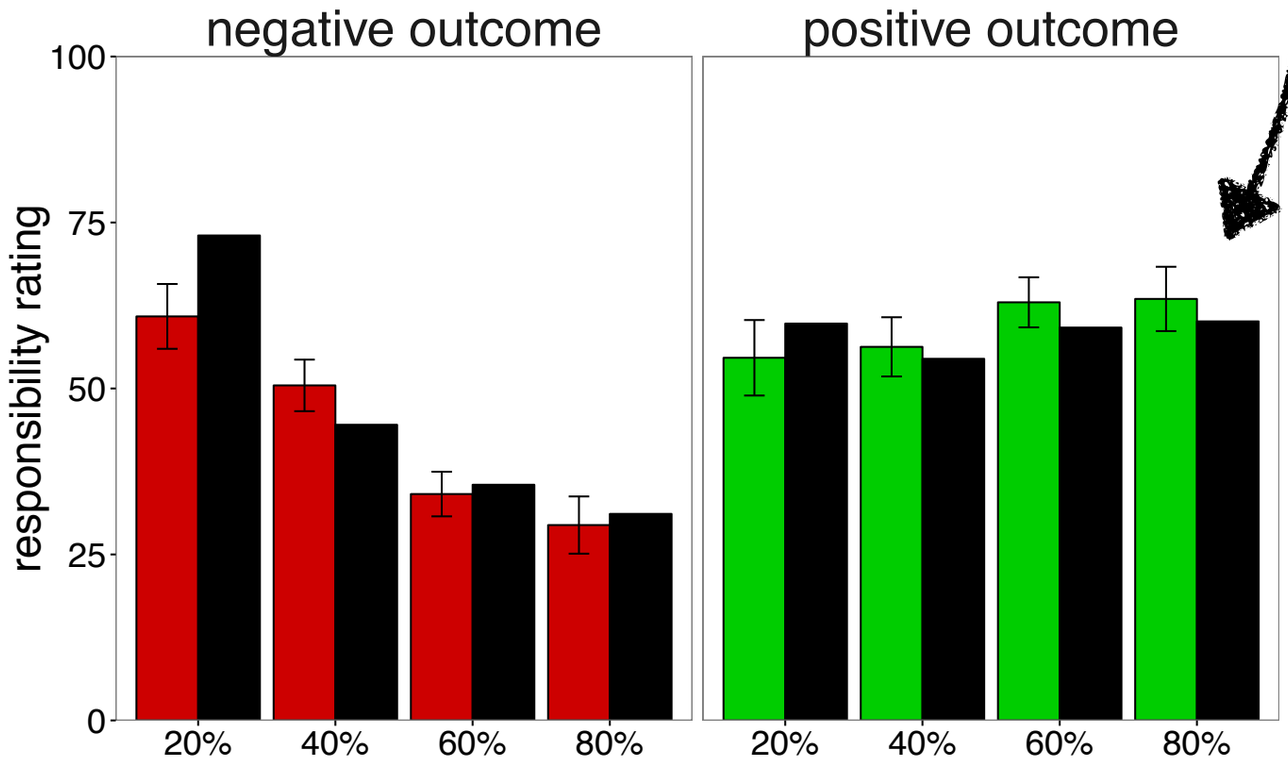
$$\mathbb{E}[r|\mathcal{W} = w, s_{\text{obs}}, a_{\text{obs}}] - \mathbb{E}[r|\mathcal{W} = w]$$



