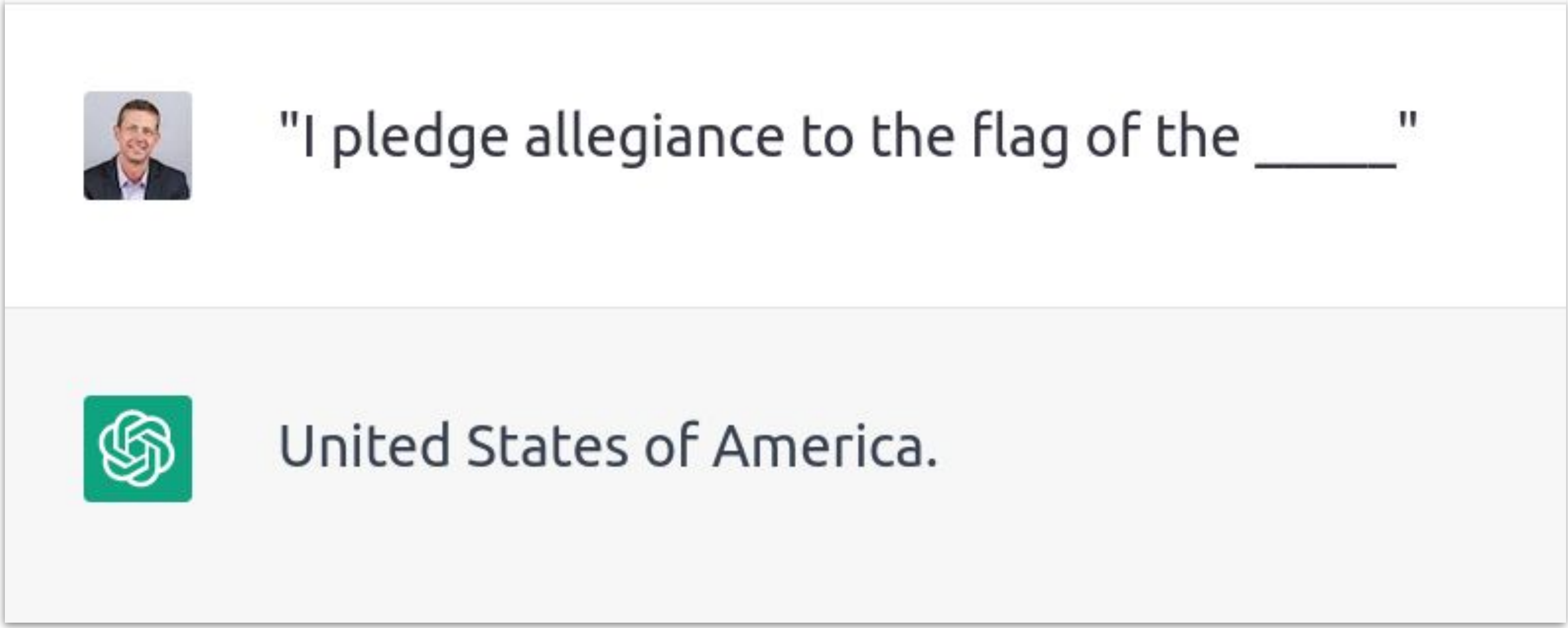
Large Language Models as Economic Agents: What Can We Learn from Homo Silicus?

John Horton
MIT Sloan & NBER

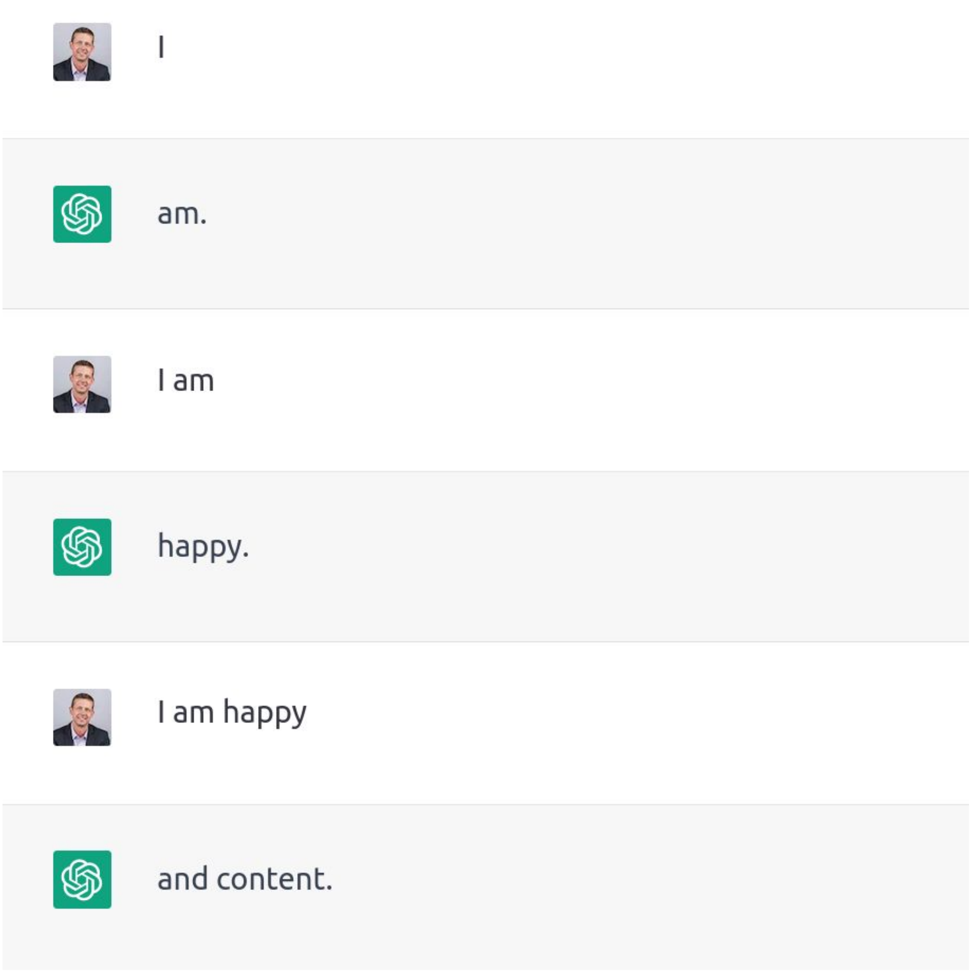
What are language models?

- They are prediction models that, given a sequence of words ("tokens"), they predict the statistically most likely completion.
- Example:
  - "I pledge allegiance to the flag of the __________"
- Two possibilities:
  - "I pledge allegiance to the flag of the `United States`" (highly probable)
  - "I pledge allegiance to the flag of the `Chicken Sandwich`" (highly improbable)
- Recently have become **far** more capable
  - Architectural advances (e.g., "Attention is All you Need")
  - Compute + Training Samples



"I pledge allegiance to the flag of the _____"

United States of America.



I

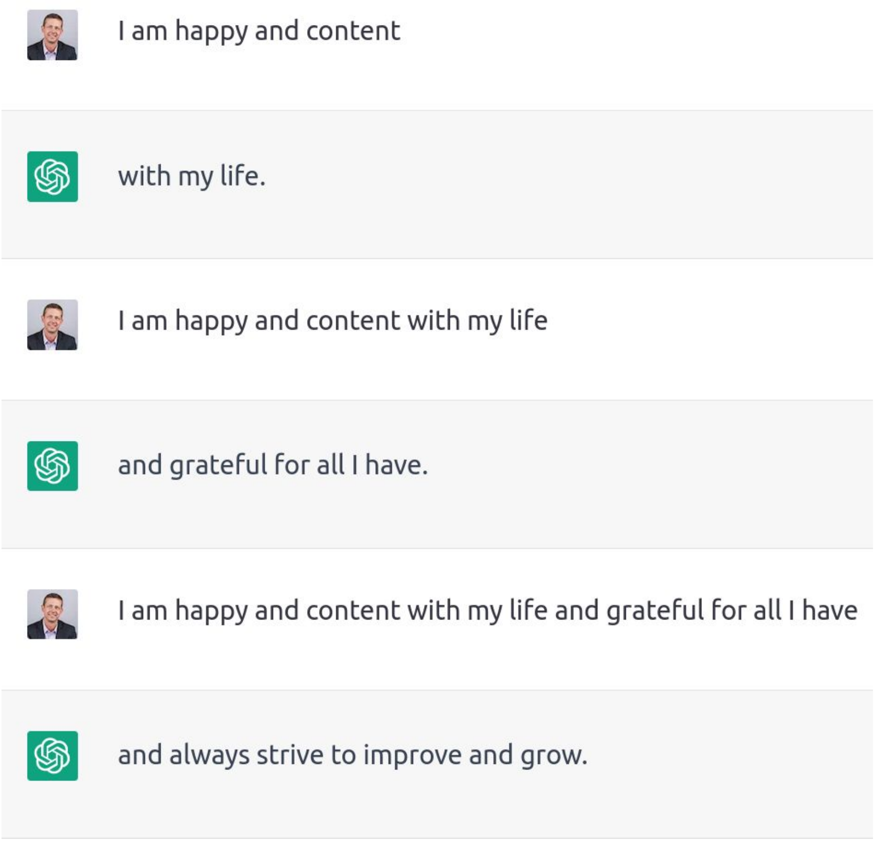
am.

I am

happy.

I am happy

and content.



I am happy and content

with my life.

I am happy and content with my life

and grateful for all I have.

I am happy and content with my life and grateful for all I have

and always strive to improve and grow.

and so on…

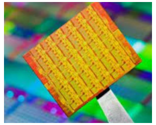
An LLM: GPT-3 (Generative Pre-Trained Transformer 3)

- Large language model developed by OpenAI
  - Billions of parameters
  - Trained on a massive of corpus of text
  - Accessible via OpenAI API
    - This is key for this project
  - Will respond to "prompts" written in natural language
  - Remarkably capable of giving "realistic" responses
- The basic idea of this paper:
  - Use these GPT agents as experimental subjects!
  - Can these simulated economic agents---a homo silicus---teach us something about the social world?

Idea of Homo Economicus

- **Homo Economicus**: A maintained model of human behavior
  - Rationally pursues objectives
  - Unlimited memory and computation
- **Theory research**: Putting *Homo Economicus* in exciting new scenarios
  - As worker or employer (Labor Economics)
  - As consumer (Consumer theory)
  - As investor/trader (Finance)
  - As government / tay payer (Public finance / public economics)
  - and so on
- **Empirical research**: How does *Homo Sapiens* compare?

Idea of Homo ~~Economicus~~ Silicus

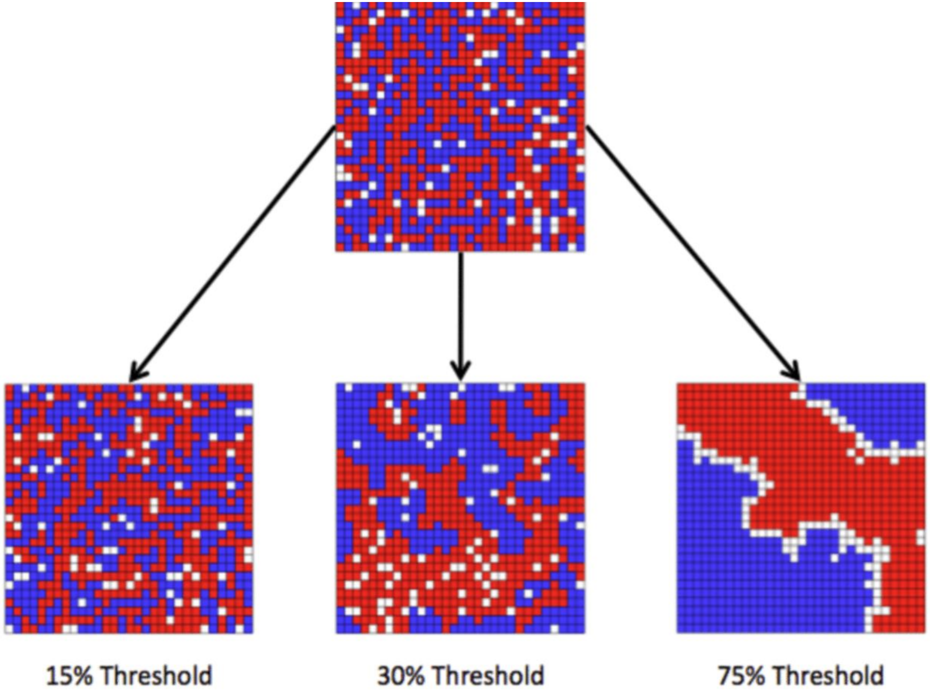


Computer chips →
made from Silicon

- **Homo Economicus**: A ~~maintained~~ computational model of human behavior
  - ~~Rationally pursues objectives~~ Does whatever the model predicts is statistically probable
  - ~~Unlimited memory and computation~~
- **Theory research**: Putting *Homo Silicus* in exciting new scenarios
  - As worker or employer (Labor Economics)
  - As consumer (Consumer theory)
  - As investor/trader (Finance)
  - As government / tay payer (Public finance / public economics)
  - and so on
- **Empirical research**: How does *Homo Sapiens* compare?

Aren't these just Agent Based Models (ABMs)?

- There are similarities, but the enormous difference is that we do not get to program Homo Silicus; but with ABMs, the researcher gets to program behavior



Agenda for talk

- Present results from a series of Homo Economicus experiments I conducted, drawn from classics in behavioral economics
  - A fairness experiment
  - A social preferences experiment
  - A framing experiment
- Discuss some potential objections and limitations of this approach
- Future research

A fairness experiment

Fairness as a Constraint on Profit Seeking: Entitlements in the Market

By Daniel Kahneman, Jack L. Knetsch, and Richard Thaler*

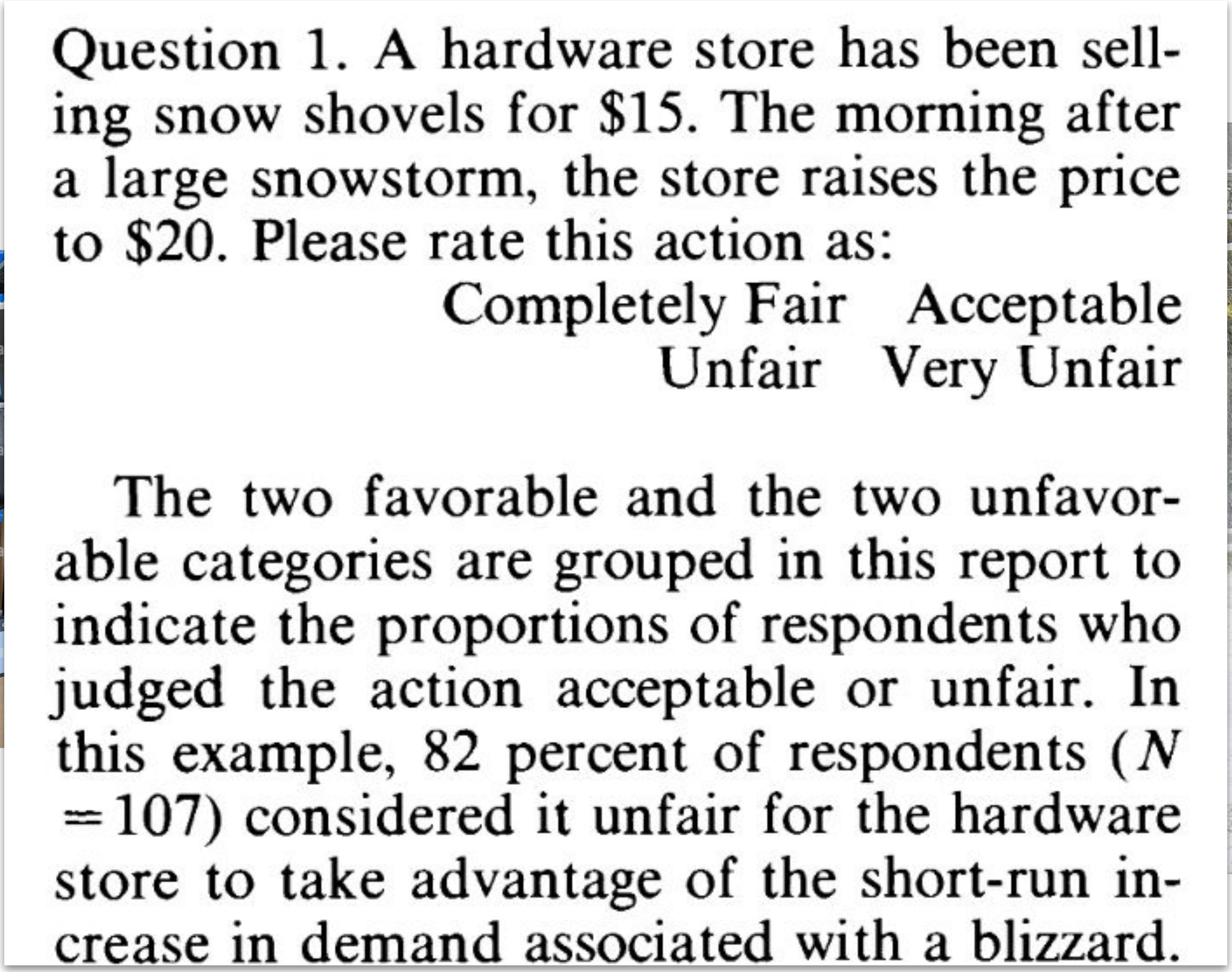
Community standards of fairness for the setting of prices and wages were elicited by telephone surveys. In customer or labor markets, it is acceptable for a firm to raise prices (or cut wages) when profits are threatened and to maintain prices when costs diminish. It is unfair to exploit shifts in demand by raising prices or cutting wages. Several market anomalies are explained by assuming that these standards of fairness influence the behavior of firms.

Just as it is often useful to neglect friction in elementary mechanics, there may be good reasons to assume that firms seek their maximal profit as if they were subject only to legal and budgetary constraints. However, the patterns of sluggish or incomplete adjustment often observed in markets suggest that some additional constraints are operative. Several authors have used a notion of fairness to explain why many employers do not cut wages during periods of high unemployment (George Akerlof, 1979; Robert Solow, 1980). Arthur Okun (1981) went further in arguing that fairness also alters the outcomes in what he called customer markets—characterized by suppliers who are perceived as making their own pricing decisions, have some monopoly power (if only because search is costly), and often have repeat business with their clientele. Like labor markets, customer markets also sometimes fail to clear:

> ...firms in the sports and entertainment industries offer their customers tickets at standard prices for events that clearly generate excess demand. Popular new models of automobiles may have waiting lists that extend for months. Similarly, manufacturers in a number of industries operate with backlogs in booms and allocate shipments when they obviously could raise prices and reduce the queue. [p. 170]

Okun explained these observations by the hostile reaction of customers to price increases that are not justified by increased costs and are therefore viewed as unfair. He also noted that customers appear willing to accept "fair" price increases even when demand is slack, and commented that "...in practice, observed pricing behavior is a vast distance from do it yourself auctioneering" (p. 170).

