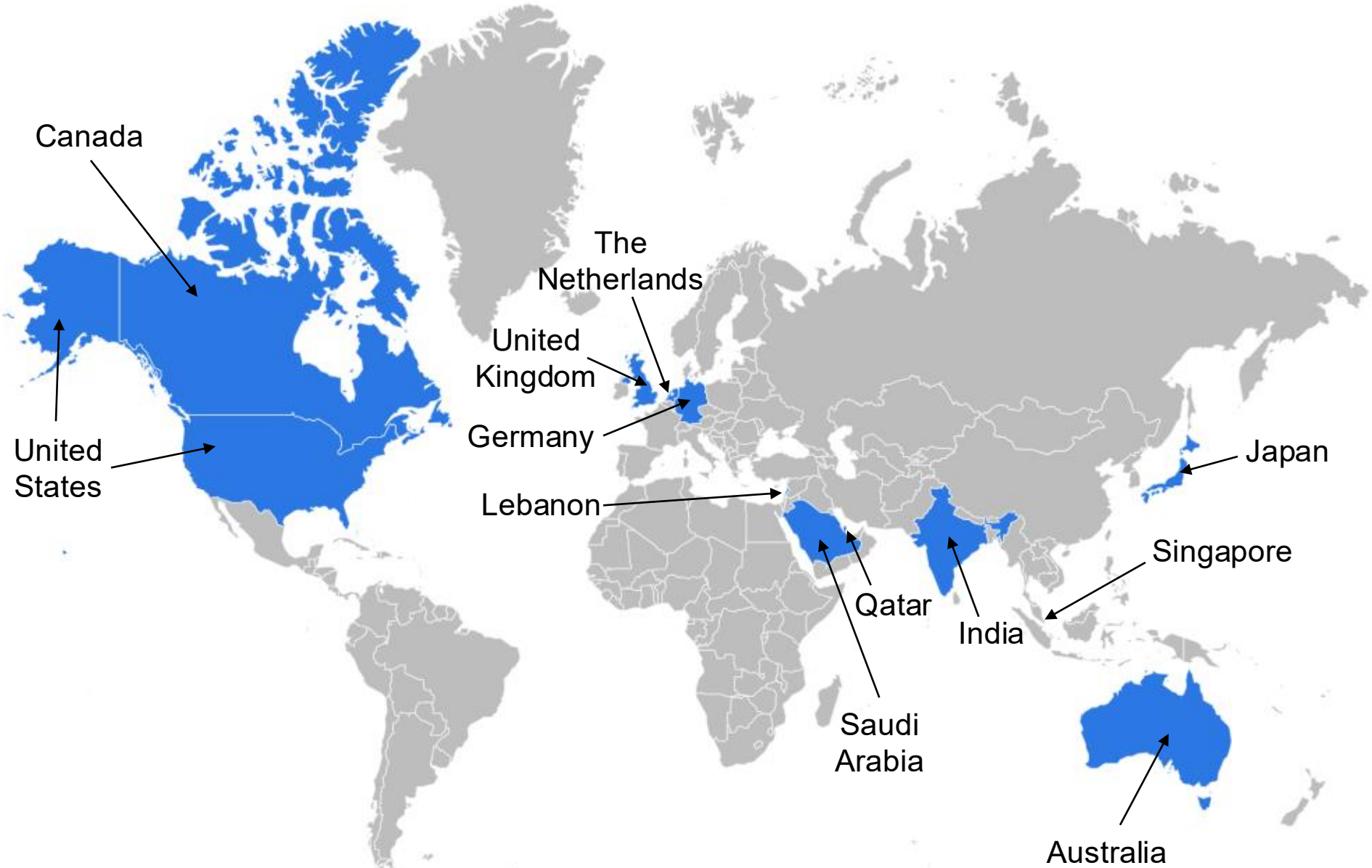
The Challenges of Behavioral Welfare Economics