Experiment 1

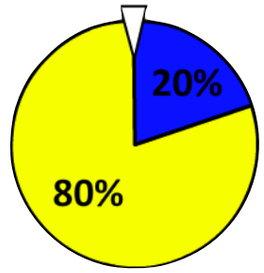




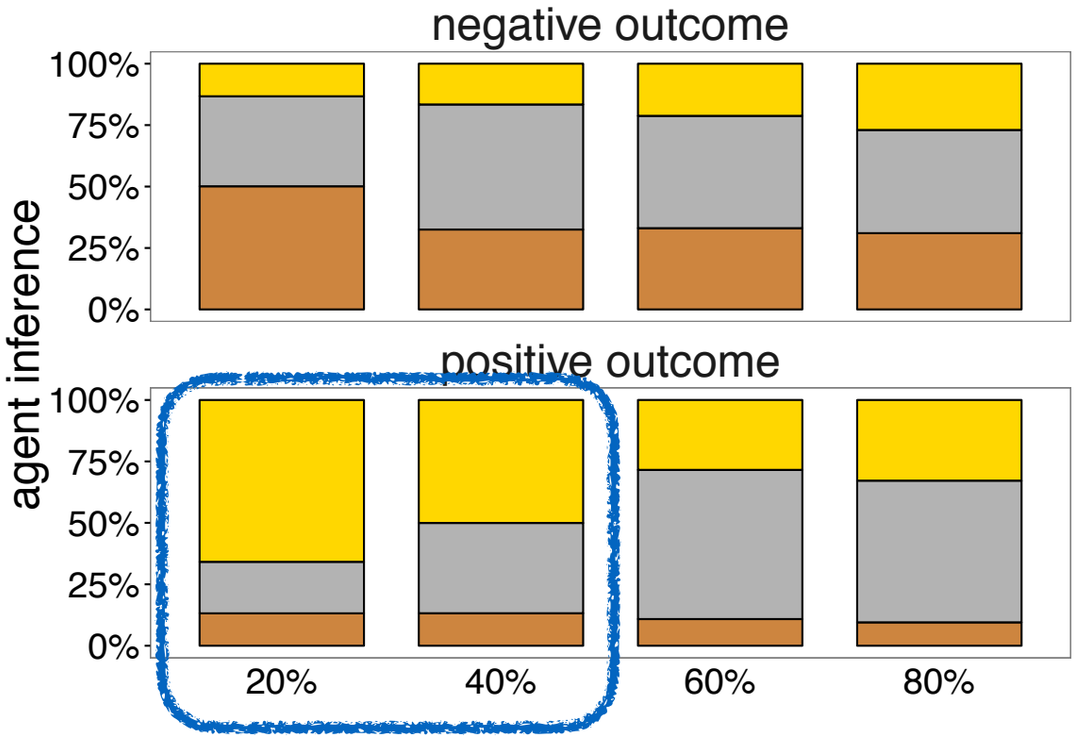


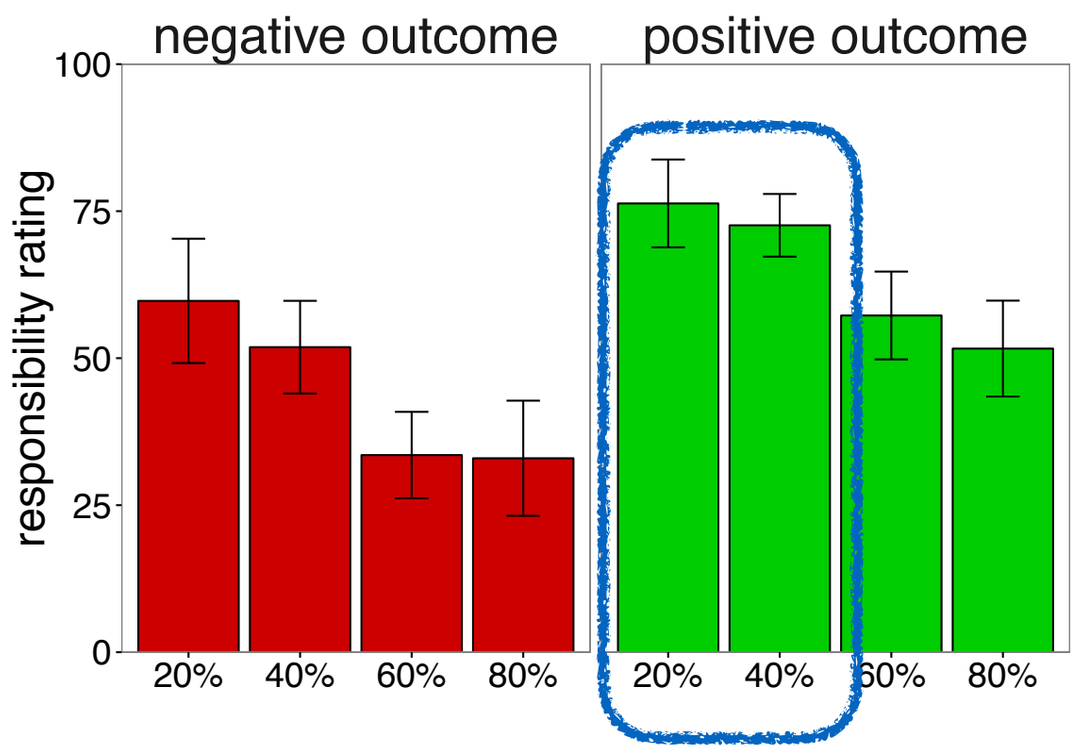