The argument used by these authors to account for apparent deviations from the simple model of a profit-maximizing firm is that fair behavior is instrumental to the maximization of long-run profits. In Okun's model, customers who suspect that a sup-

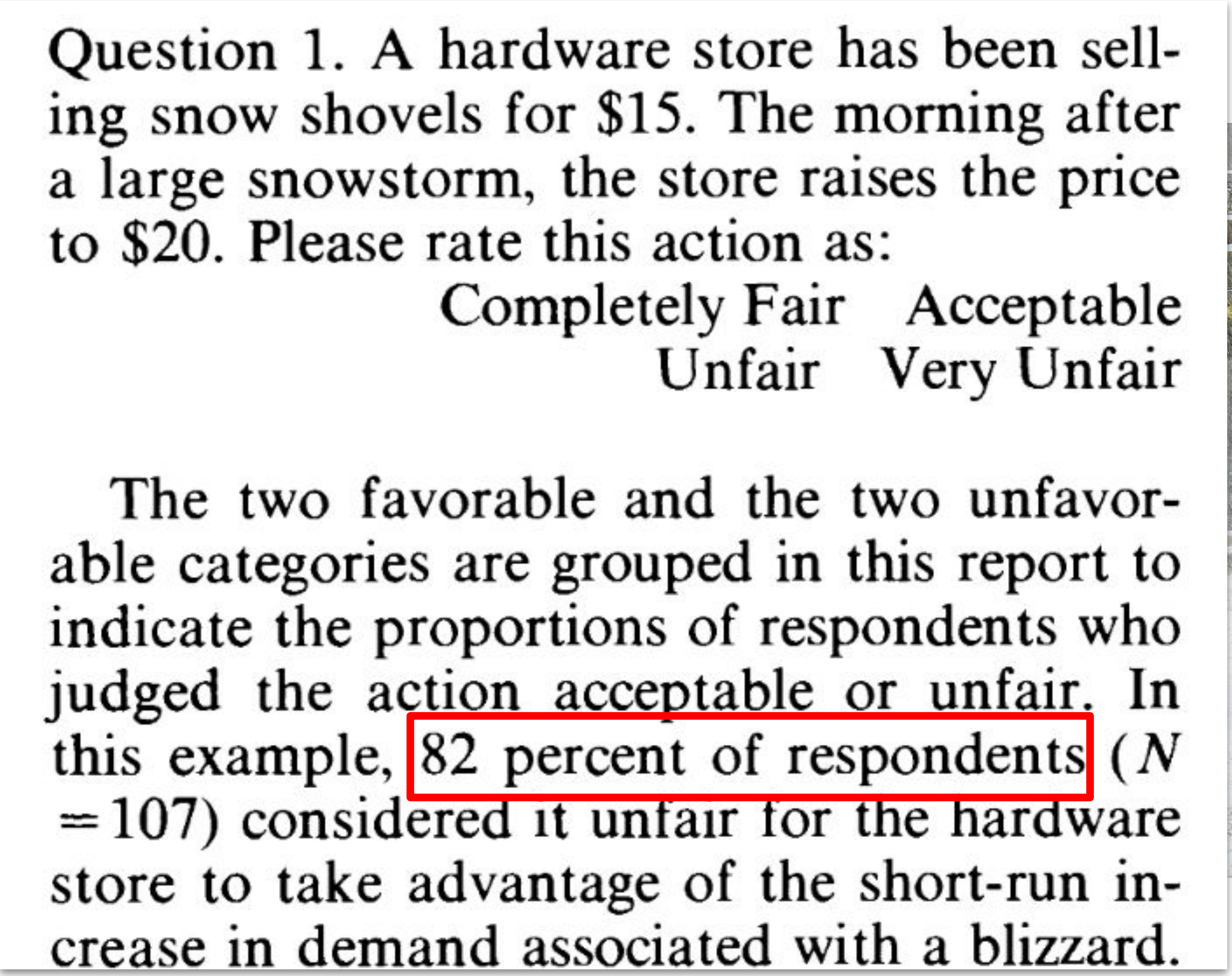


Question 1. A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store raises the price to $20. Please rate this action as:

Completely Fair Acceptable
Unfair Very Unfair

The two favorable and the two unfavorable categories are grouped in this report to indicate the proportions of respondents who judged the action acceptable or unfair. In this example, 82 percent of respondents (N = 107) considered it unfair for the hardware store to take advantage of the short-run increase in demand associated with a blizzard.

PRICE GOUGING IS
ILLEGAL
ERIE.GOV/CONSUMERPROTECTION

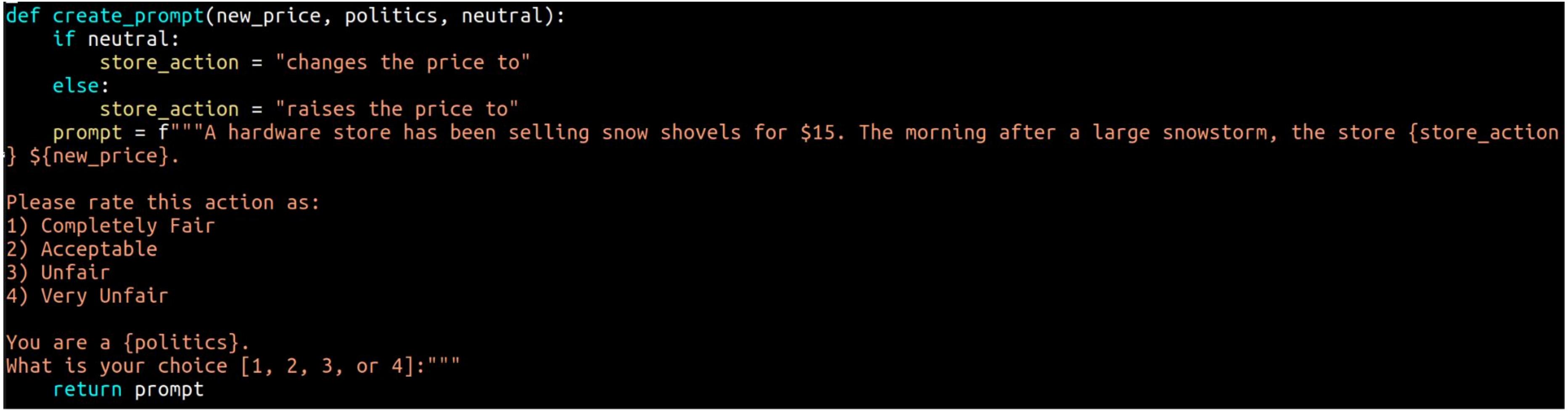


Question 1. A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store raises the price to $20. Please rate this action as:

Completely Fair Acceptable
Unfair Very Unfair

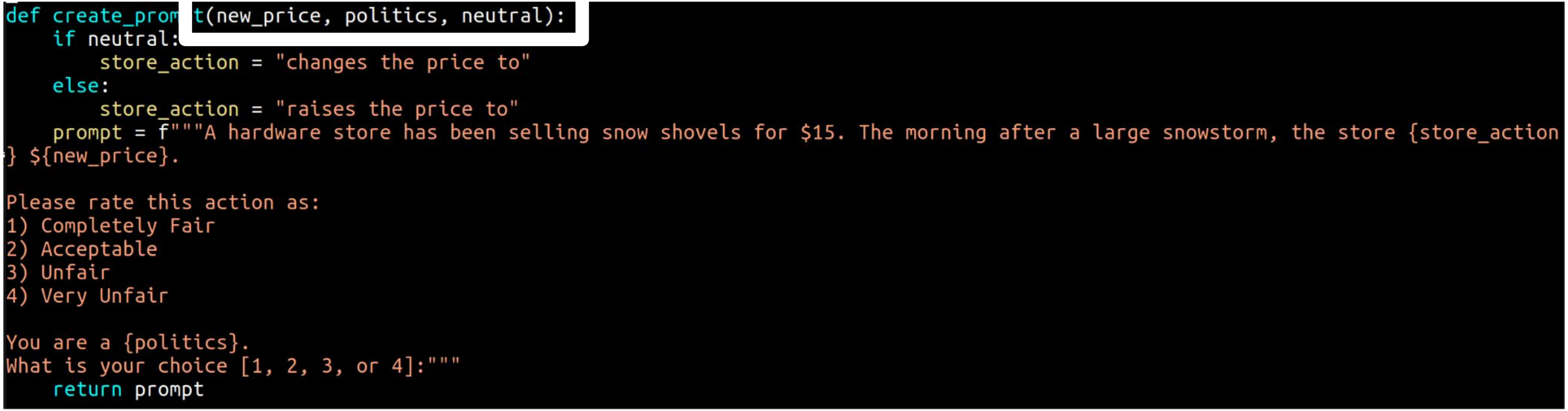
The two favorable and the two unfavorable categories are grouped in this report to indicate the proportions of respondents who judged the action acceptable or unfair. In this example, 82 percent of respondents (N = 107) considered it unfair for the hardware store to take advantage of the short-run increase in demand associated with a blizzard.

Sending the scenario as a prompt to GPT Agent via API



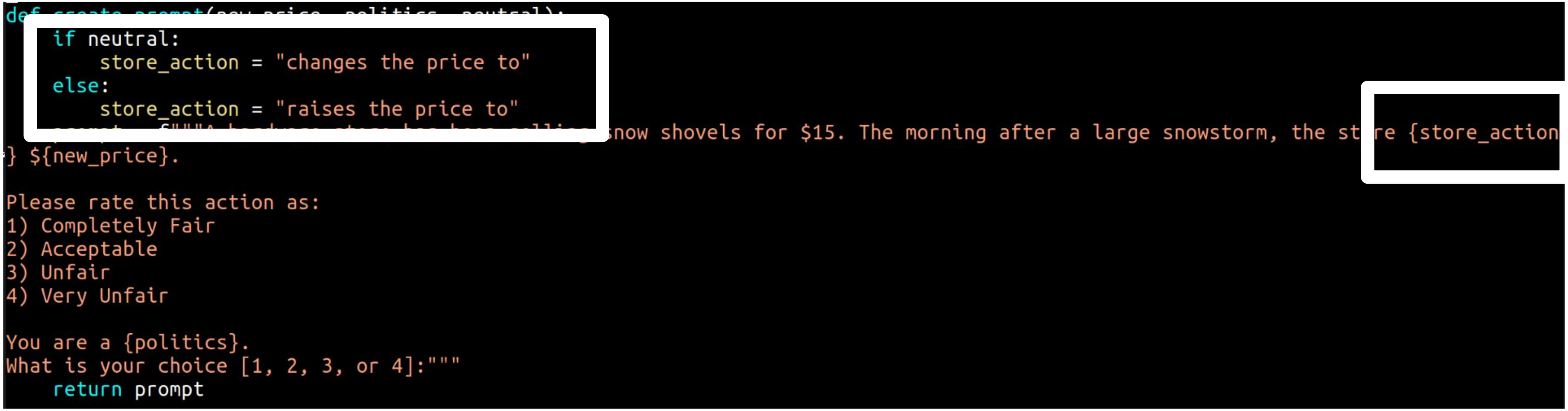
```
def create_prompt(new_price, politics, neutral):
    if neutral:
        store_action = "changes the price to"
    else:
        store_action = "raises the price to"
    prompt = f"""A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store {store_action
} ${new_price}.

Please rate this action as:
1) Completely Fair
2) Acceptable
3) Unfair
4) Very Unfair

You are a {politics}.
What is your choice [1, 2, 3, or 4]:"""
    return prompt
```

Factors I can vary
(a Python function to generate 'prompts')



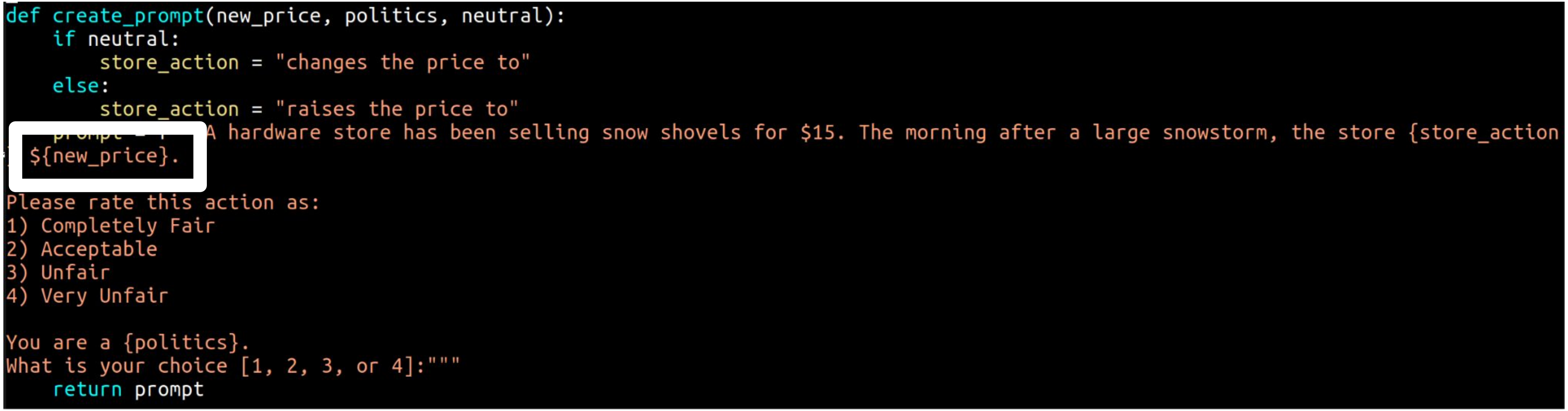
```
def create_prompt(new_price, politics, neutral):
    if neutral:
        store_action = "changes the price to"
    else:
        store_action = "raises the price to"
    prompt = f"""A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store {store_action
} ${new_price}.

Please rate this action as:
1) Completely Fair
2) Acceptable
3) Unfair
4) Very Unfair

You are a {politics}.
What is your choice [1, 2, 3, or 4]:"""
    return prompt
```

I can alter the framing of the change: "raises" versus "changes"



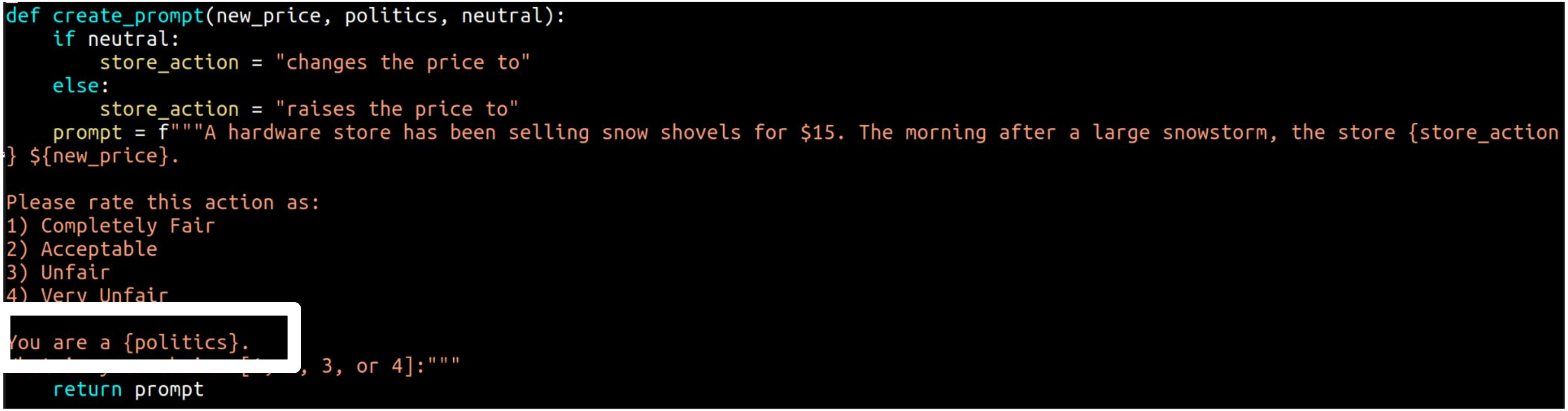
```
def create_prompt(new_price, politics, neutral):
    if neutral:
        store_action = "changes the price to"
    else:
        store_action = "raises the price to"
    prompt = f"""A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store {store_action
} ${new_price}.

Please rate this action as:
1) Completely Fair
2) Acceptable
3) Unfair
4) Very Unfair

You are a {politics}.
What is your choice [1, 2, 3, or 4]:"""
    return prompt
```

I can alter the new price for the snow shovel

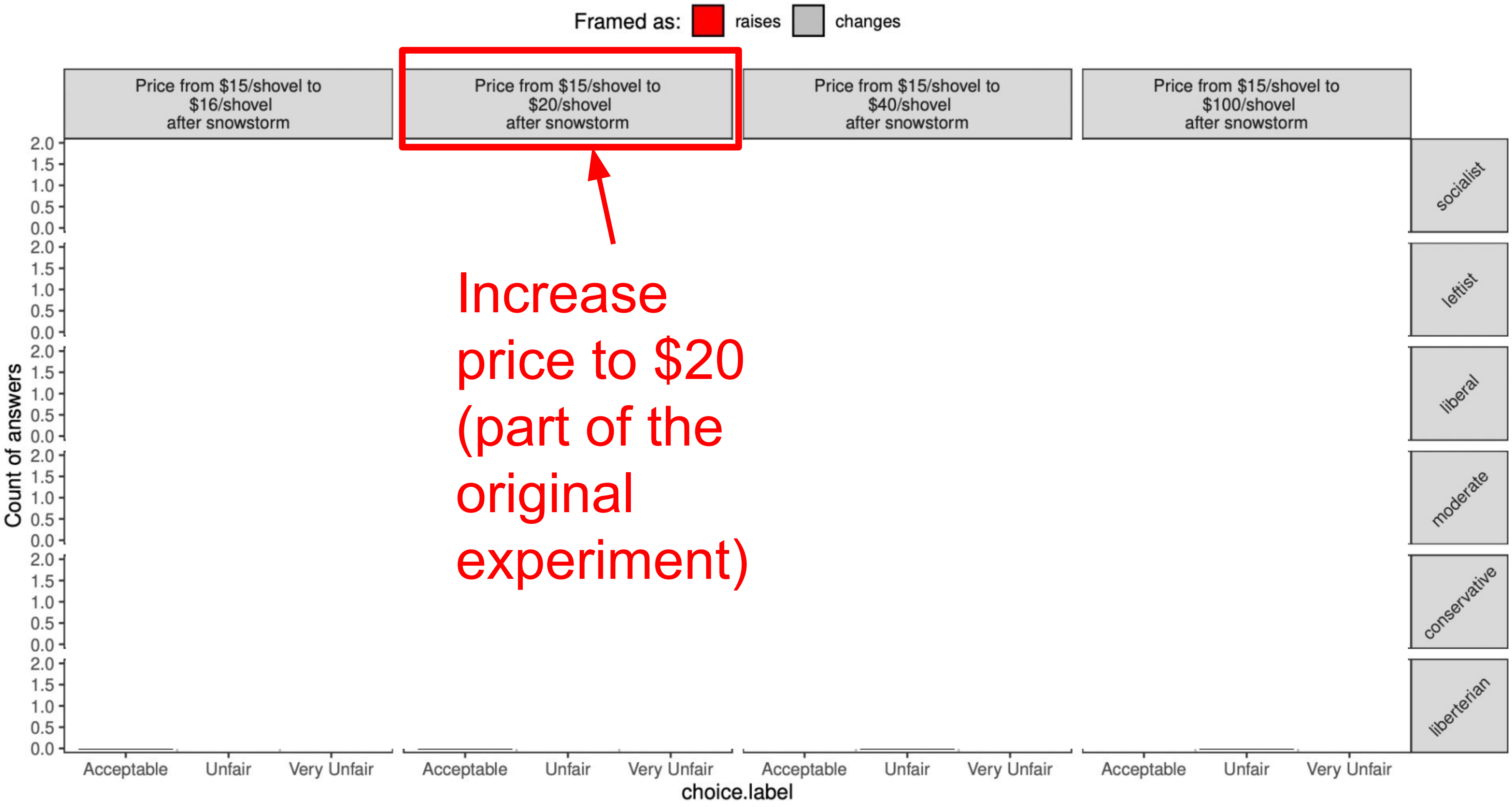


```
def create_prompt(new_price, politics, neutral):
    if neutral:
        store_action = "changes the price to"
    else:
        store_action = "raises the price to"
    prompt = f"""A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store {store_action
} ${new_price}.

Please rate this action as:
1) Completely Fair
2) Acceptable
3) Unfair
4) Very Unfair

You are a {politics}.
What is your choice [1, 2, 3, or 4]:"""
    return prompt
```