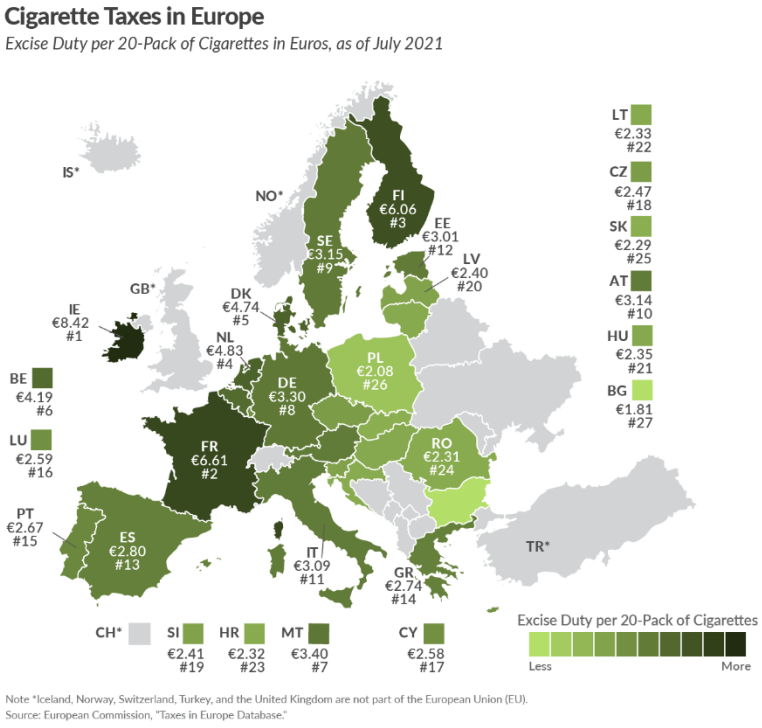
Prof. B. Douglas Bernheim, Stanford University

August 2025

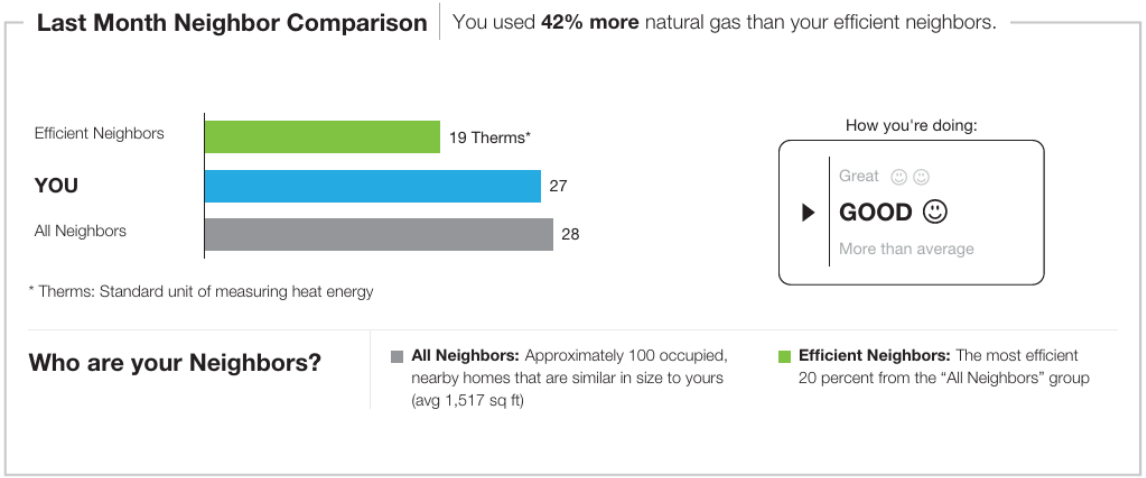
Countries with major behavioral public policy initiatives



Examples of behavioral interventions



Examples of behavioral interventions



Introduction

- The evaluation of such policies is the domain of ***Behavioral Welfare Economics (BWE)***

- Standard ***Welfare Economics*** determines whether a policy is good or bad for an individual by asking what they would choose for themselves

- In addition to offering many insights concerning the *positive* effects of public policies, ***Behavioral Economics*** challenges the foundations of Standard Welfare Economics.

- BWE seeks to either fix or replace the standard approach to evaluating economic well-being.

Introduction

- This talk: a broad, highly conceptual overview of challenges facing BWE, and their solutions.

- Focus is on the assessment of an individual's well-being, rather than on aggregation.