bad **average** **skilled**

Cluster differences

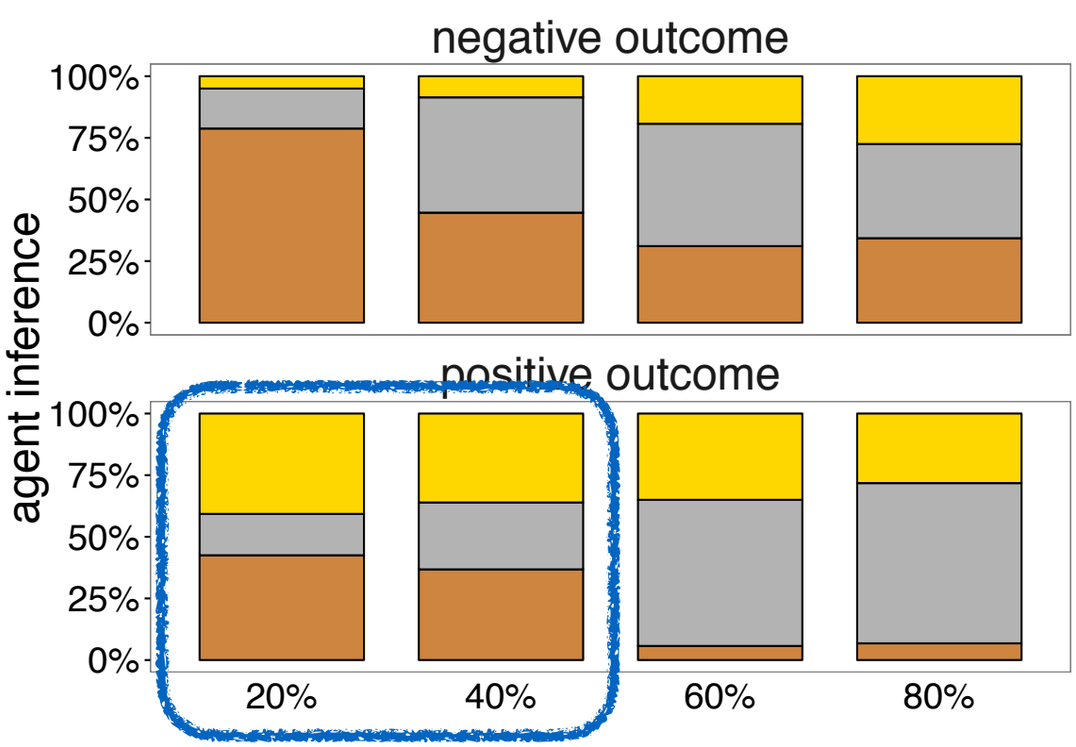


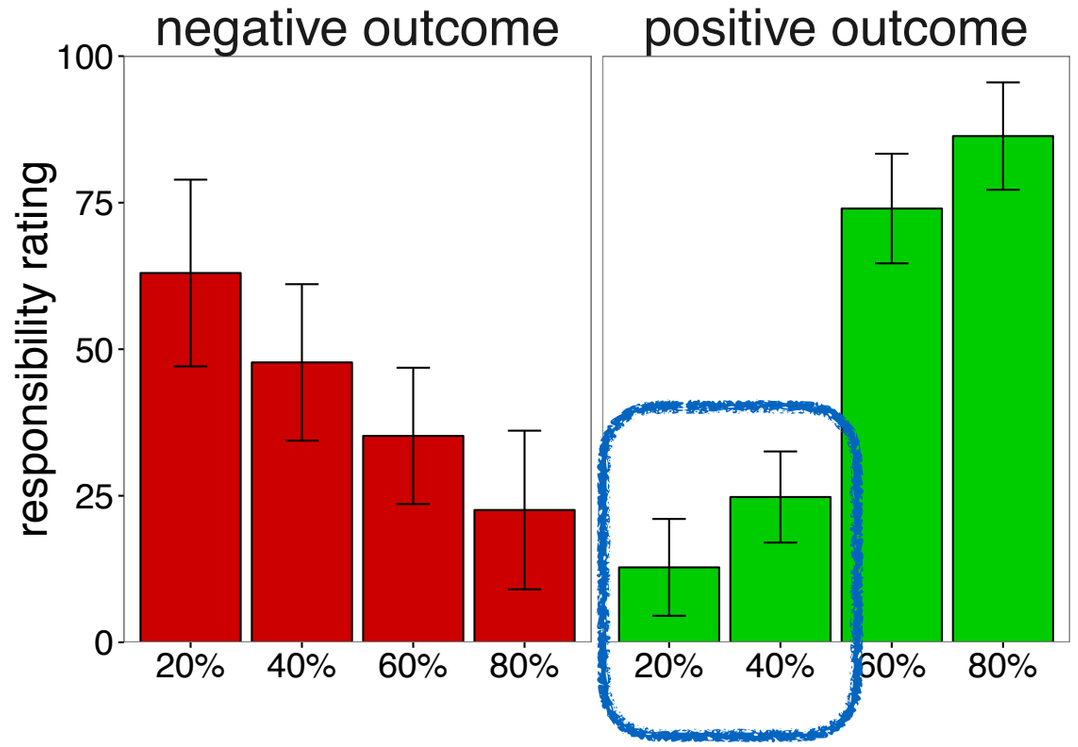
N = 27