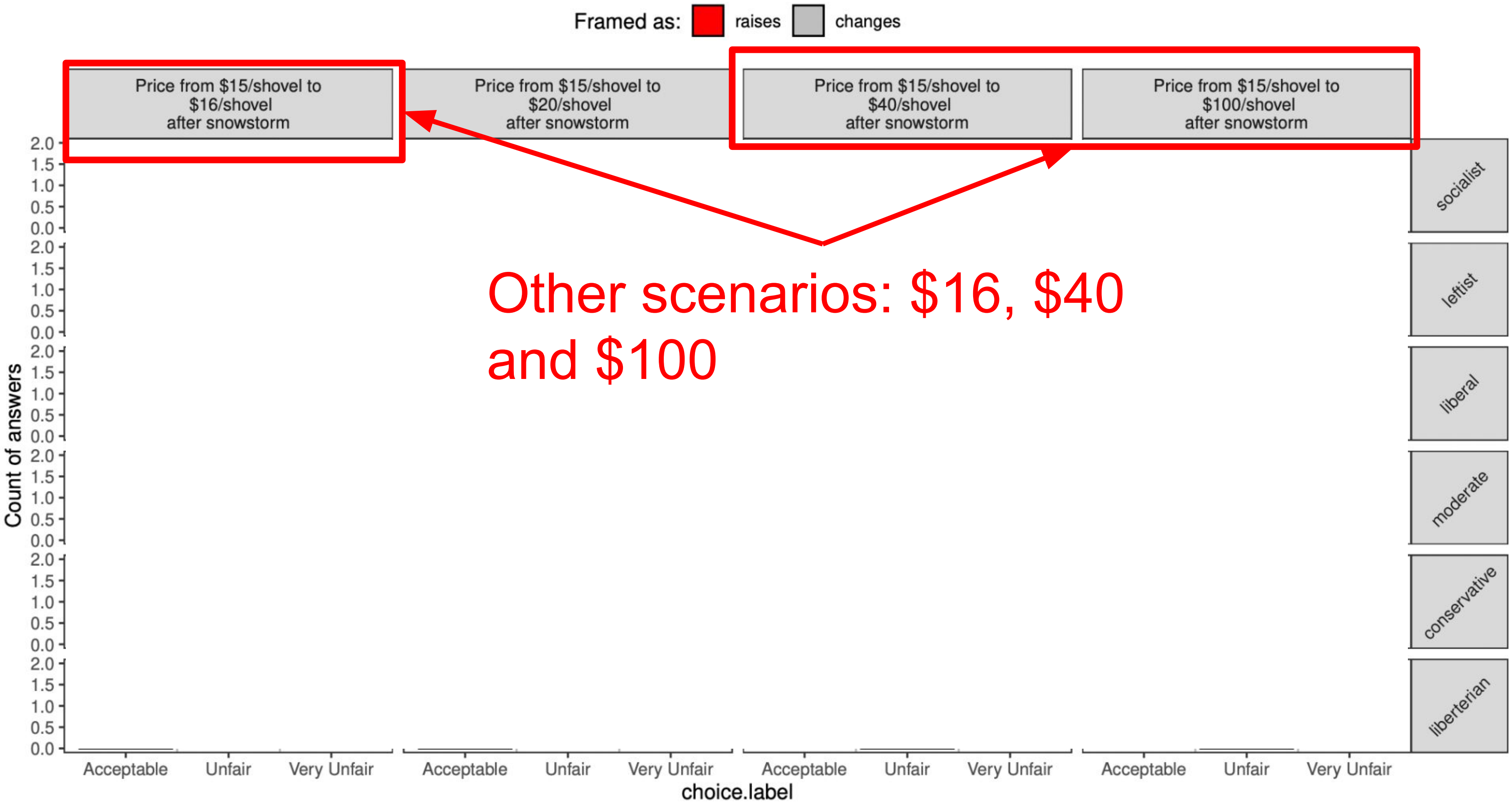
I can alter the "politics" of the GPT3 agent (Liberal, conservative, etc.)



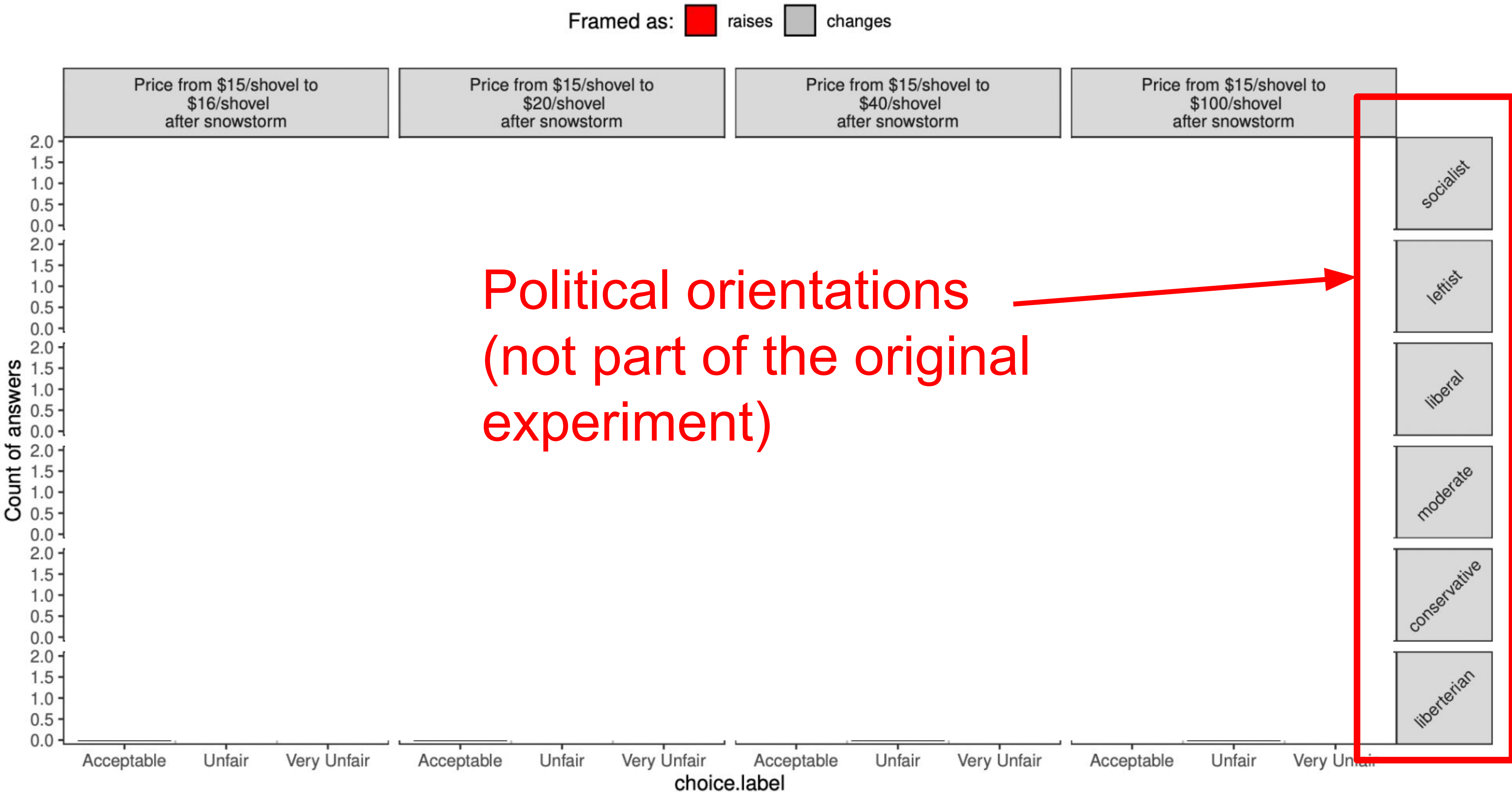
```
def create_prompt(new_price, politics, neutral):
    if neutral:
        store_action = "changes the price to"
    else:
        store_action = "raises the price to"
    prompt = f"""A hardware store has been selling snow shovels for $15. The morning after a large snowstorm, the store {store_action} ${new_price}.

Please rate this action as:
1) Completely Fair
2) Acceptable
3) Unfair
4) Very Unfair

You are a {politics}.
                                   , 3, or 4]:"""
    return prompt
```



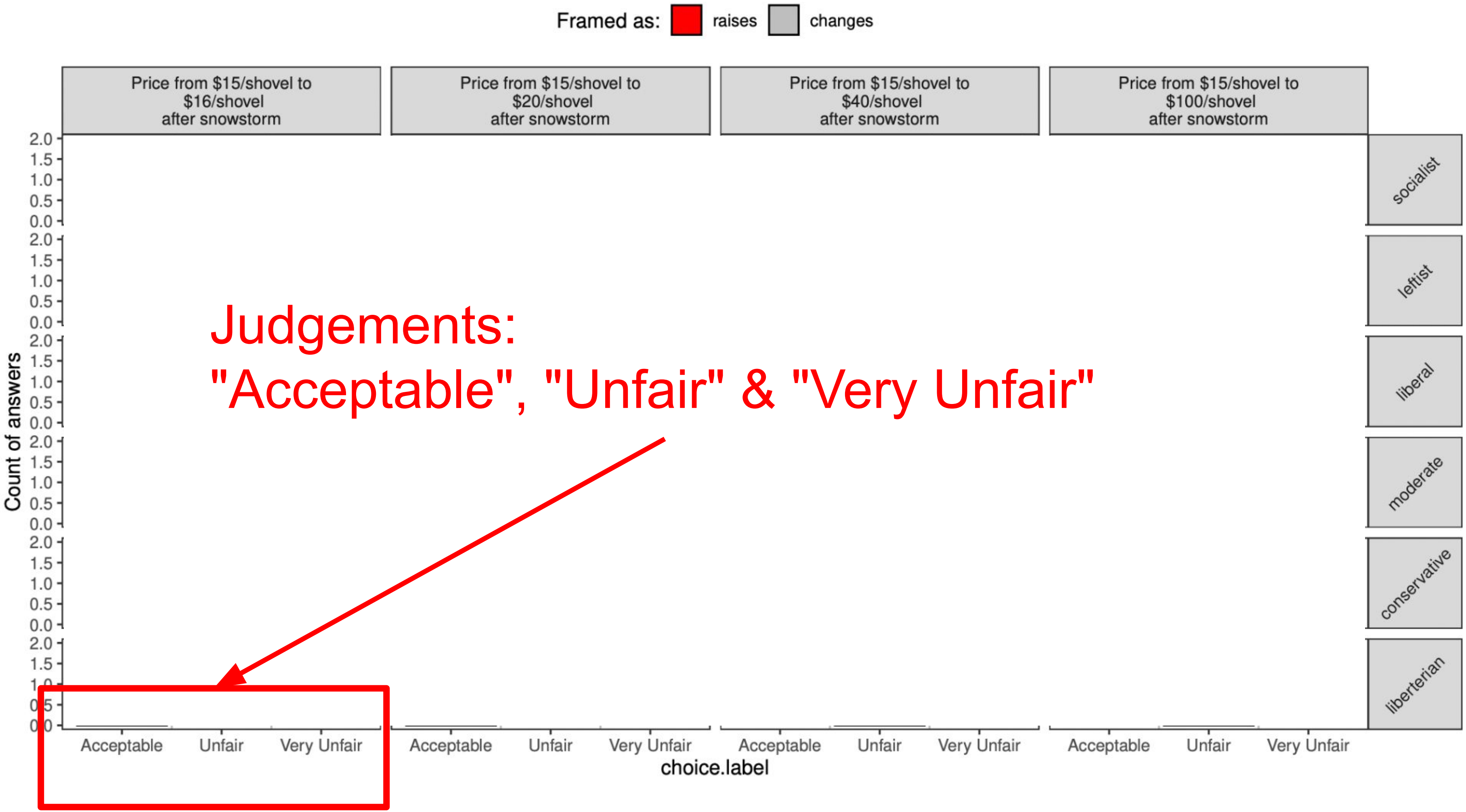
Increase price to $20 (part of the original experiment)



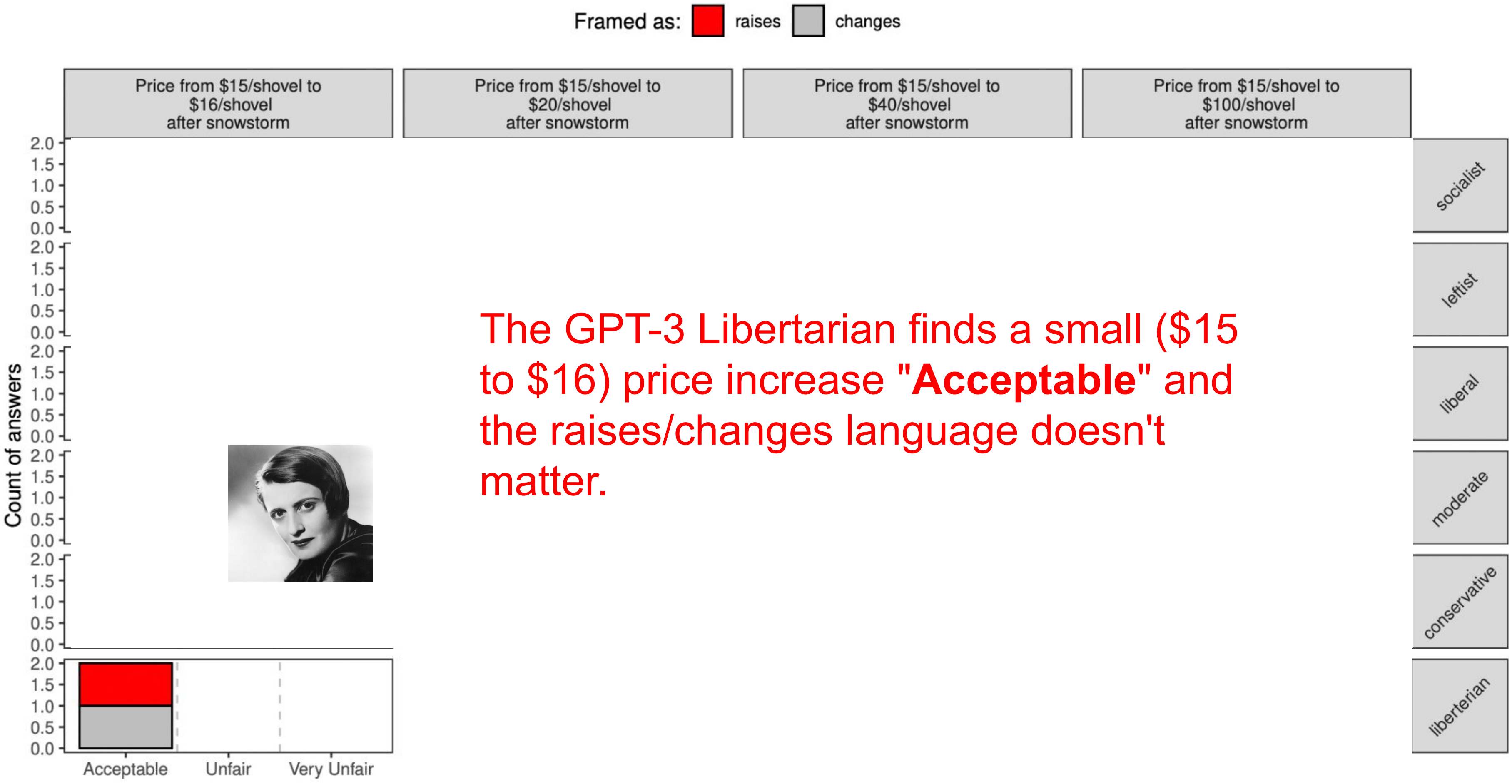
Other scenarios: $16, $40 and $100



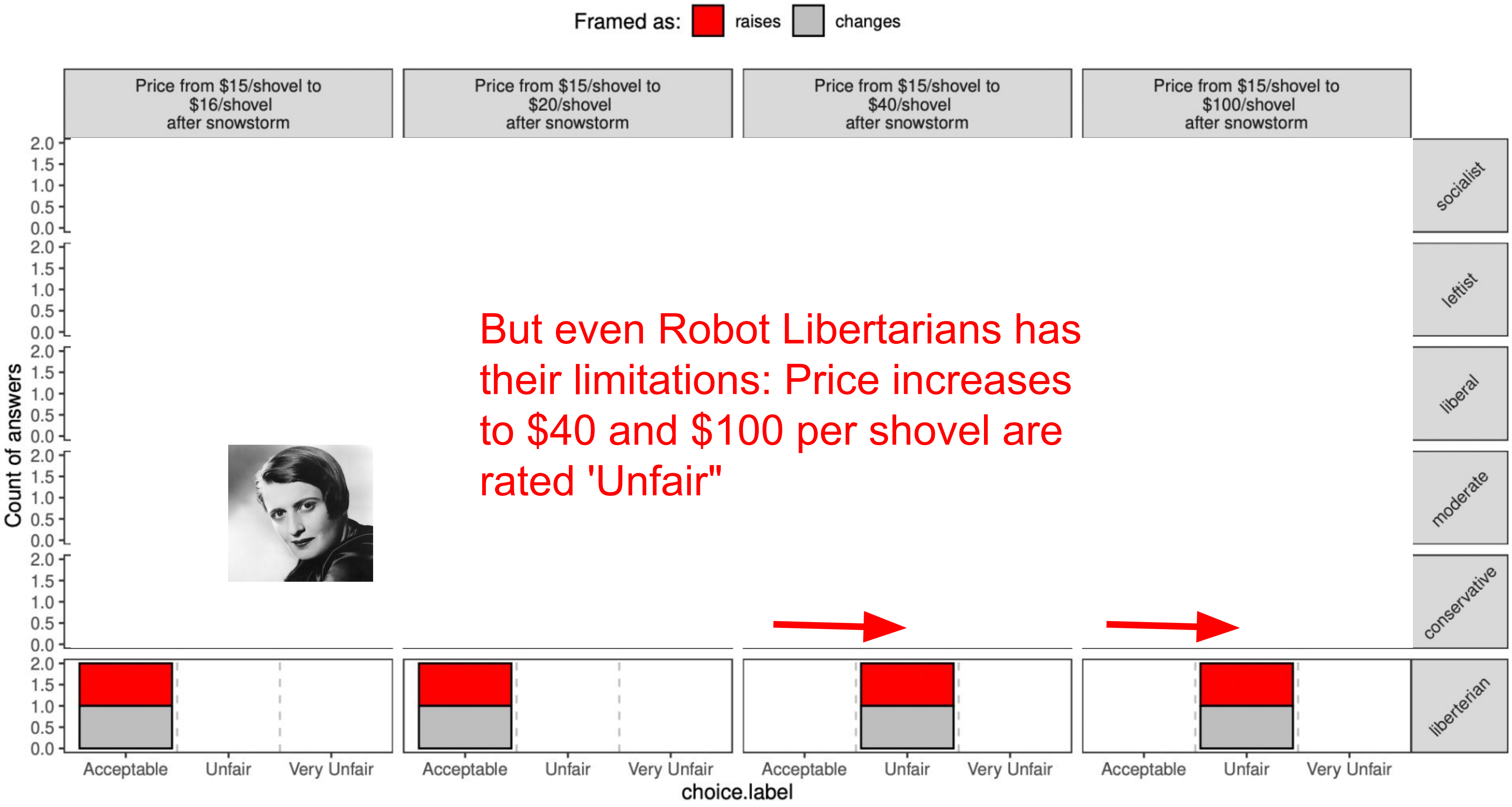
Political orientations (not part of the original experiment)