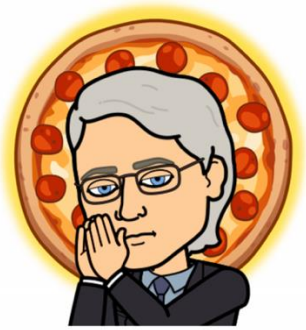
Standard Welfare Economics

Standard Welfare Economics

The planner's task

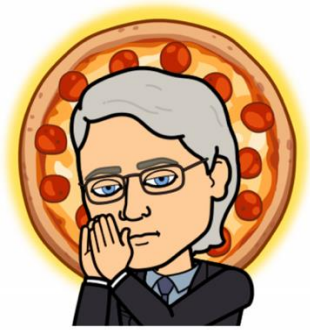
Standard Welfare Economics

The planner's task



Standard Welfare Economics

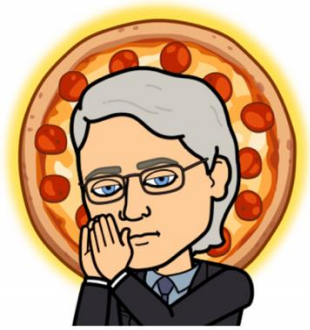
The planner's task



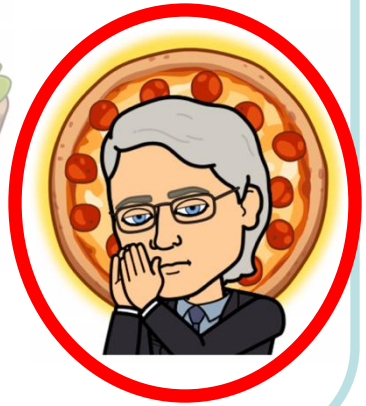
My task

Standard Welfare Economics

The planner's task

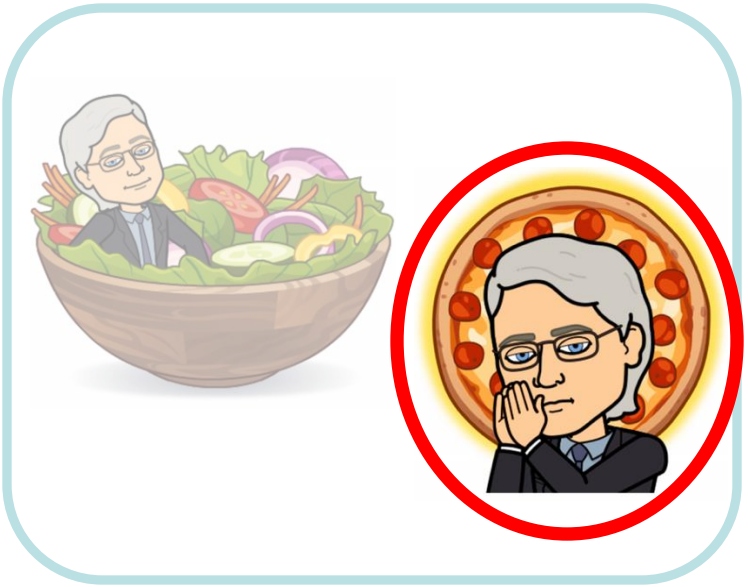


My task

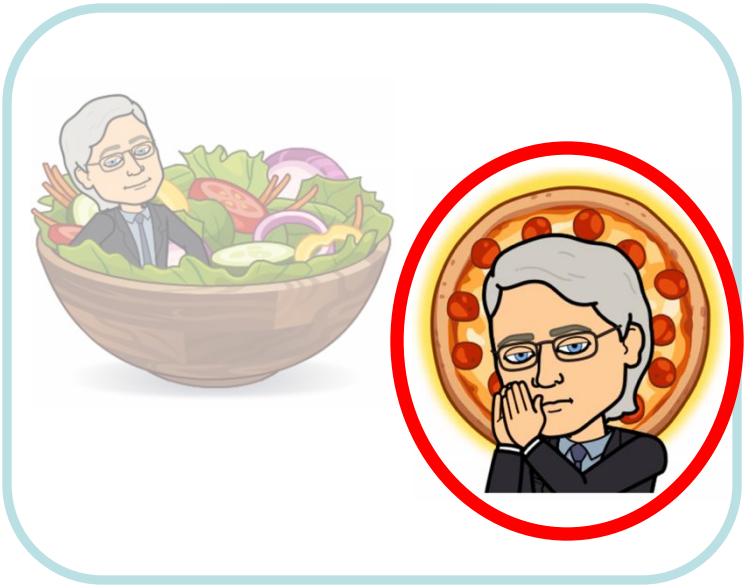


Standard Welfare Economics

The planner's task



My task



The Premises Standard Welfare Economics

The Premises Standard Welfare Economics

- ***Premise 1:*** *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*
  - $\succsim$ is a well-behaved (complete, transitive) preference relation

The Premises Standard Welfare Economics

- ***Premise 1:*** *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*

- ***Premise 2:*** *Each individual is the best judge of their own well-being.*

  – Philosophical justifications: (i) arguments for self-determination in the tradition of classical liberalism; (ii) Cartesian principle that experience is inherently private and not directly observable

  – Implication: $\succsim$ is *normative*.

The Premises Standard Welfare Economics

- ***Premise 1:*** *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*

- ***Premise 2:*** *Each individual is the best judge of their own well-being.*

- ***Premise 3:*** *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.*

  – From any choice set, the consumer selects a maximal element according to $\succsim$. It follows that $\succsim$ is discoverable from choices.

The Premises Standard Welfare Economics

- ***Premise 1:*** *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*

- ***Premise 2:*** *Each individual is the best judge of their own well-being.*

- ***Premise 3:*** *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.*

- ***Premise 4:*** *The consequences of the planner's actions for a particular individual are reproducible as consequences of actions when that individual is the decision maker.*

  – Changing the decision maker from the individual to the planner does not change the nature of the options in any other consequential way.