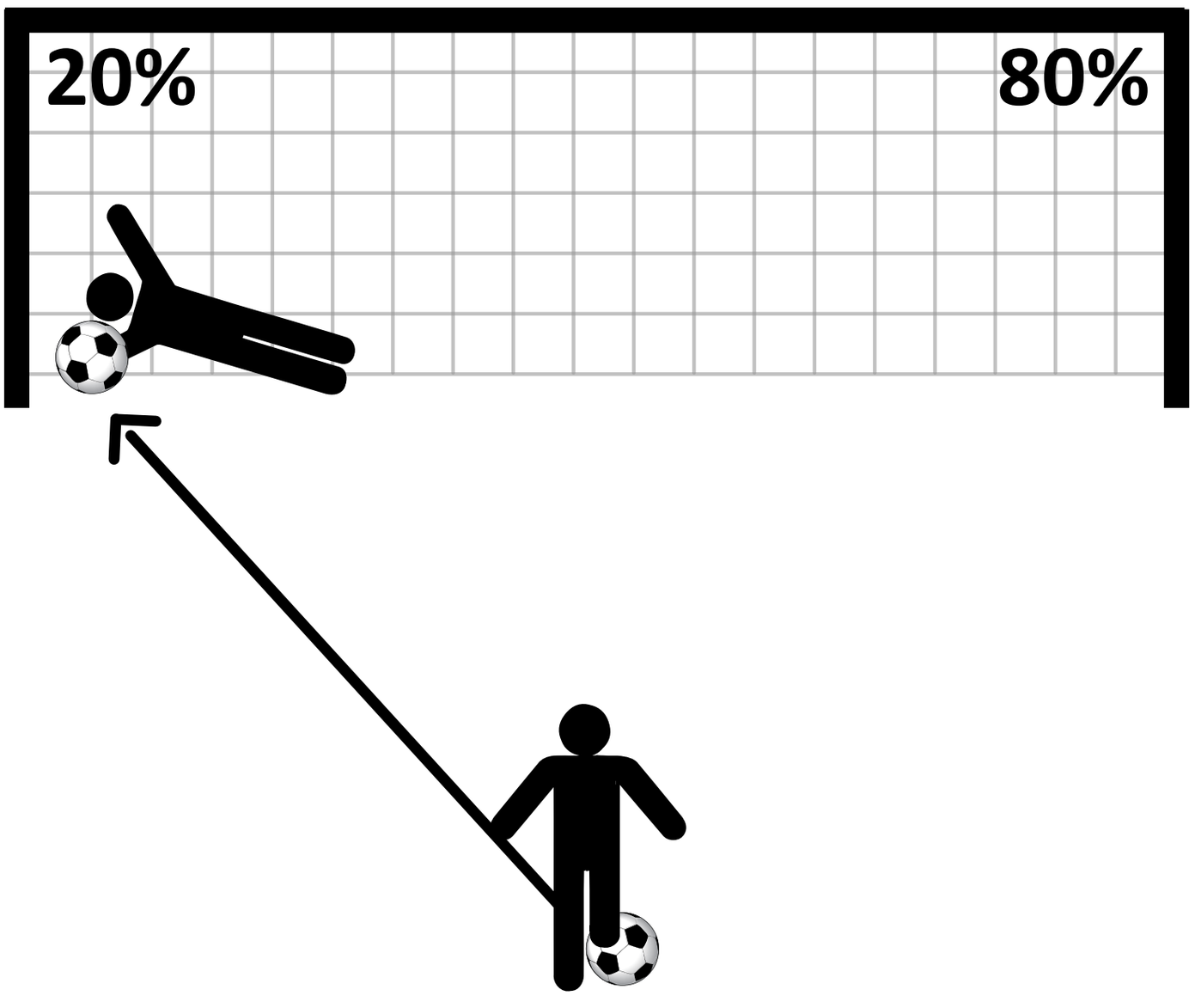


N = 14



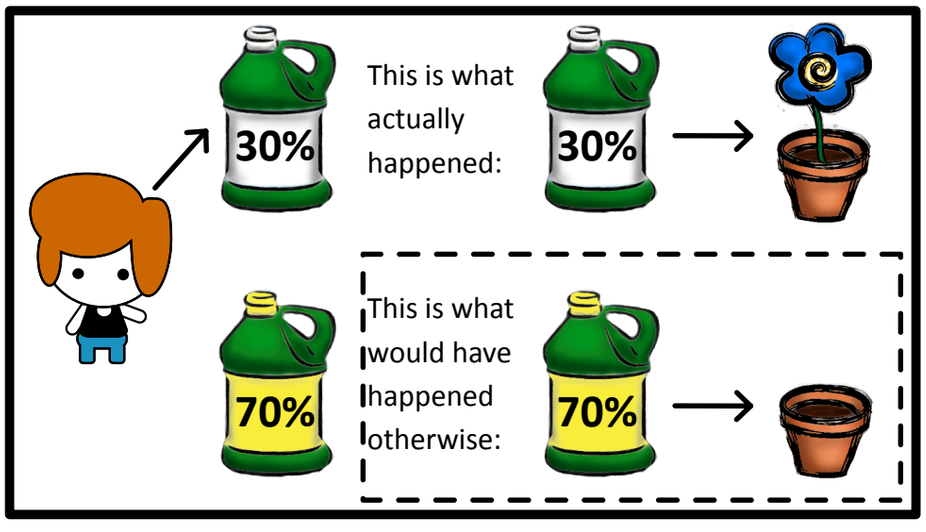


$$\text{Responsibility} = f\left(\text{Difference in