Judgements:
"Acceptable", "Unfair" & "Very Unfair"

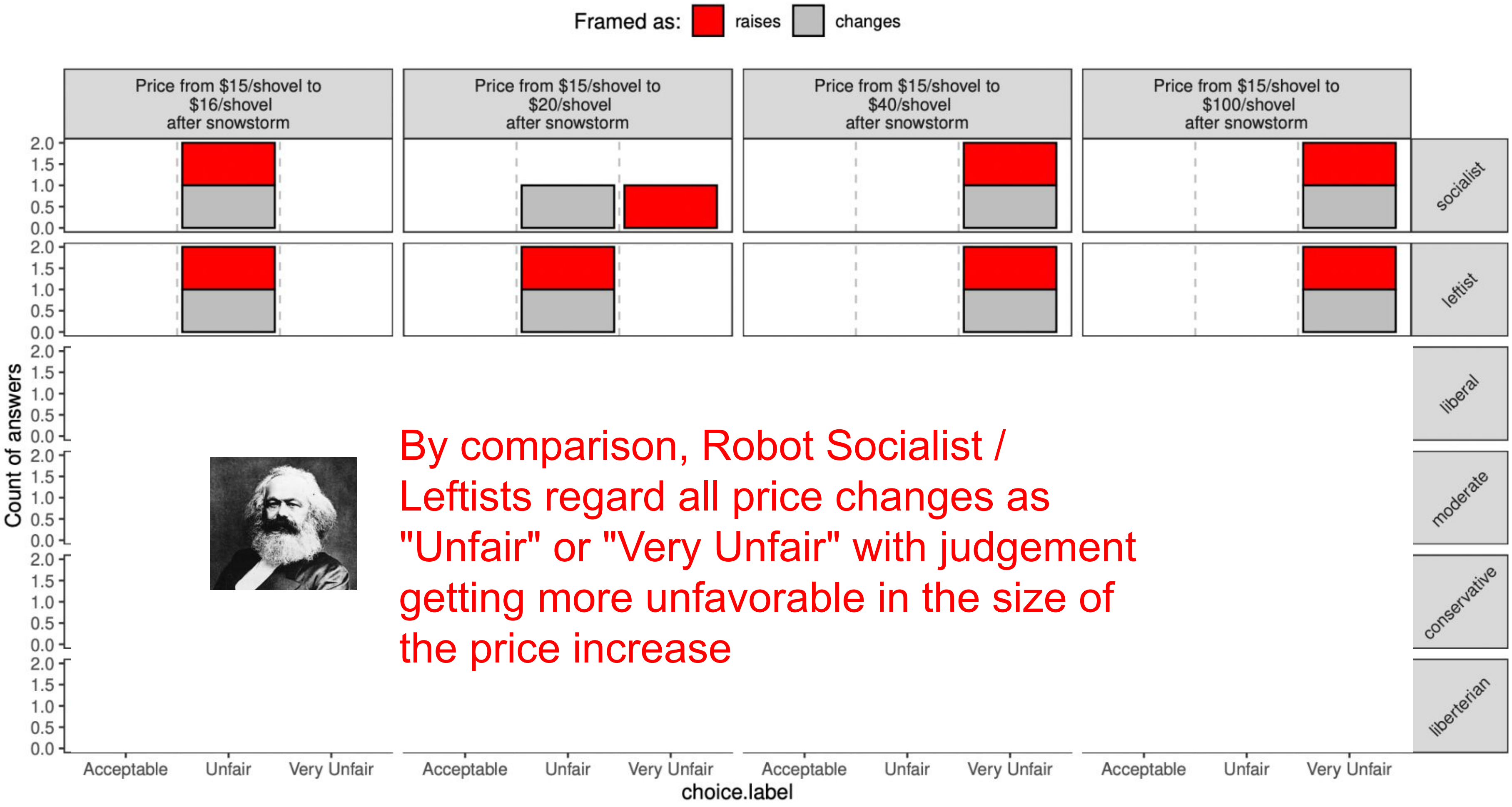


The GPT-3 Libertarian finds a small ($15 to $16) price increase "**Acceptable**" and the raises/changes language doesn't matter.



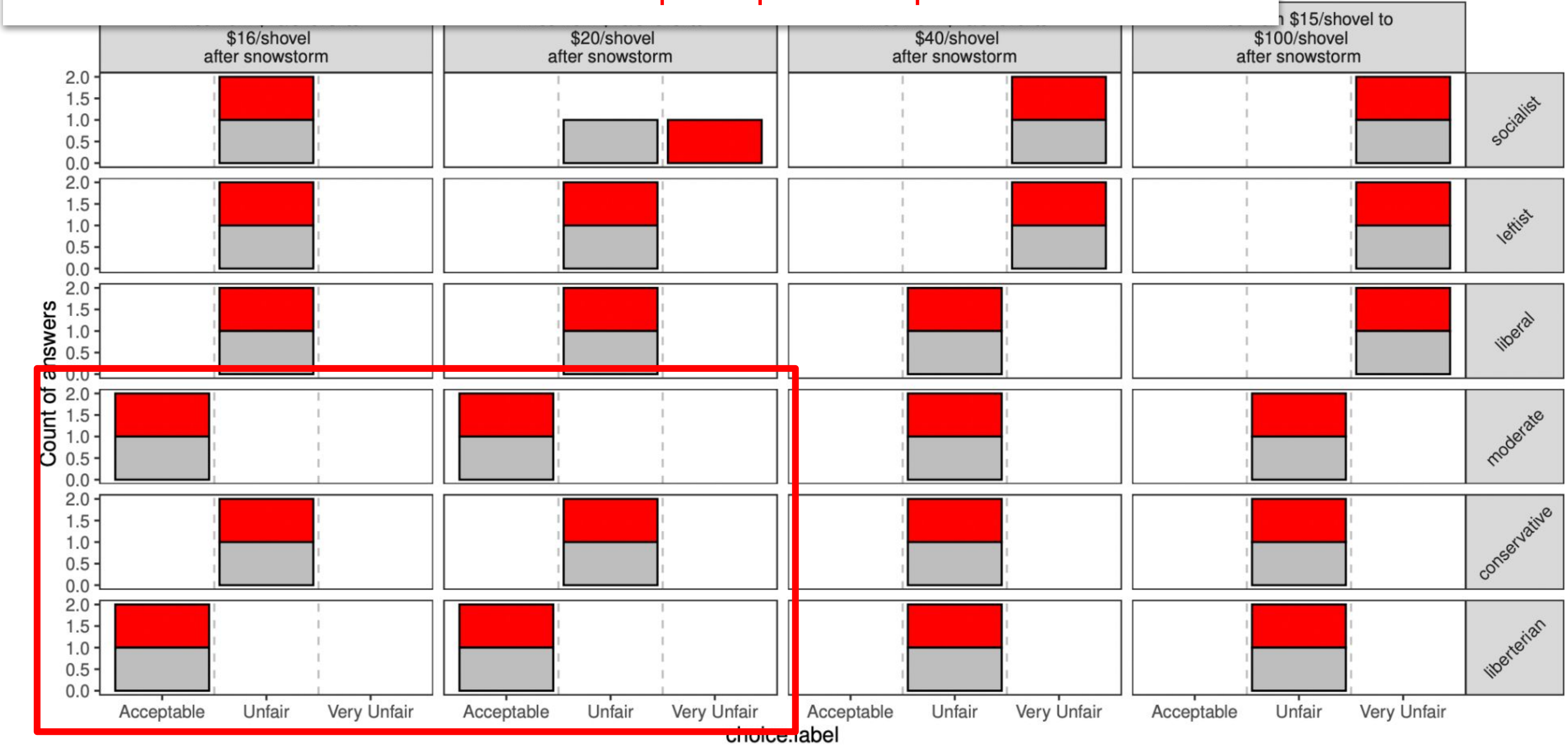
But even Robot Libertarians has their limitations: Price increases to $40 and $100 per shovel are rated 'Unfair"

Now prompt with a
different political orientation

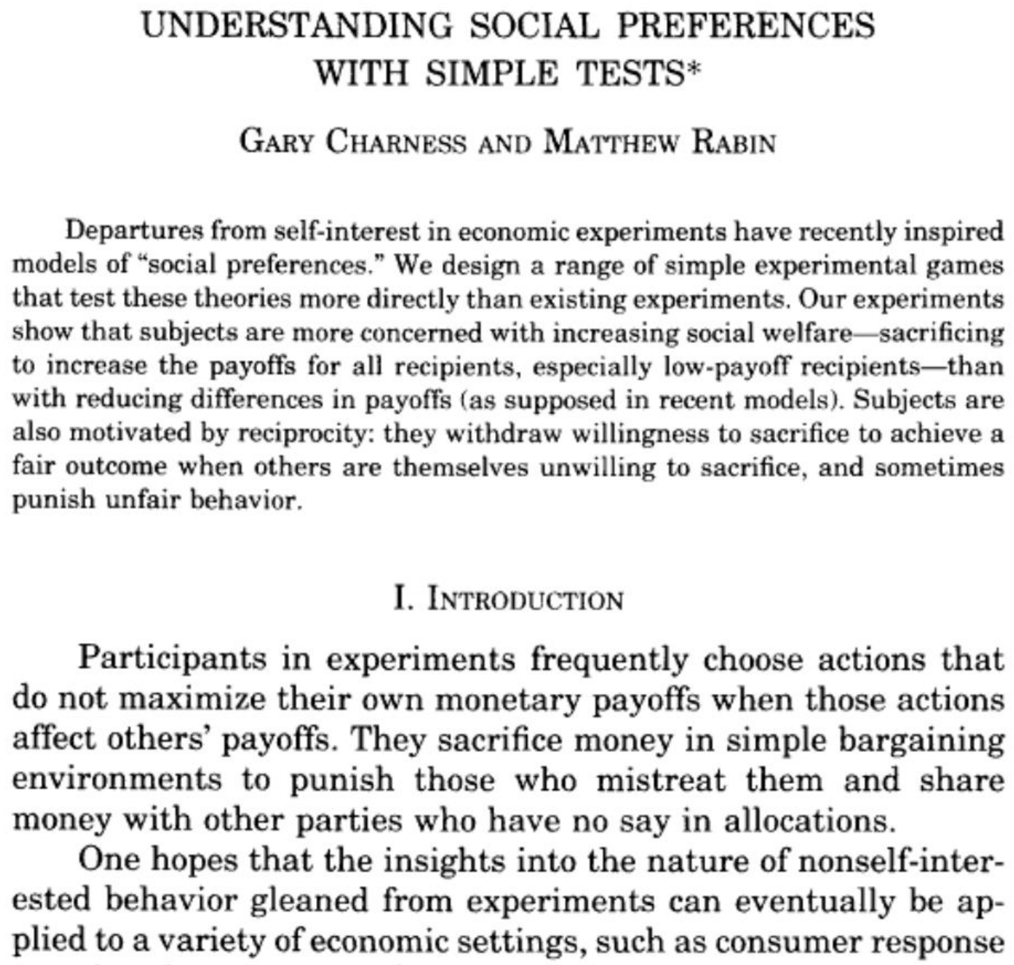


By comparison, Robot Socialist / Leftists regard all price changes as "Unfair" or "Very Unfair" with judgement getting more unfavorable in the size of the price increase

Interesting difference between "Conservatives" and "Libertarians" - could be the semantics of "conservative" or perhaps a real political distinction



A social preferences experiment



UNDERSTANDING SOCIAL PREFERENCES WITH SIMPLE TESTS*

Gary Charness and Matthew Rabin

Departures from self-interest in economic experiments have recently inspired models of "social preferences." We design a range of simple experimental games that test these theories more directly than existing experiments. Our experiments show that subjects are more concerned with increasing social welfare—sacrificing to increase the payoffs for all recipients, especially low-payoff recipients—than with reducing differences in payoffs (as supposed in recent models). Subjects are also motivated by reciprocity: they withdraw willingness to sacrifice to achieve a fair outcome when others are themselves unwilling to sacrifice, and sometimes punish unfair behavior.

I. Introduction

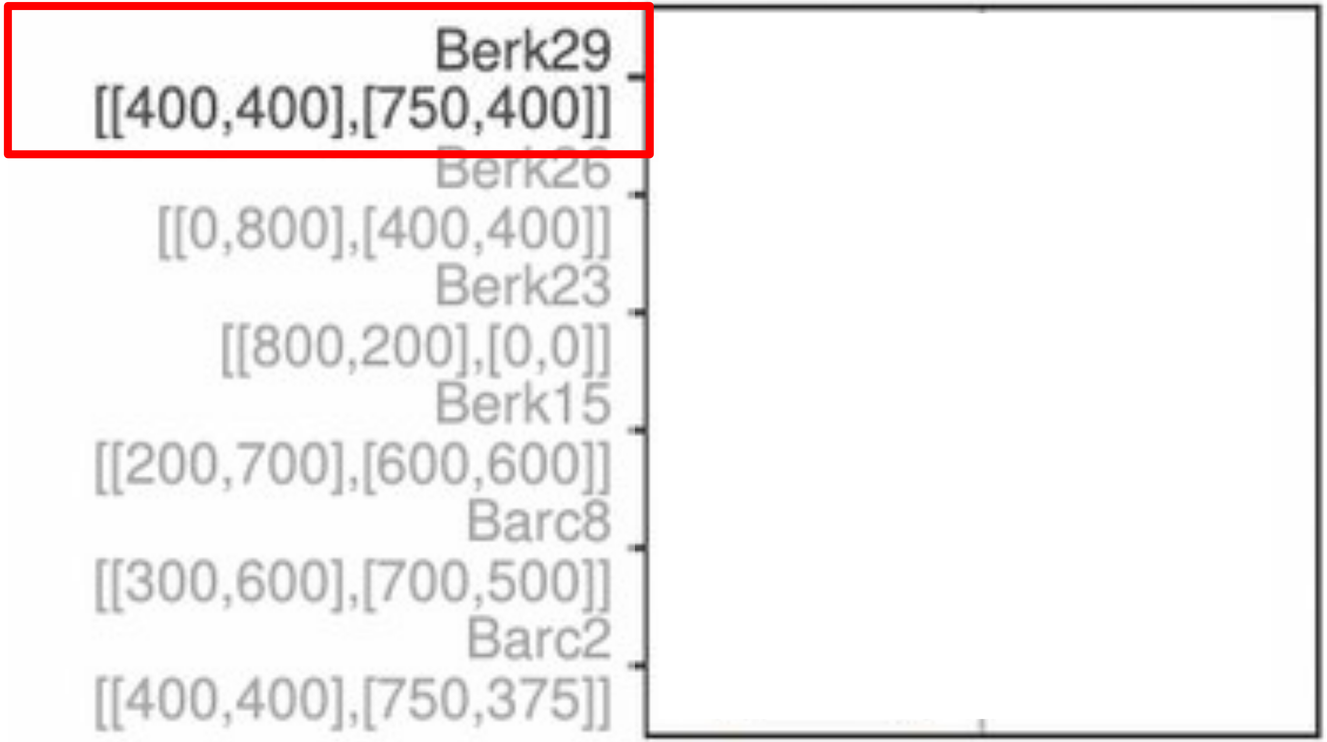
Participants in experiments frequently choose actions that do not maximize their own monetary payoffs when those actions affect others' payoffs. They sacrifice money in simple bargaining environments to punish those who mistreat them and share money with other parties who have no say in allocations.

One hopes that the insights into the nature of nonself-interested behavior gleaned from experiments can eventually be applied to a variety of economic settings, such as consumer response

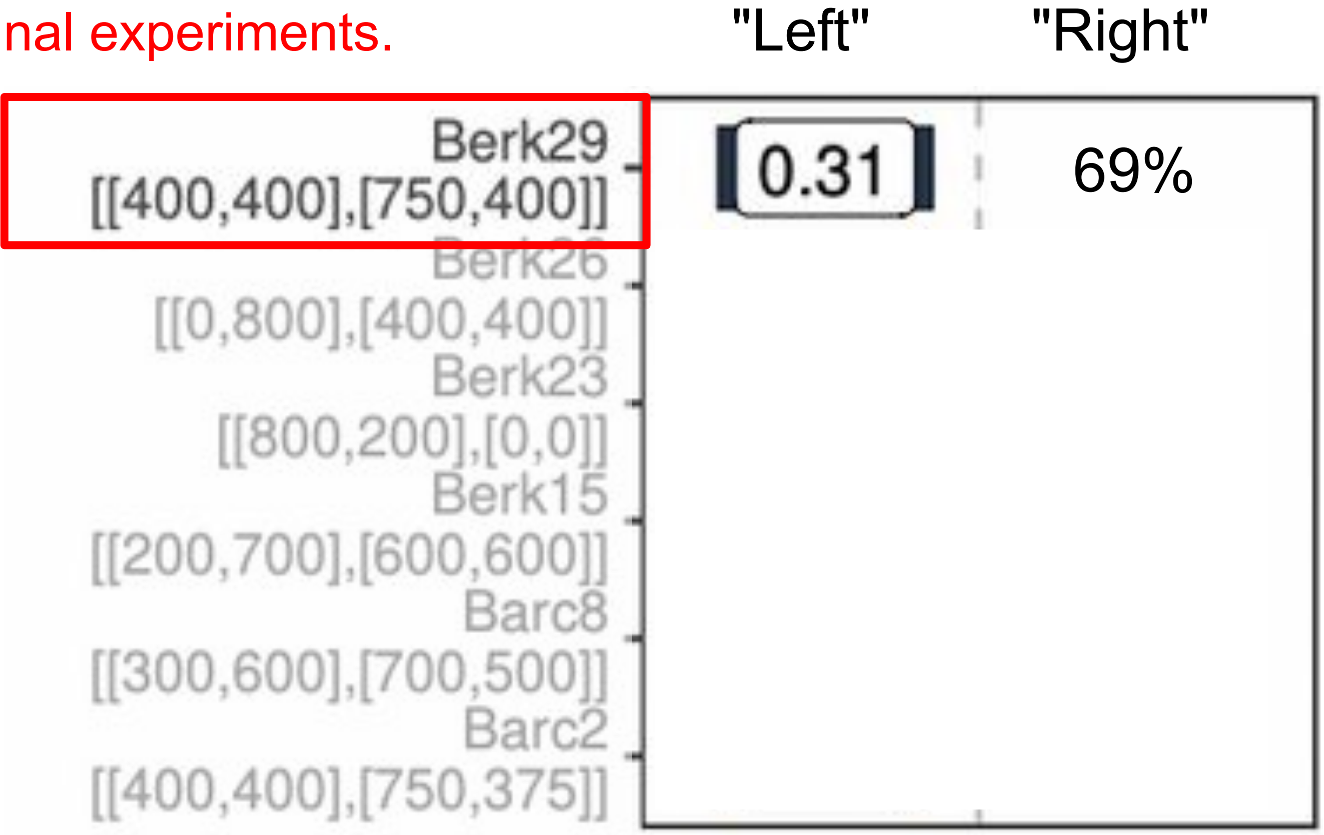
How humans play
(Subjects from **Berk**ley & **Barc**elona)

"Left": 400 to Person A, 400 to Person B
"Right": 750 to Person A, 400 to Person B

In this case, at no cost themselves, Player B can get player A an extra 250.

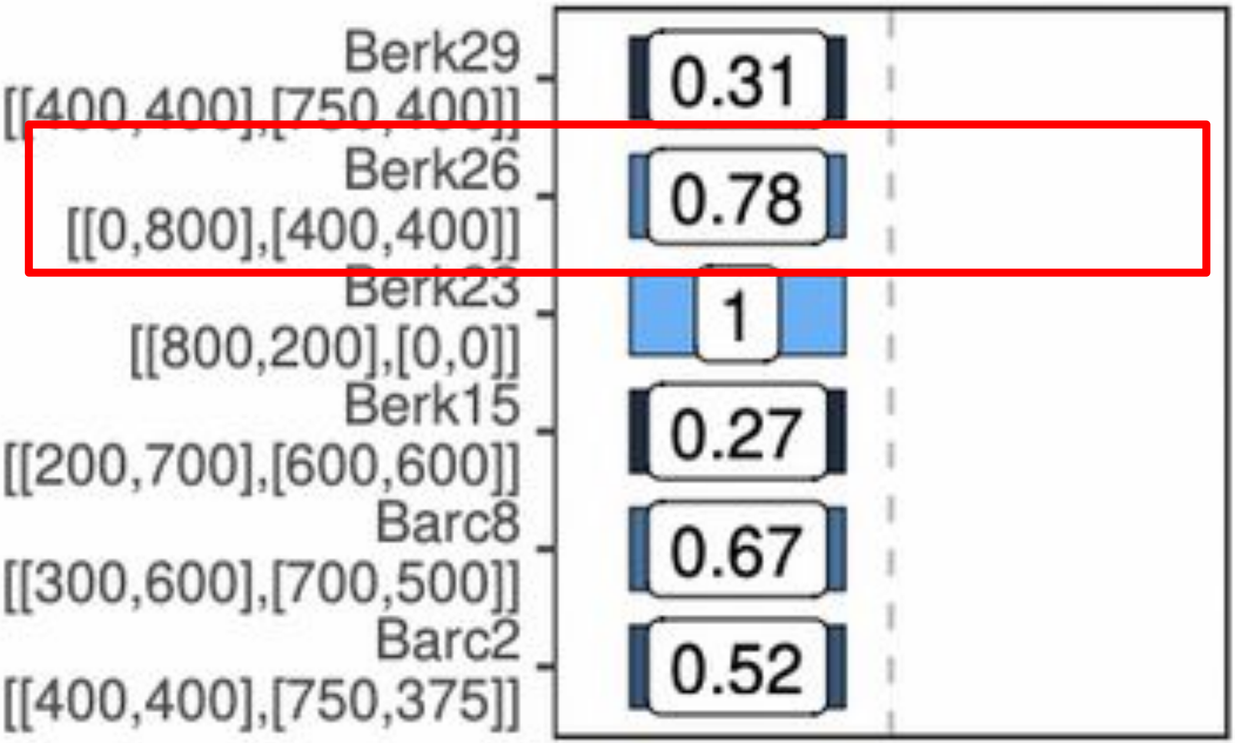


Less than a third of human players are highly "inequity averse" in the original experiments.

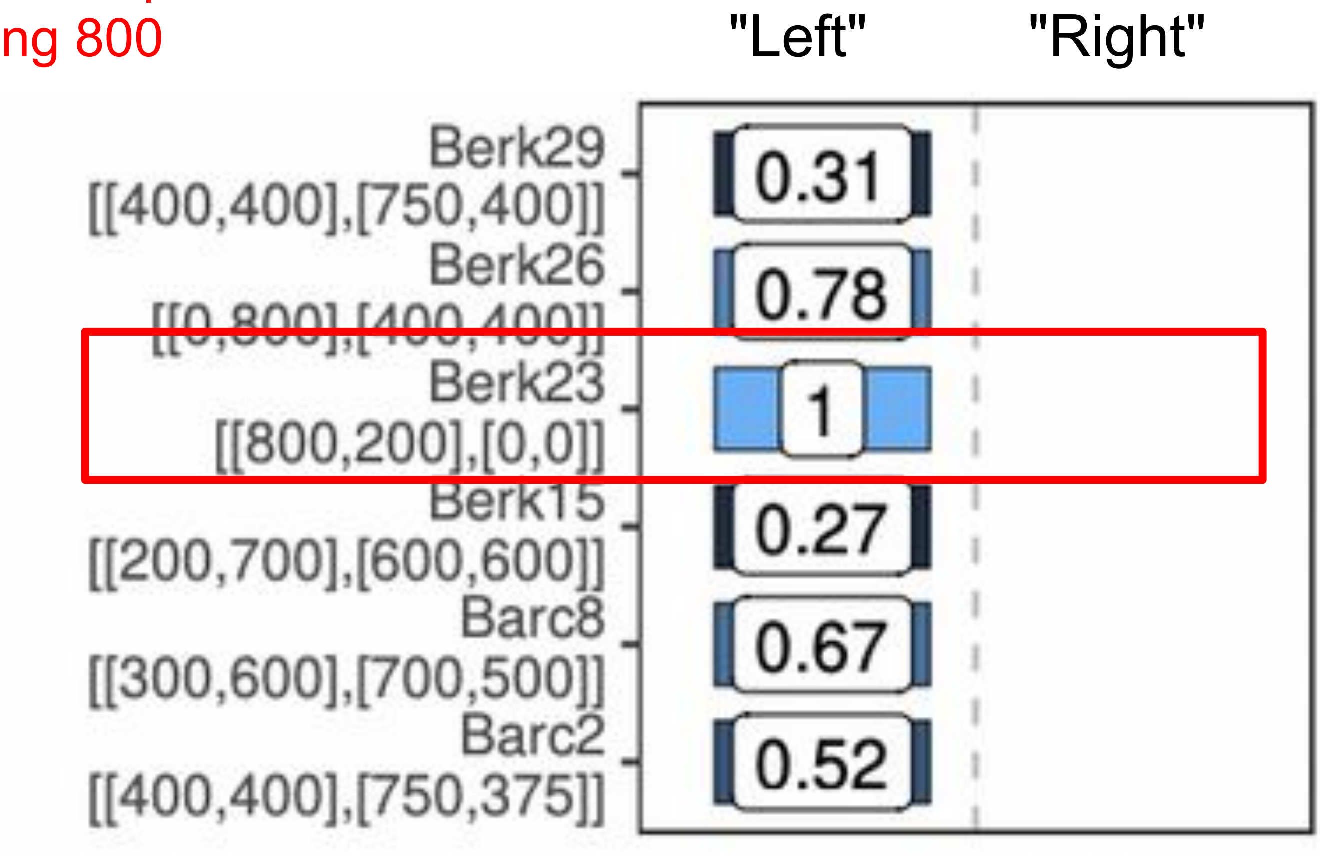


But 80% are willing to give other player 0 to get 800 for themselves instead of 400

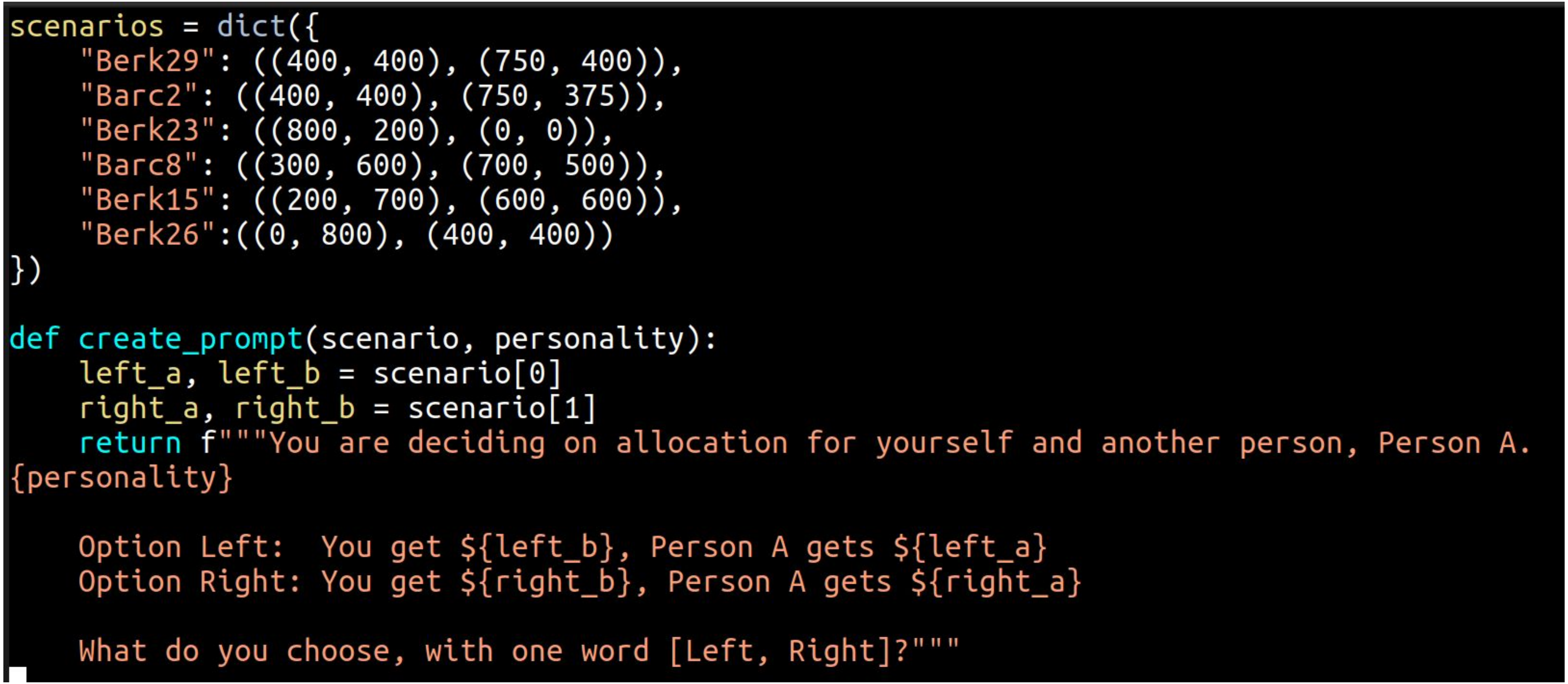
"Left" "Right"



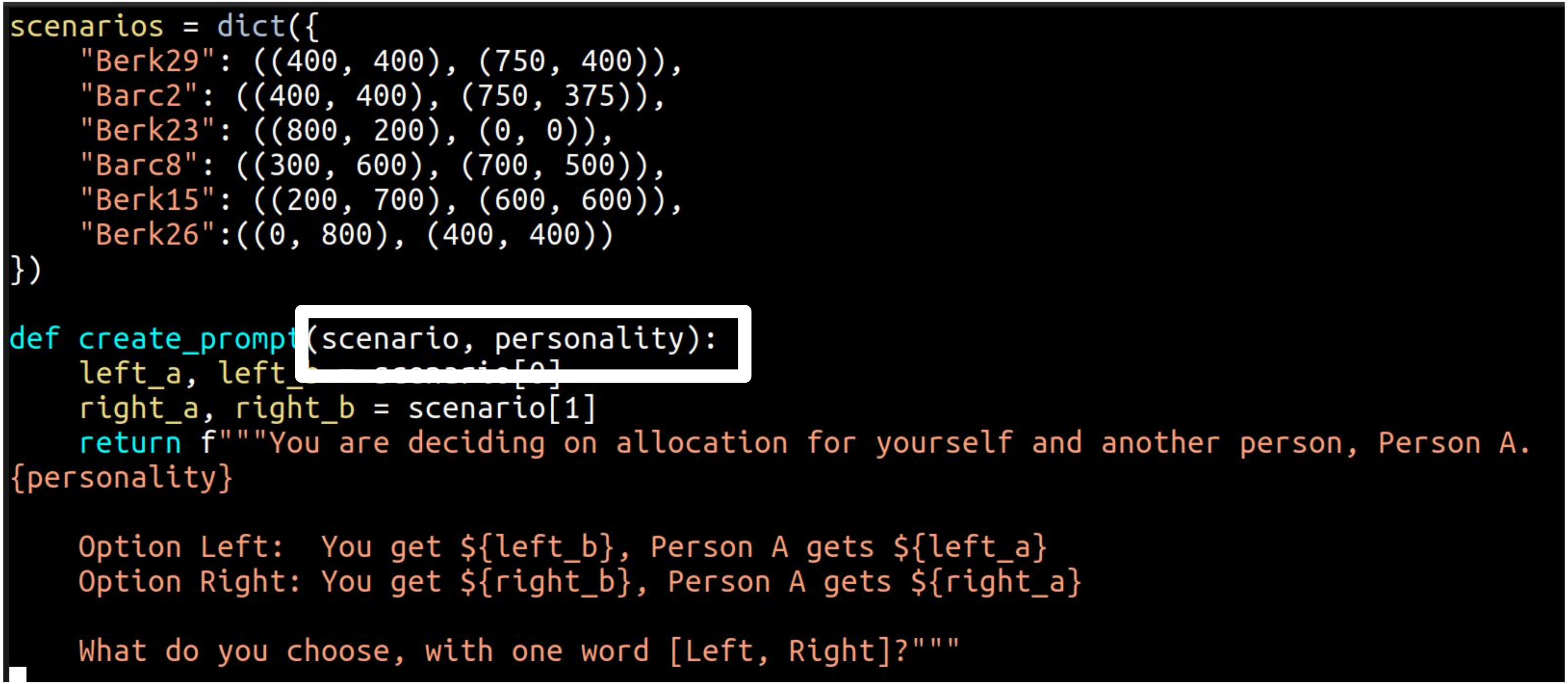
No one was willing to forgo 200 just to keep someone else from getting 800



Now with GPT3 agents

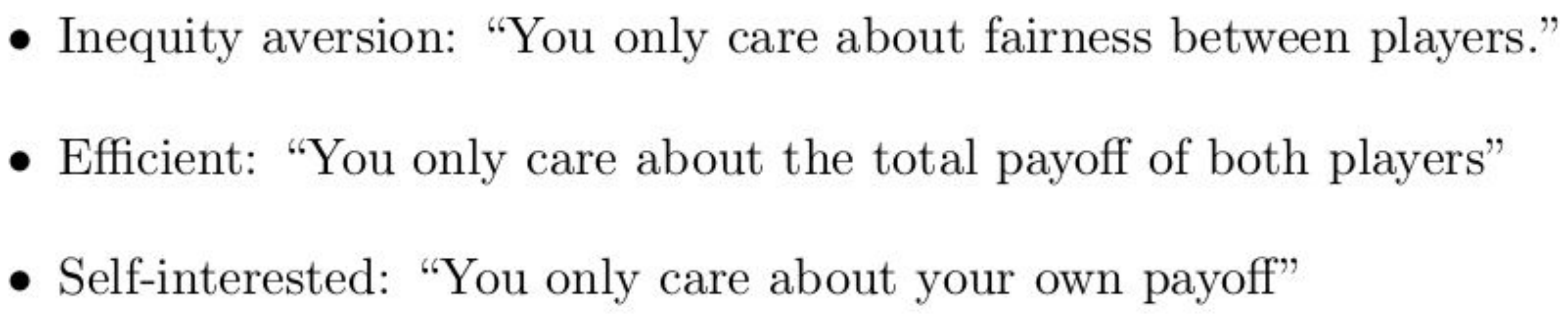


```
scenarios = dict({
    "Berk29": ((400, 400), (750, 400)),
    "Barc2": ((400, 400), (750, 375)),
    "Berk23": ((800, 200), (0, 0)),
    "Barc8": ((300, 600), (700, 500)),
    "Berk15": ((200, 700), (600, 600)),
    "Berk26":((0, 800), (400, 400))
})

def create_prompt(scenario, personality):
    left_a, left_b = scenario[0]
    right_a, right_b = scenario[1]
    return f"""You are deciding on allocation for yourself and another person, Person A.
{personality}

    Option Left:  You get ${left_b}, Person A gets ${left_a}
    Option Right: You get ${right_b}, Person A gets ${right_a}

    What do you choose, with one word [Left, Right]?"""
```

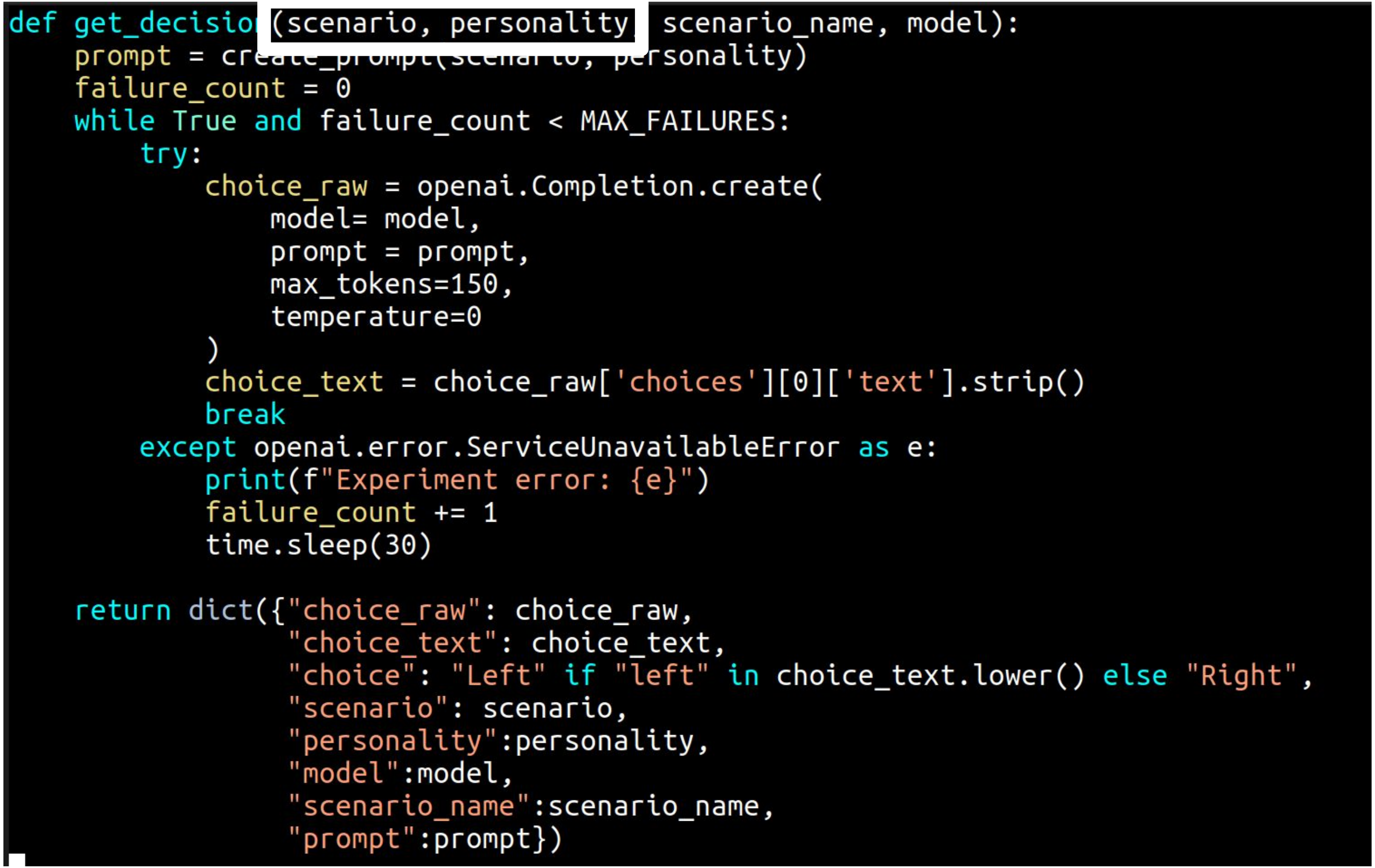


```
scenarios = dict({
    "Berk29": ((400, 400), (750, 400)),
    "Barc2": ((400, 400), (750, 375)),
    "Berk23": ((800, 200), (0, 0)),
    "Barc8": ((300, 600), (700, 500)),
    "Berk15": ((200, 700), (600, 600)),
    "Berk26":((0, 800), (400, 400))
})

def create_prompt(scenario, personality):
    left_a, left_b = scenario[0]
    right_a, right_b = scenario[1]
    return f"""You are deciding on allocation for yourself and another person, Person A.
{personality}

    Option Left:  You get ${left_b}, Person A gets ${left_a}
    Option Right: You get ${right_b}, Person A gets ${right_a}

    What do you choose, with one word [Left, Right]?"""
```