Expectation}, \text{Pivotality}\right)$$

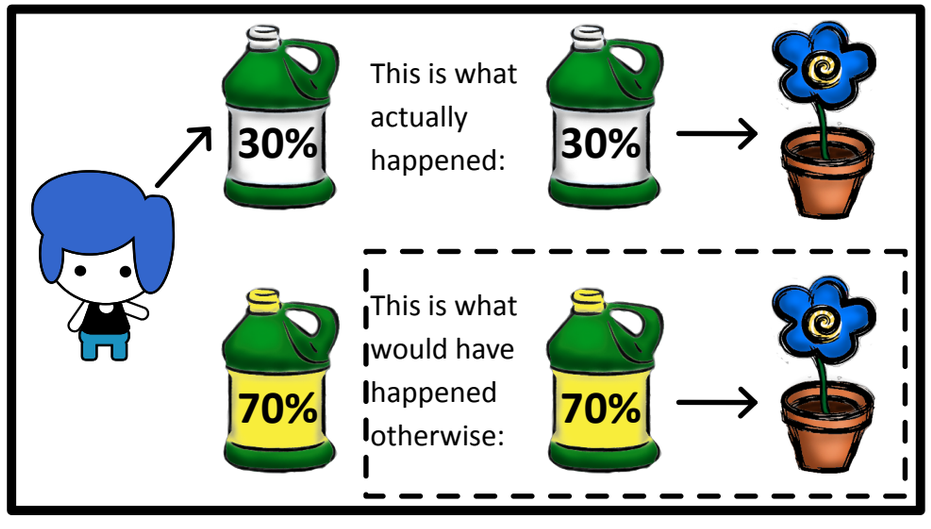


Responsibility = f(Difference in Expectation , Pivotality)

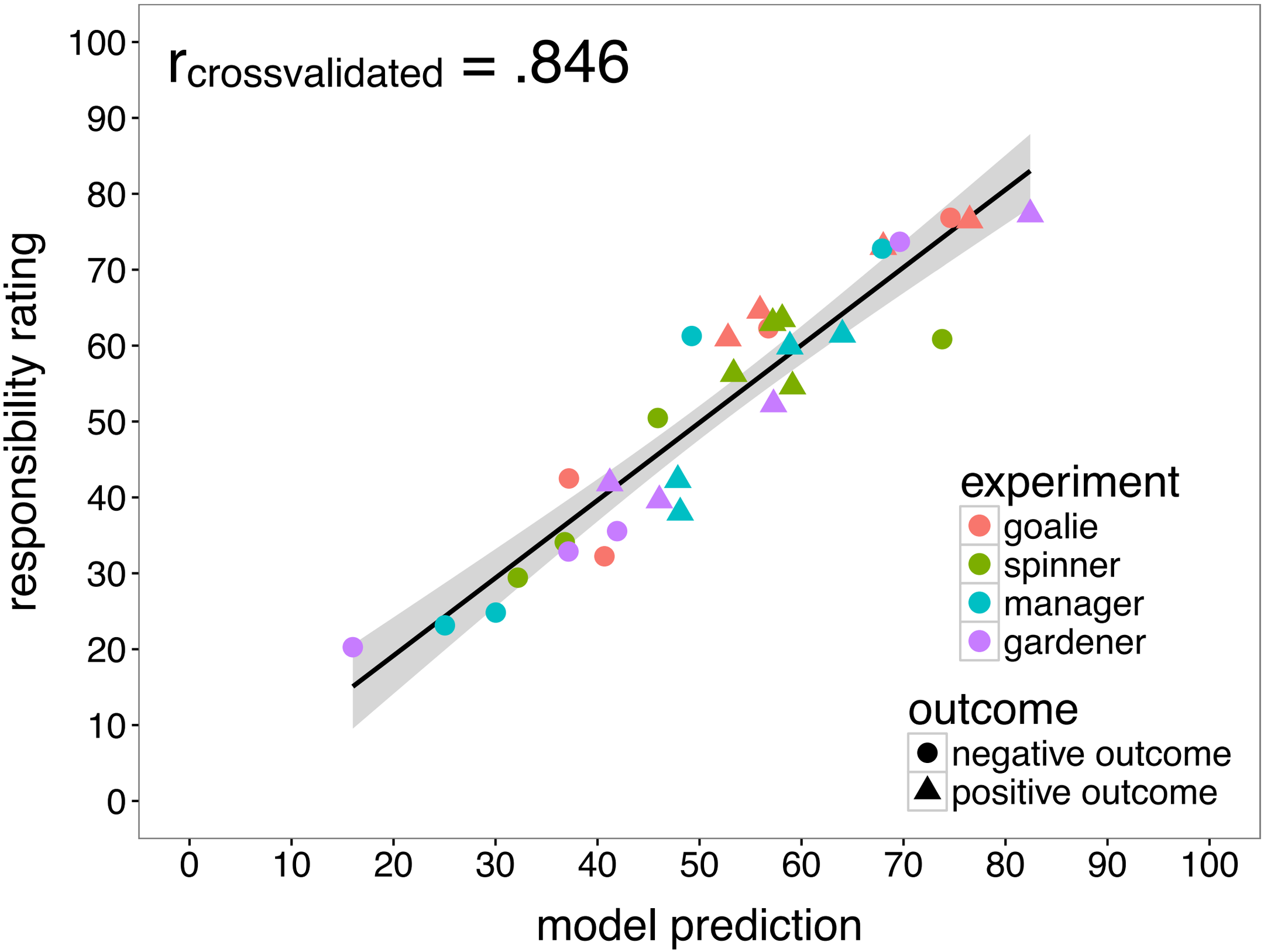
pivotal



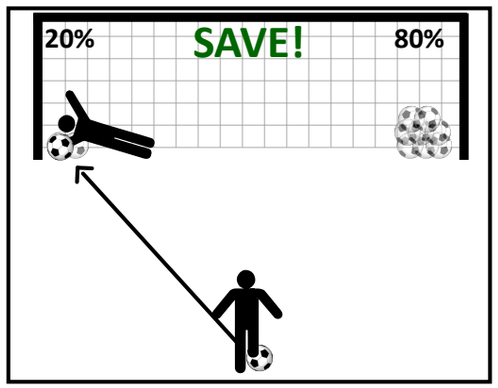
non-pivotal



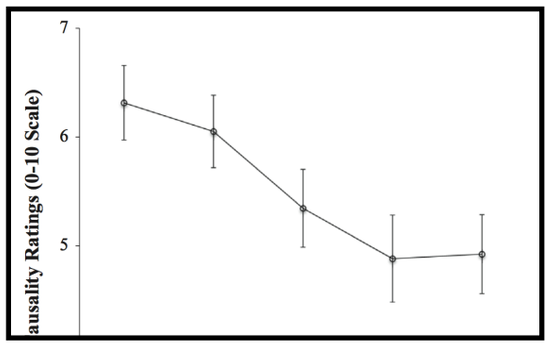
$$\text{Responsibility} = \alpha + w_1 \cdot (\text{Difference in expected reward}) + w_2 \cdot (\text{Pivotality for outcome})$$



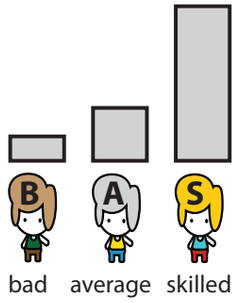
Summary



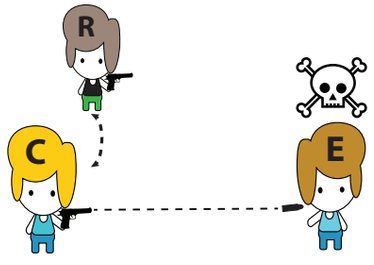