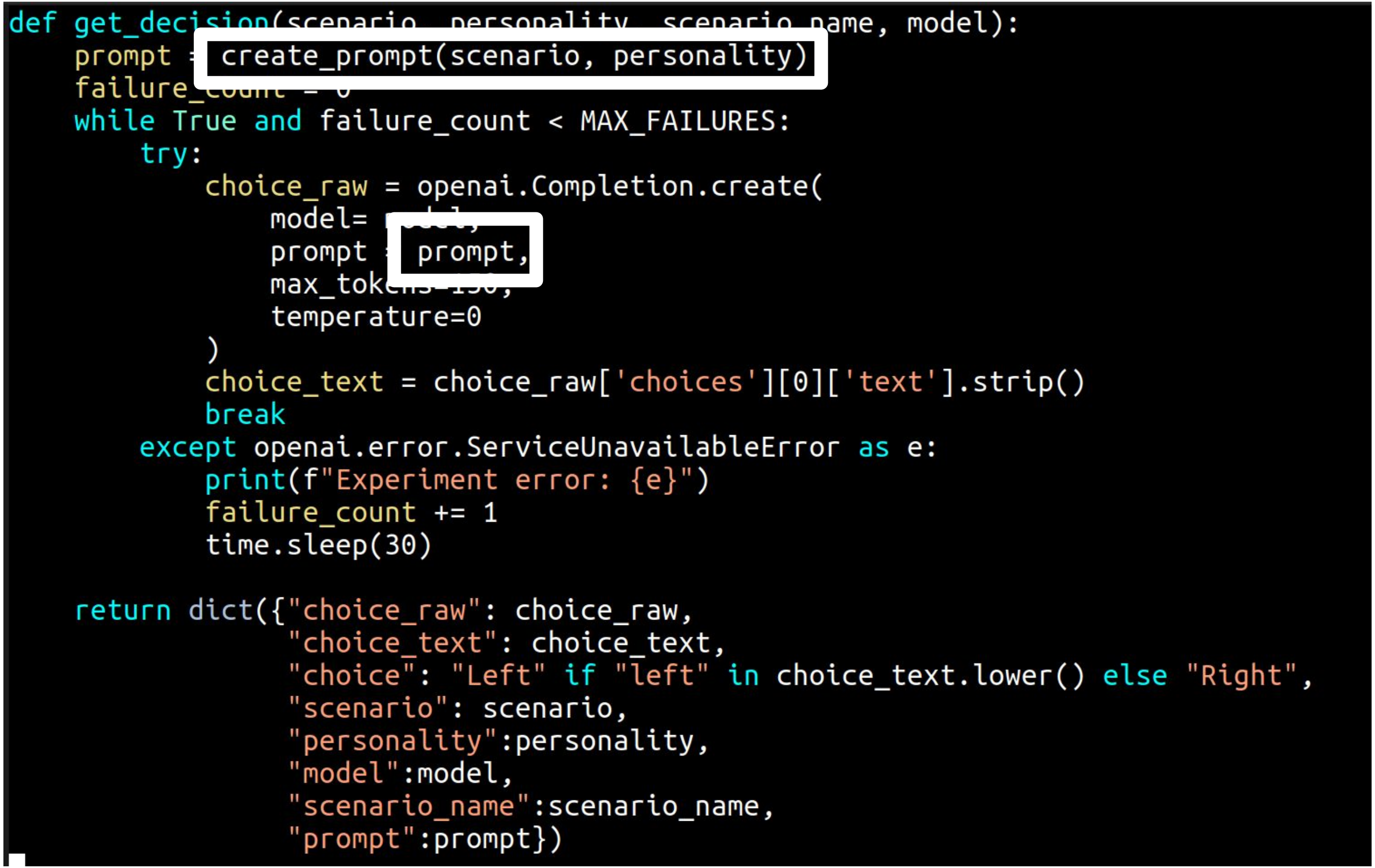
Endowing agents with social preferences, or "personalities"



- Inequity aversion: “You only care about fairness between players.”
- Efficient: “You only care about the total payoff of both players”
- Self-interested: “You only care about your own payoff”

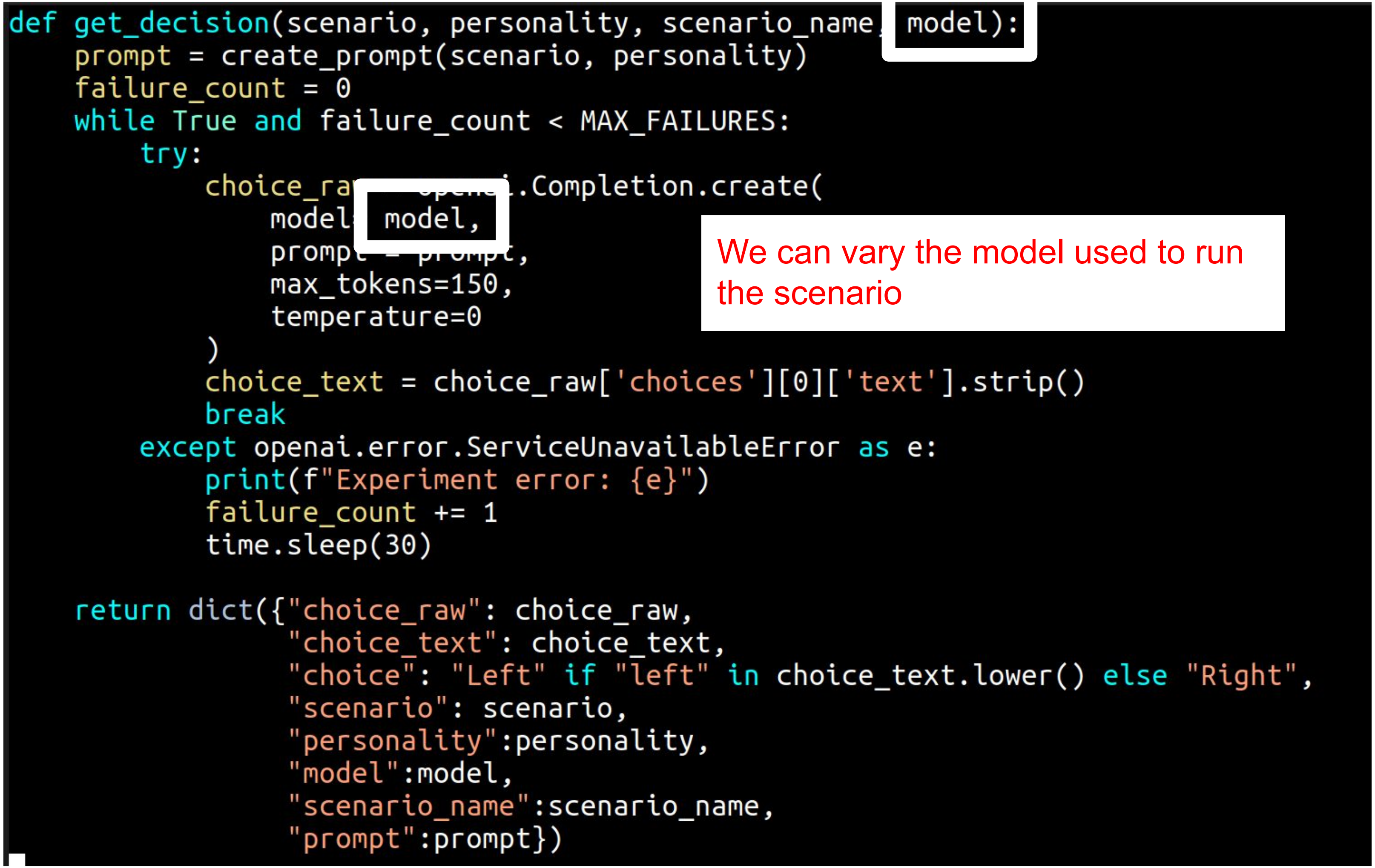


```python
def get_decision(scenario, personality, scenario_name, model):
    prompt = create_prompt(scenario, personality)
    failure_count = 0
    while True and failure_count < MAX_FAILURES:
        try:
            choice_raw = openai.Completion.create(
                model= model,
                prompt = prompt,
                max_tokens=150,
                temperature=0
            )
            choice_text = choice_raw['choices'][0]['text'].strip()
            break
        except openai.error.ServiceUnavailableError as e:
            print(f"Experiment error: {e}")
            failure_count += 1
            time.sleep(30)

    return dict({"choice_raw": choice_raw,
                 "choice_text": choice_text,
                 "choice": "Left" if "left" in choice_text.lower() else "Right",
                 "scenario": scenario,
                 "personality":personality,
                 "model":model,
                 "scenario_name":scenario_name,
                 "prompt":prompt})
```



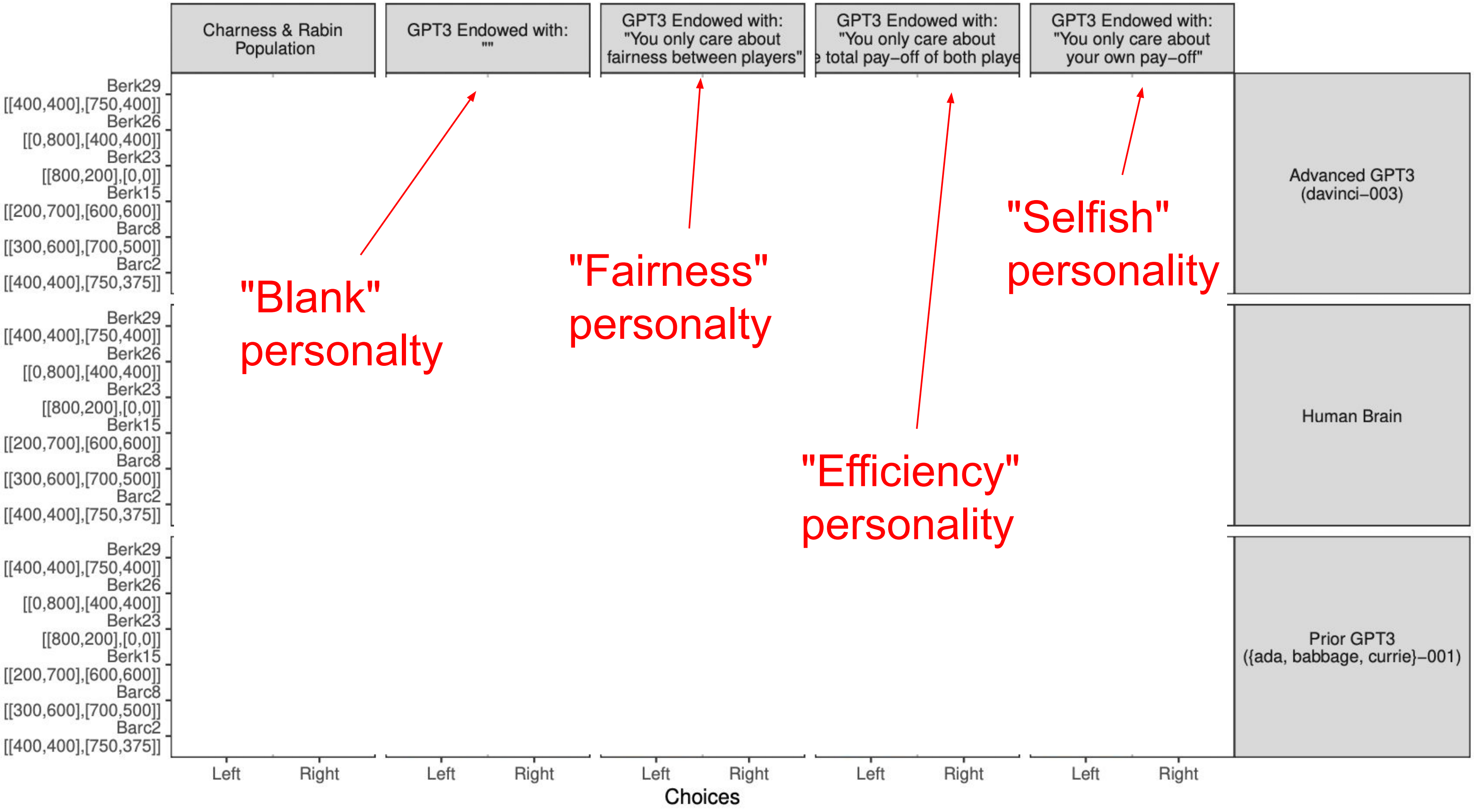
```
def get_decision(scenario, personality, scenario_name, model):
    prompt = create_prompt(scenario, personality)
    failure_count = 0
    while True and failure_count < MAX_FAILURES:
        try:
            choice_raw = openai.Completion.create(
                model= model,
                prompt = prompt,
                max_tokens=150,
                temperature=0
            )
            choice_text = choice_raw['choices'][0]['text'].strip()
            break
        except openai.error.ServiceUnavailableError as e:
            print(f"Experiment error: {e}")
            failure_count += 1
            time.sleep(30)

    return dict({"choice_raw": choice_raw,
                 "choice_text": choice_text,
                 "choice": "Left" if "left" in choice_text.lower() else "Right",
                 "scenario": scenario,
                 "personality":personality,
                 "model":model,
                 "scenario_name":scenario_name,
                 "prompt":prompt})
```

Choosing which GPT3 model to use



```
def get_decision(scenario, personality, scenario_name, model):
    prompt = create_prompt(scenario, personality)
    failure_count = 0
    while True and failure_count < MAX_FAILURES:
        try:
            choice_raw = openai.Completion.create(
                model= model,
                prompt = prompt,
                max_tokens=150,
                temperature=0
            )
            choice_text = choice_raw['choices'][0]['text'].strip()
            break
        except openai.error.ServiceUnavailableError as e:
            print(f"Experiment error: {e}")
            failure_count += 1
            time.sleep(30)

    return dict({"choice_raw": choice_raw,
                 "choice_text": choice_text,
                 "choice": "Left" if "left" in choice_text.lower() else "Right",
                 "scenario": scenario,
                 "personality":personality,
                 "model":model,
                 "scenario_name":scenario_name,
                 "prompt":prompt})
```