- holding others responsible for decisions under uncertainty



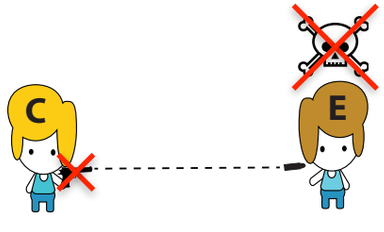
- no simple mapping from action expectations to responsibility



- inference over what type of person a person is



- person-based counterfactual contrast

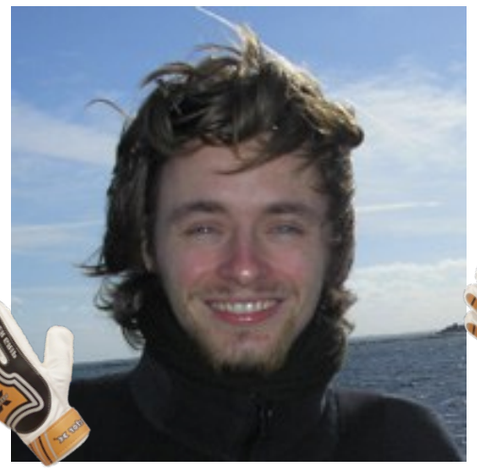


- action-based counterfactual contrast

Thank you!

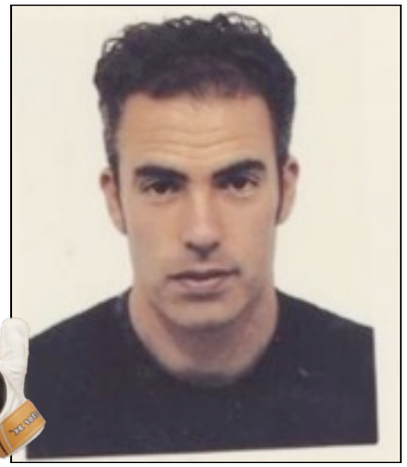
Tomer Ullman

Max Kleiman-Weiner

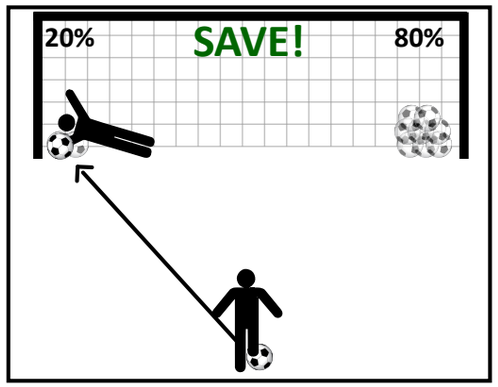


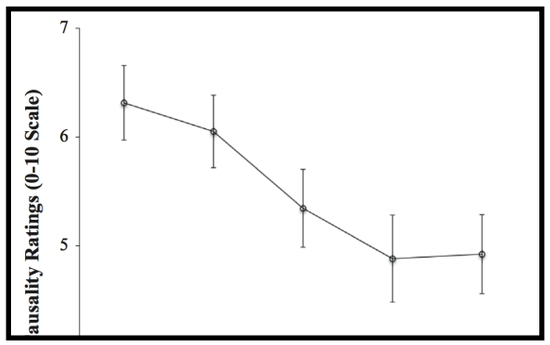
Jonas Nagel

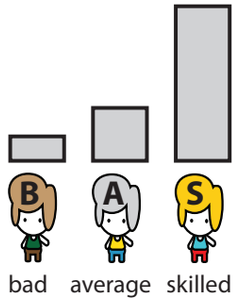
Josh Tenenbaum

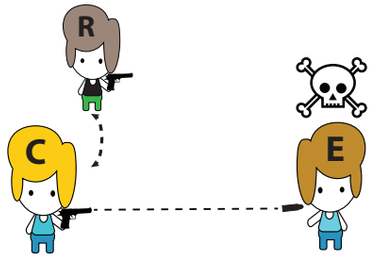


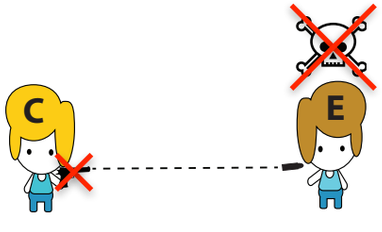
Dave Lagnado







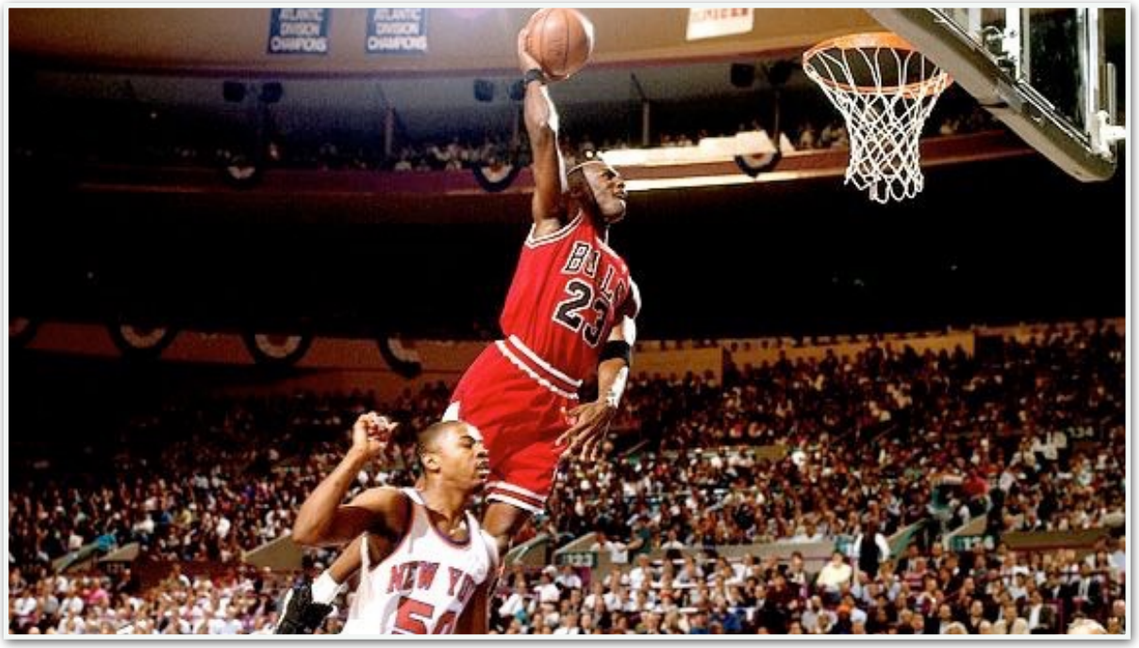




Summary

- holding others responsible for decisions under uncertainty
- no simple mapping from action expectations to responsibility
- inference over what type of person a person is
- person-based counterfactual contrast
- action-based counterfactual contrast

Which norms guide our expectations?

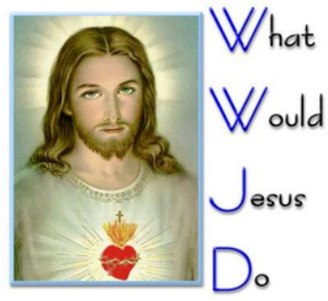


population norm

reference group norm

self-as-norm

idealistic norm



person norm