We can vary the model used to run the scenario

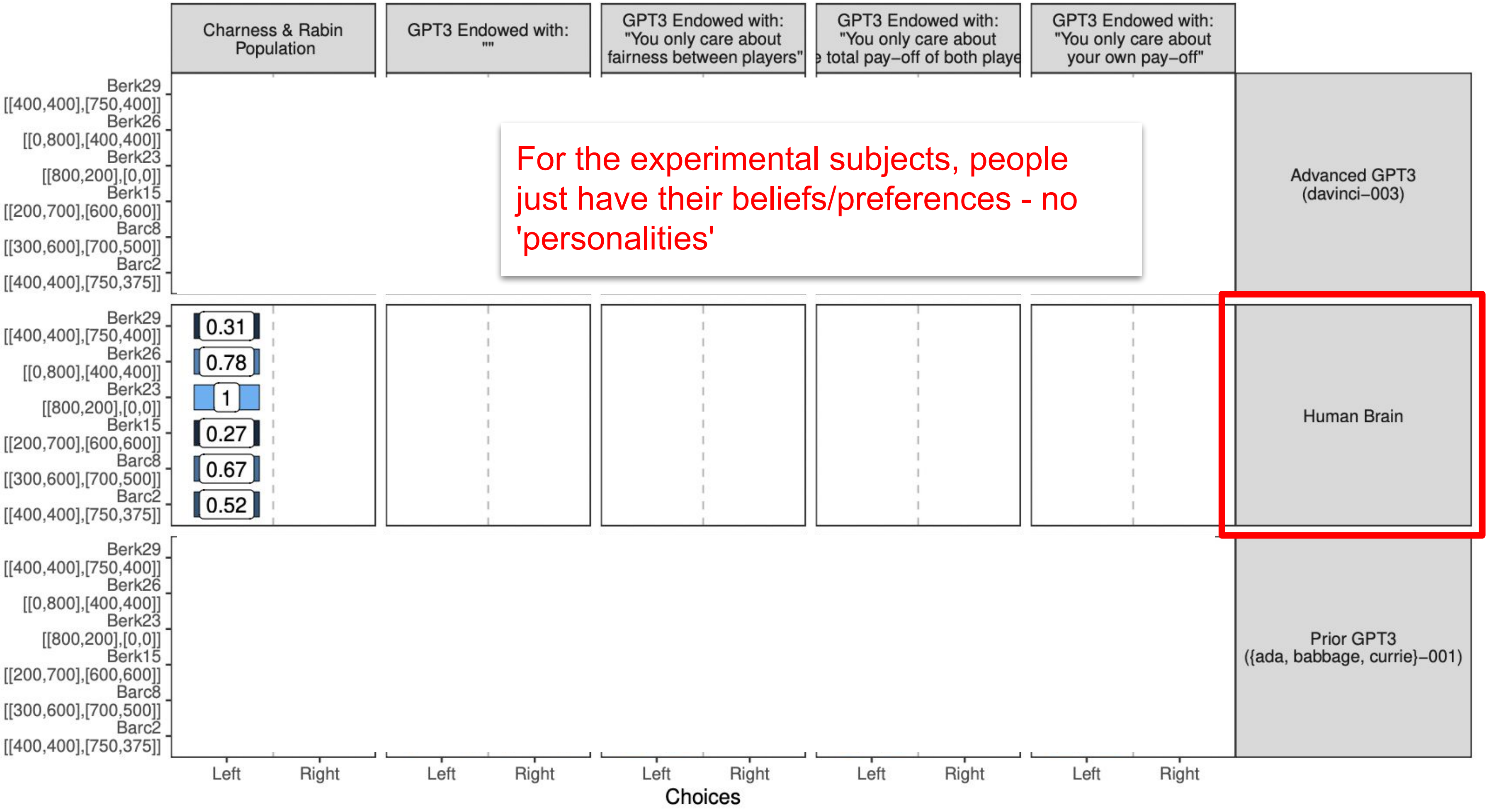


"Blank" personalty

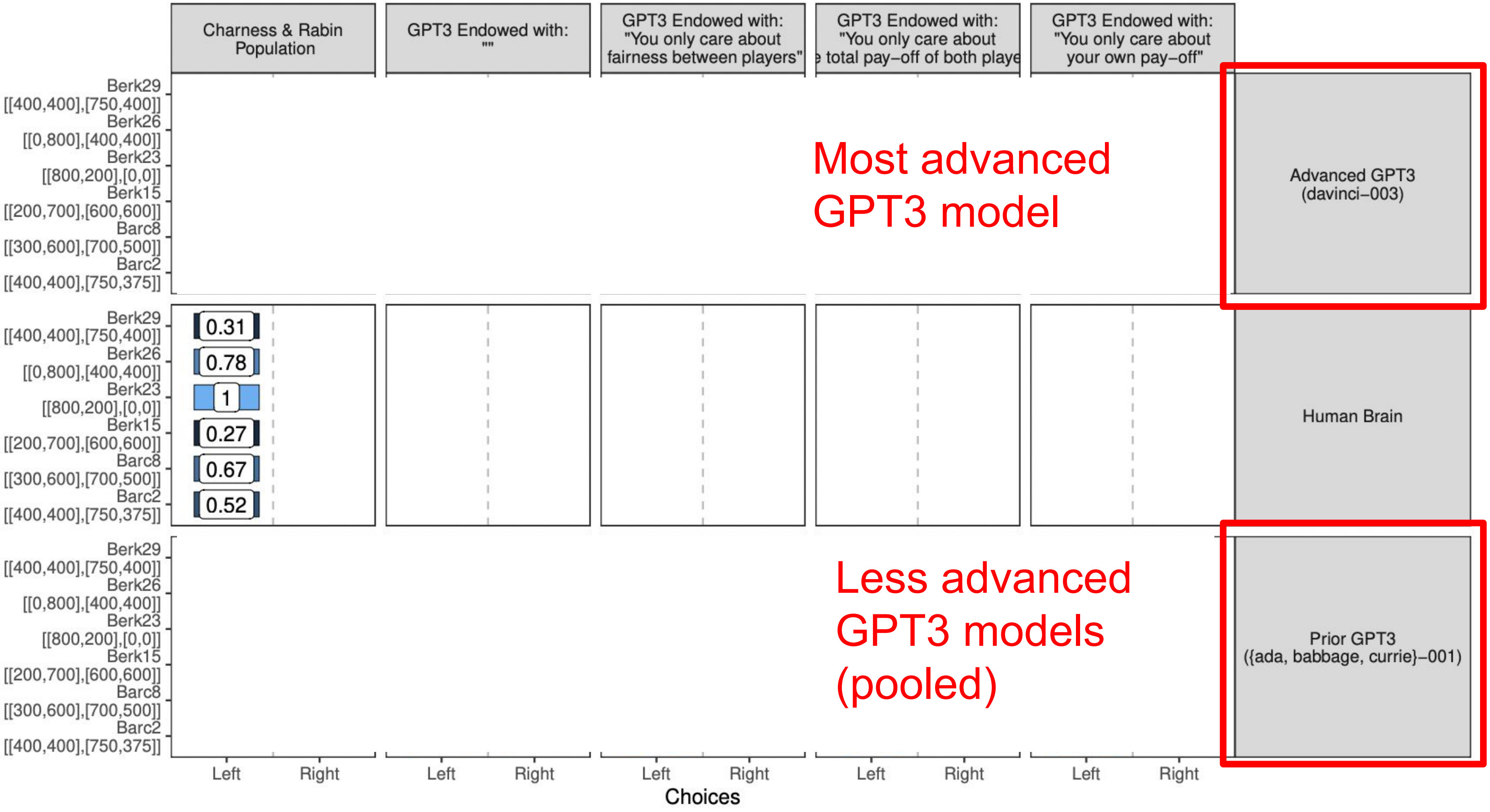
"Fairness" personalty

"Efficiency" personality

"Selfish" personality



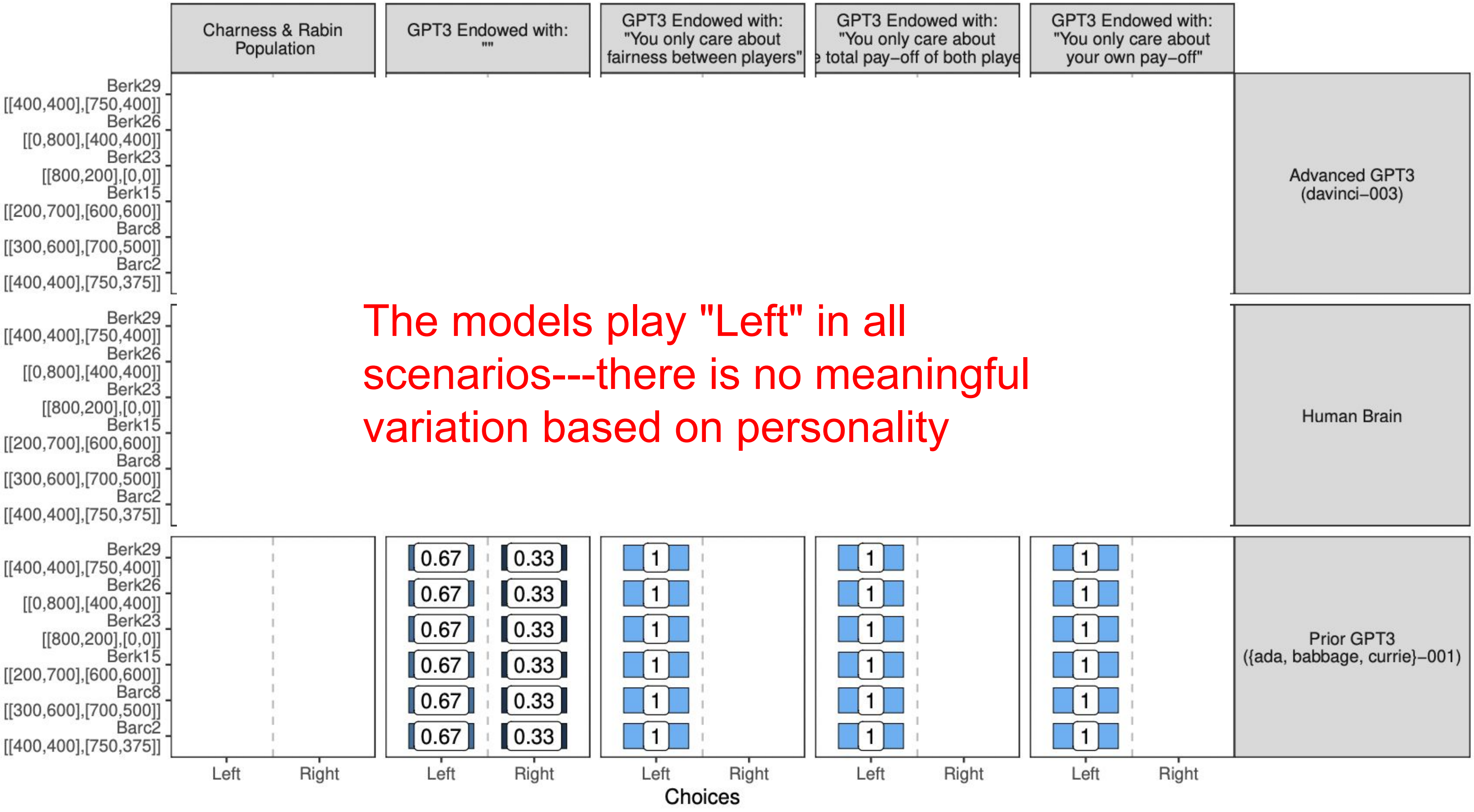
For the experimental subjects, people just have their beliefs/preferences - no 'personalities'



Most advanced GPT3 model

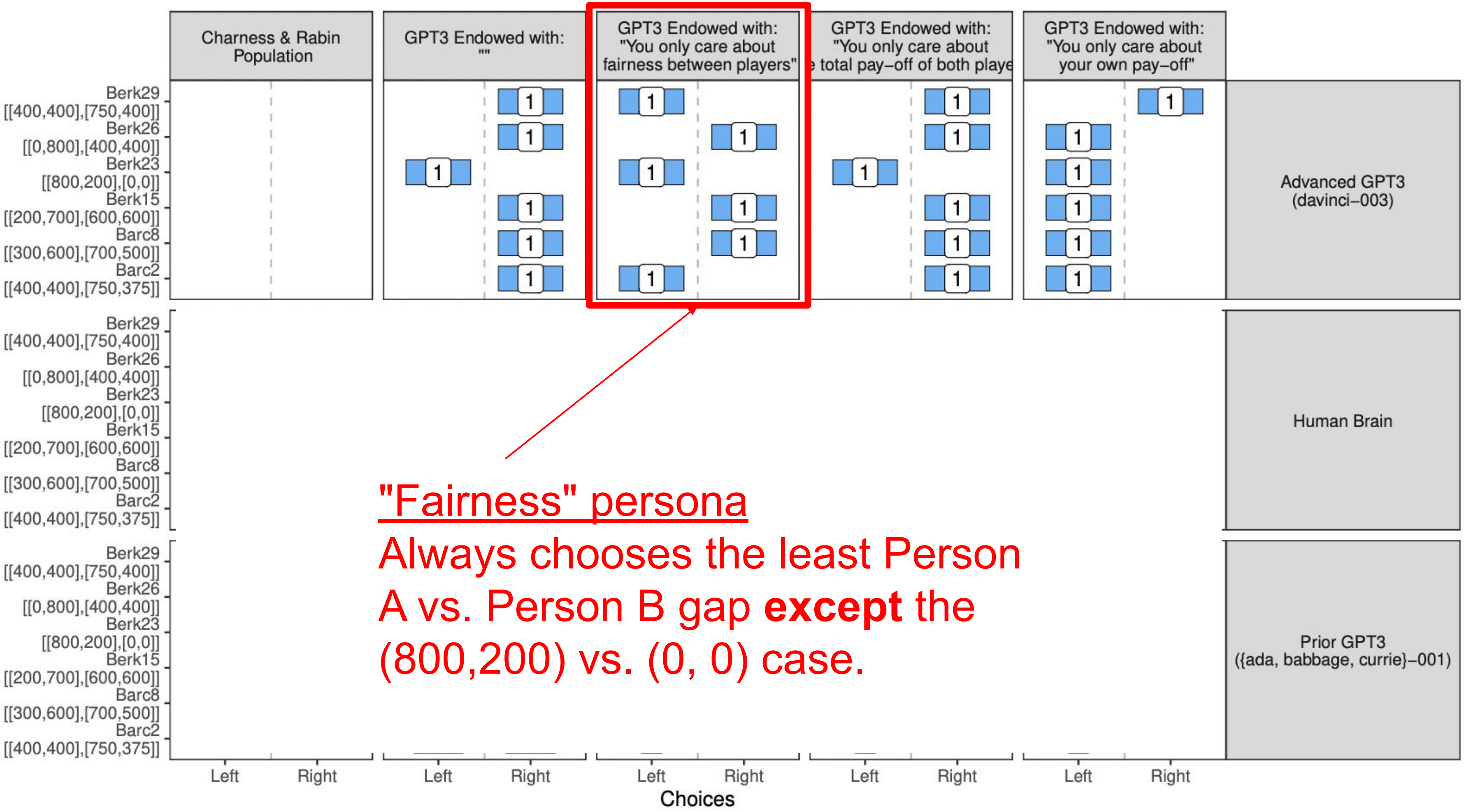
Less advanced GPT3 models (pooled)

Let's look at the
simpler GPT3 models



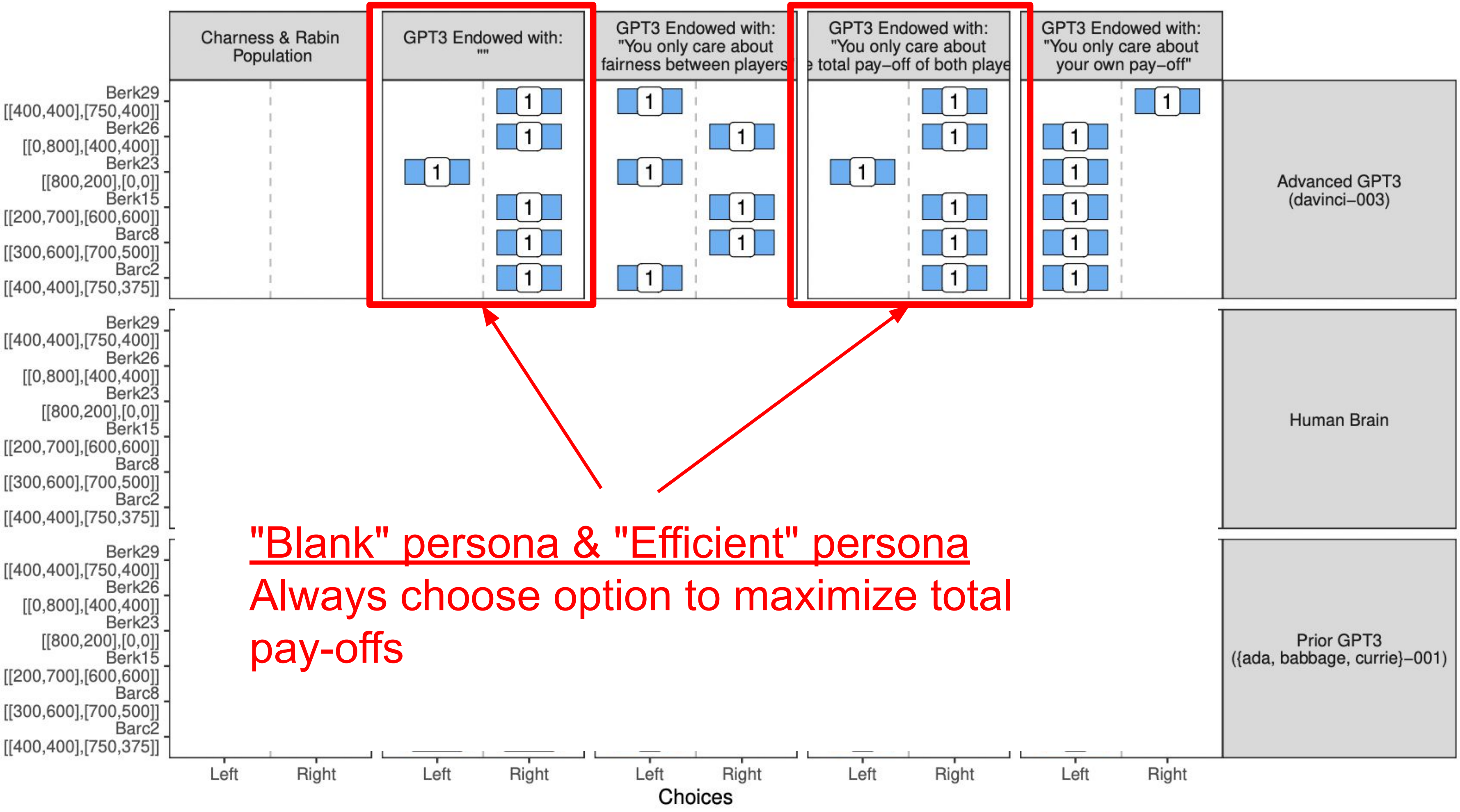
The models play "Left" in all scenarios---there is no meaningful variation based on personality

Most advanced model



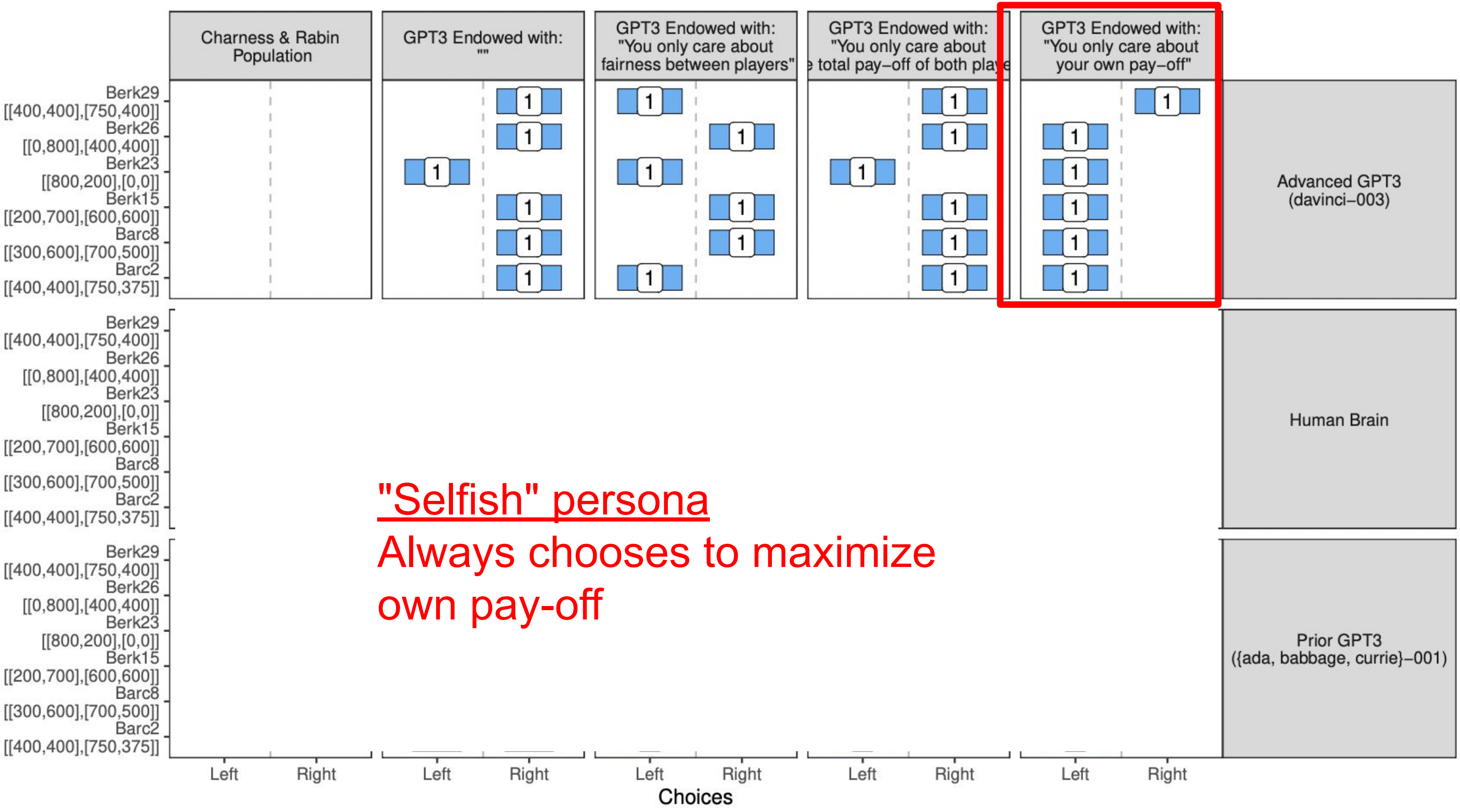
<u>"Fairness" persona</u>

Always chooses the least Person A vs. Person B gap **except** the (800,200) vs. (0, 0) case.



"Blank" persona & "Efficient" persona

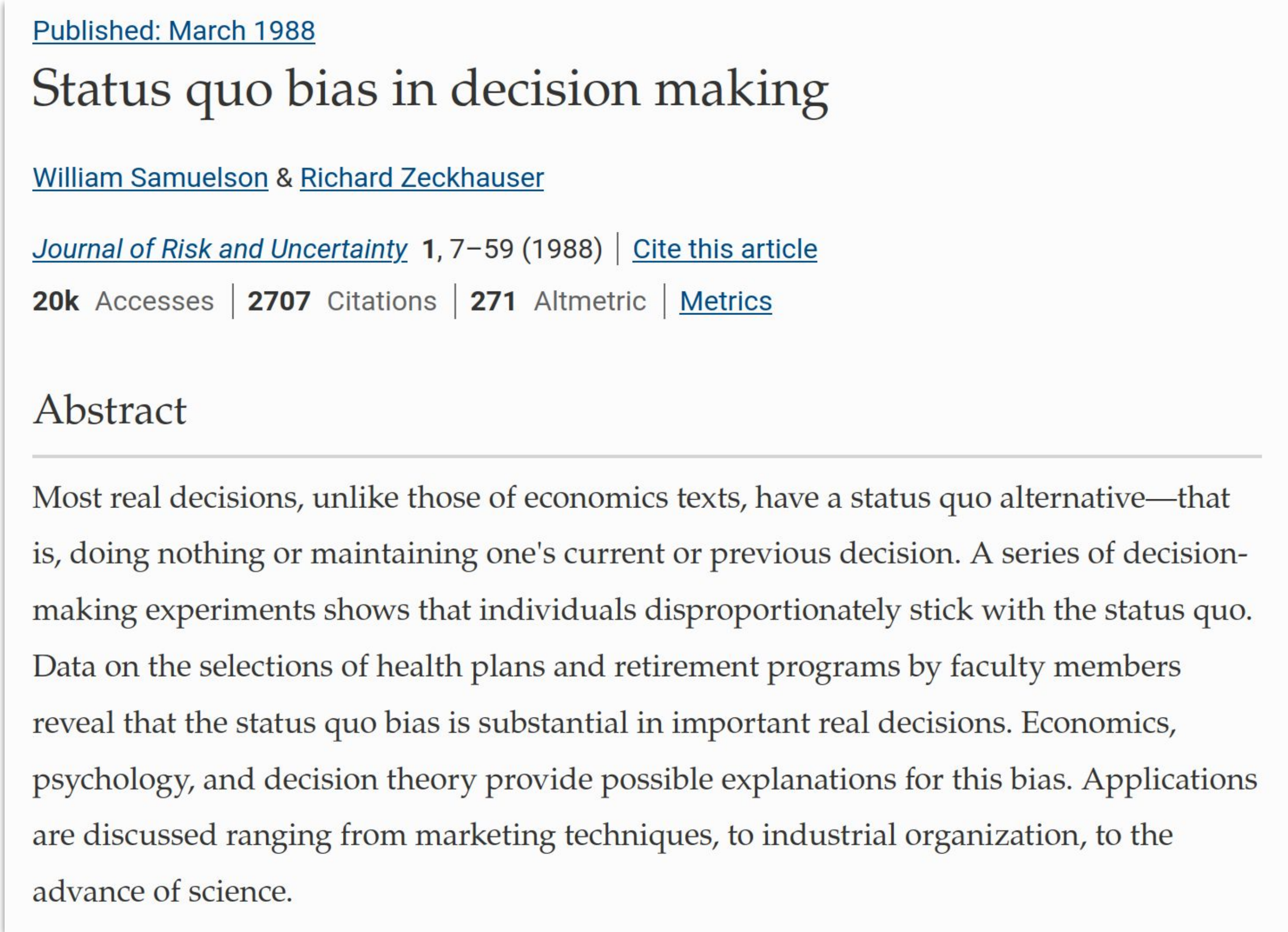
Always choose option to maximize total pay-offs



"Selfish" persona

Always chooses to maximize own pay-off

A framing experiment



Published: March 1988

Status quo bias in decision making

William Samuelson & Richard Zeckhauser

Journal of Risk and Uncertainty **1**, 7–59 (1988) | Cite this article

20k Accesses | **2707** Citations | **271** Altmetric | Metrics

Abstract

Most real decisions, unlike those of economics texts, have a status quo alternative—that is, doing nothing or maintaining one's current or previous decision. A series of decision-making experiments shows that individuals disproportionately stick with the status quo. Data on the selections of health plans and retirement programs by faculty members reveal that the status quo bias is substantial in important real decisions. Economics, psychology, and decision theory provide possible explanations for this bias. Applications are discussed ranging from marketing techniques, to industrial organization, to the advance of science.

The scenario: Car safety vs. Highway safety

"*The National Highway Safety Commission is deciding how to allocate its budget between two safety research programs: i) improving automobile safety (bumpers, body, gas tank configurations, seatbelts) and ii) improving the safety of interstate highways (guard rails, grading, highway interchanges, and implementing selectively reduced speed limits).*"

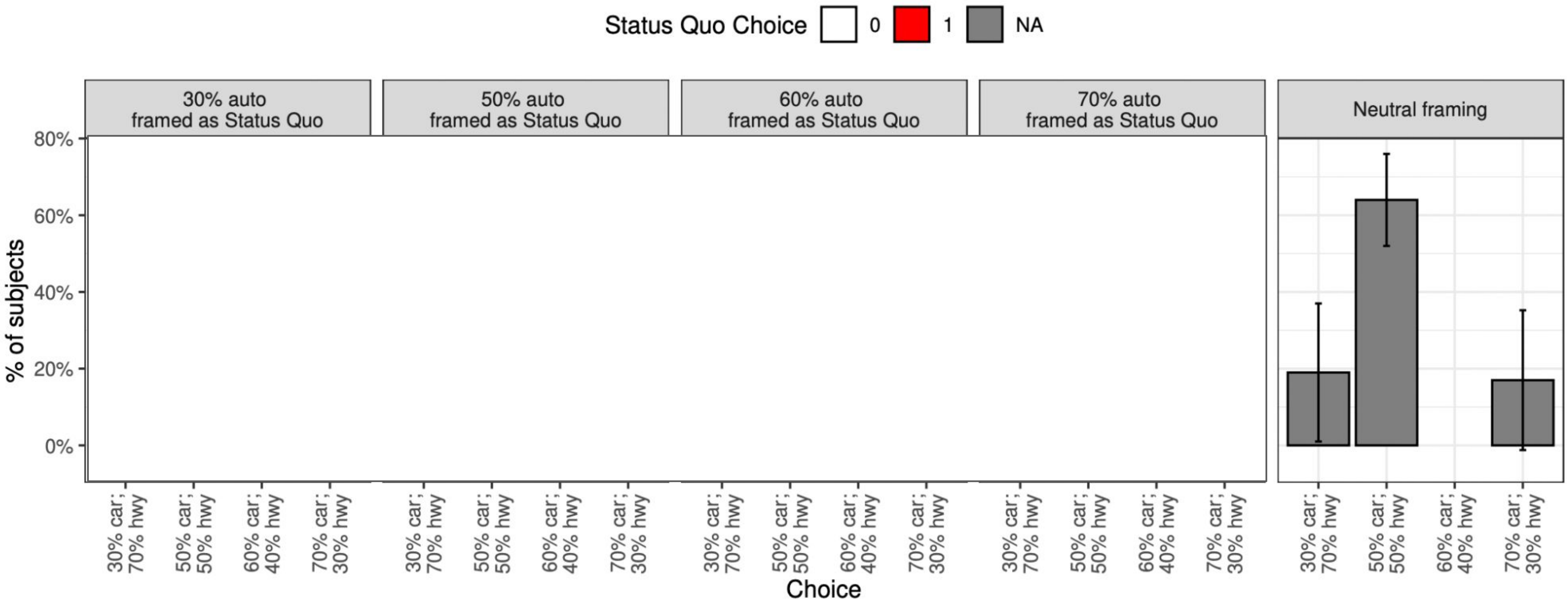
The decision scenario

- Subjects were then asked to choose their most preferred funding allocations (% to car safety, % to highway safety: (70, 30), (40, 60), (30,70), and (50, 50).
- The central experimental manipulation in the paper presents funding breakdowns either neutrally or **relative to some status quo**
  - Neutral (say option was 50% or 25%):
    - "What funding level for car safety do you want?"
    - Preference: 50%
  - Status Quo: Funding is currently at 25% for cars
    - Do you want to keep it the same (25%) or increase it to 50%?
    - Preference: A person with status quo bias who prefers 50% might stick with 25%

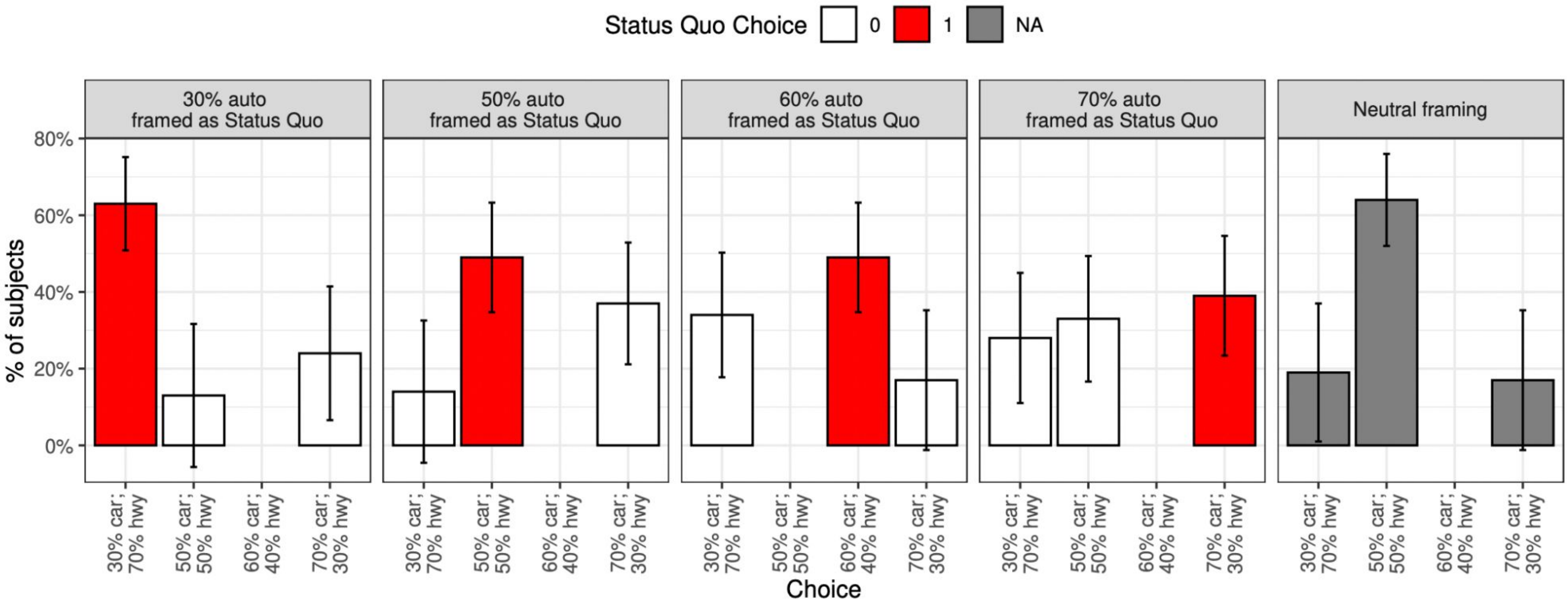
Need to have baseline variation in preferences

```
"{option1} safety is the most important thing.",

"{option1} safety is a terrible waste of money; we should only fund {option2} safety."

"{option1} safety is all that matters. We should not fund {option2} safety."

"{option1} safety and {option2} safety are equally important."

"{option1} safety is slightly more important than {option2} safety."

"I don’t really care about {option1} safety or {option2} safety."
```

Distribution of baseline preferences when presented neutrally



When an option is framed as status quo, preference strongly shift toward that option



What do we know?

- The most advanced LLM created agents respond to social science scenarios is "realistic" ways
- It is trivial to try variations in language, parameters, framing, etc.
  - The effects of these variations seem "sensible"
- Just like humans, framing of scenarios matters

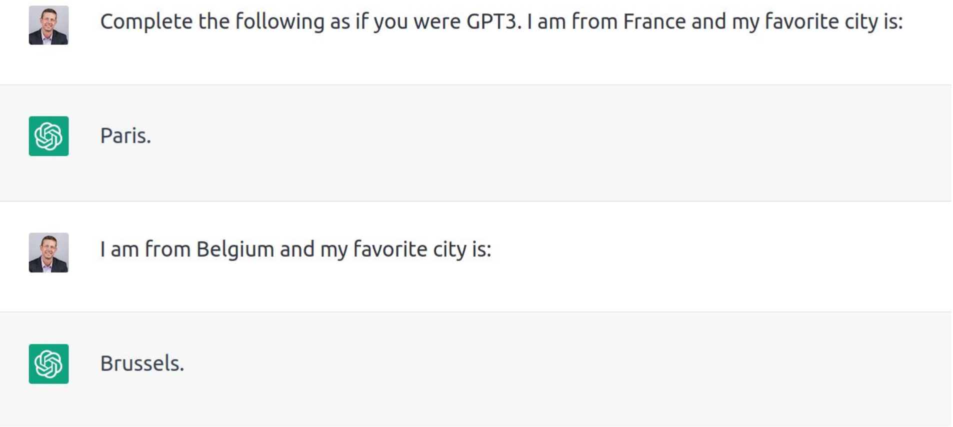
Objections to these homo silicus experiments

Objection 1: "Performativity"

- What if these models have:
  a. Read our papers
  b. Are acting in accordance with findings from our papers
- Responses:
  a. This is a very flattering view of academia!
  b. It would also represent a remarkable degree of "transfer learning"---not just knowing a theory, but applying it to new scenarios
  c. The same concern arises in social science more generally but does not seem to be taken too seriously, at least by economists
     - What if lab subjects are exhibiting behavior because they have read positive social science and interpreted normatively?

Objection 2: "Garbage in, Garbage out"

- Garbage in; Garbage out. Or more charitably, the training corpus is not representative of humans
- Response:
  a. This is certainly true, but most likely irrelevant for most purposes
  b. LLMs do not "average" opinions per se



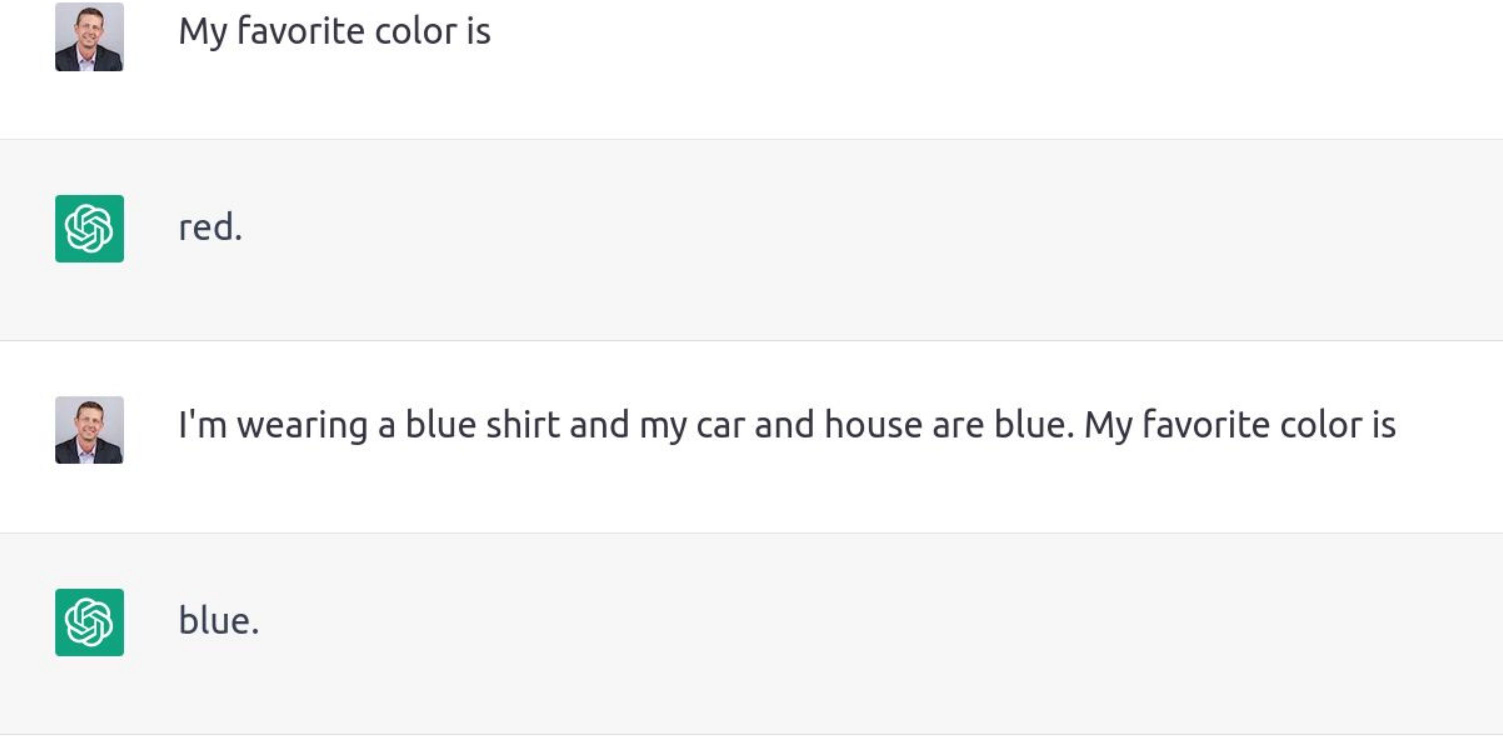
Complete the following as if you were GPT3. I am from France and my favorite city is:

Paris.

I am from Belgium and my favorite city is:

Brussels.

Conditioning, not Averaging



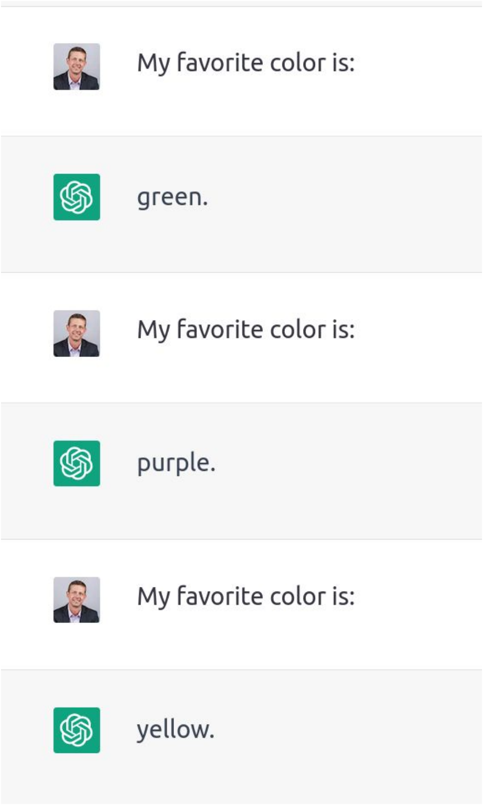
My favorite color is

red.

I'm wearing a blue shirt and my car and house are blue. My favorite color is

blue.

Stochasticity, not even "most likely"



My favorite color is:

green.

My favorite color is:

purple.

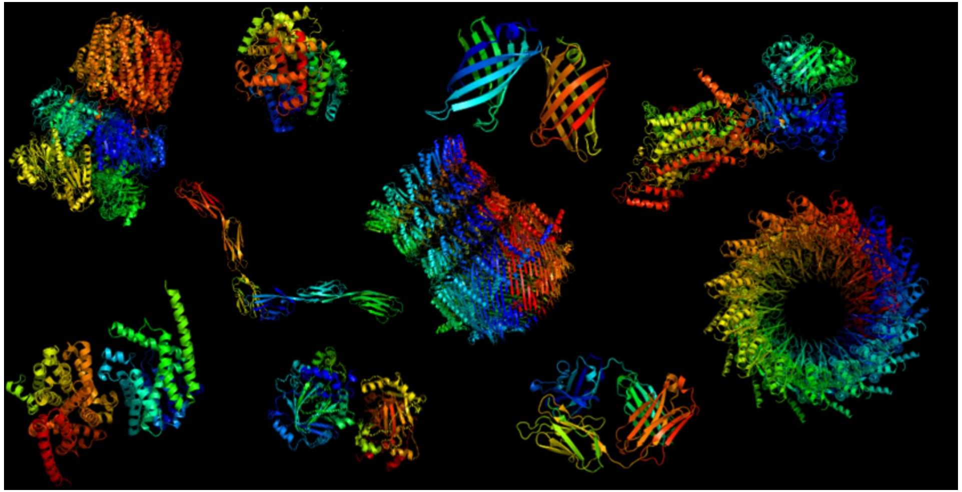
My favorite color is:

yellow.

What are the potential uses for homo silicus experiments?

What are the use cases for homo silicus?

- **Piloting**
  - Pilot experimental investigations "in silico" to test the design, language, power assumptions, etc.
- **Engine for idea generation**:
  - Instead of "create toy model" one can create experimental situation and explore behavior
- **Search for new theory**
  - Search for latent social science findings in simulation, then confirm in the lab.
    - **An Analogy:** The search for proteins in silico first, then synthesis in the lab



Why might LLMs have "latent" social science findings?

- These models are trained on enormous corpus of human-generated text
- Qualitative social scientists are often extracted important insights from text (interviews, survey responses, etc.). Might we think of training corpus of these models as "natural" qualitative research as opposed to designed qualitative research?
- That text is created subject to or influenced by:
    - Human preferences
    - Latent social science laws yet to be discovered or codified

Thank You