The behavioral critique

Premise 1: *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*

Premise 2: *Each individual is the best judge of their own well-being.*

Premise 3: *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.*

Premise 4: *The consequences of the planner's actions for a particular individual are reproducible as consequences of actions when that individual is the decision maker.*

The behavioral critique

Premise 1: *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*

Premise 2: *Each individual is the best judge of their own well-being.*

Premise 3: *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.*

Implementation Critiques

Premise 4: *The consequences of the planner's actions for a particular individual are reproducible as consequences of actions when that individual is the decision maker.*

The behavioral critique

Implementation Critiques:

- Conditional on the information they possess, people sometimes hold false beliefs about the consequences of their actions.
- People sometimes ignore options that are available to them.
- People sometimes cope with complexity by taking shortcuts – in other words, they deploy a heuristic or solve a problem that's simpler than the one they actually face.

The behavioral critique

Implementation Critiques:

- Conditional on the information they possess, people sometimes hold false beliefs about the consequences of their actions.
- People sometimes ignore options that are available to them.
- People sometimes cope with complexity by taking shortcuts – in other words, they deploy a heuristic or solve a problem that's simpler than the one they actually face.

Implication:

- Choice may be a poor guide to well-being

The behavioral critique

Premise 1: *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*

Coherence Critiques

Premise 2: *Each individual is the best judge of their own well-being.*

Premise 3: *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.*

Implementation Critiques

Premise 4: *The consequences of the planner's actions for a particular individual are reproducible as consequences of actions when that individual is the decision maker.*

The behavioral critique

Coherence Critiques:

- People may not have well-defined preferences. Instead, their preferences may be *constructed* contextually.
- People may have well-defined preferences, but they may not be well-behaved.
- People may have *endogenous preferences*

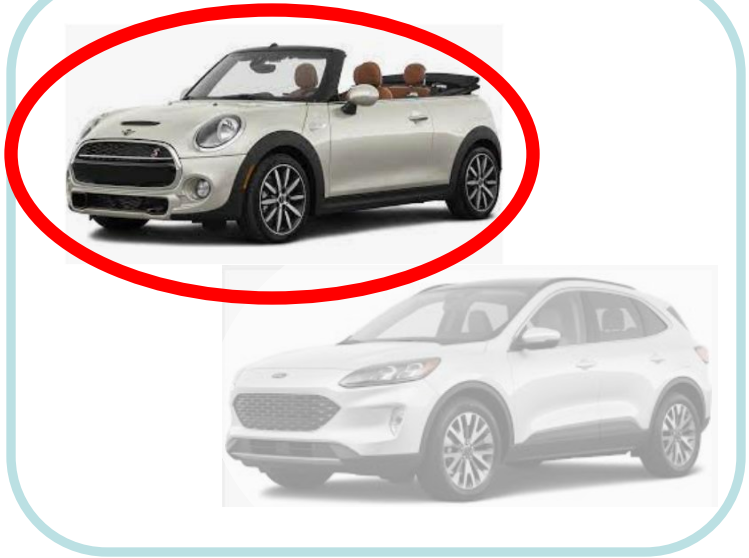
Sources of Context Dependence

Based on Busse, Pope, Pope, and Silva-Risso (2015)

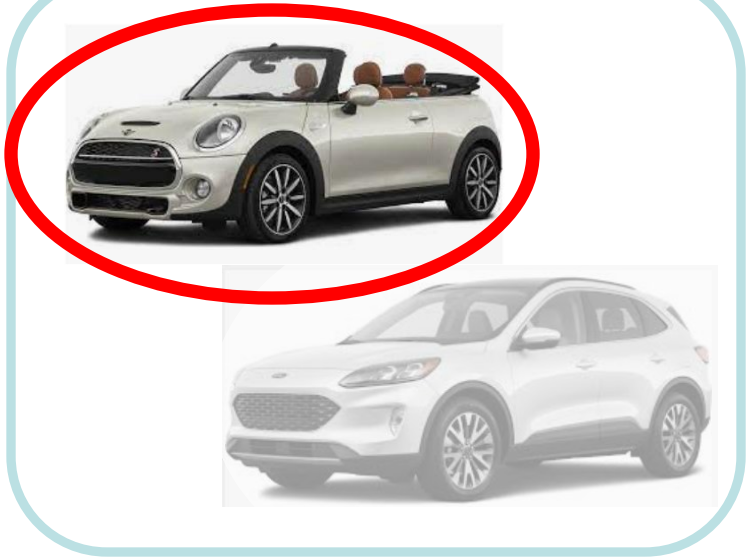
Sources of Context Dependence

Sources of Context Dependence

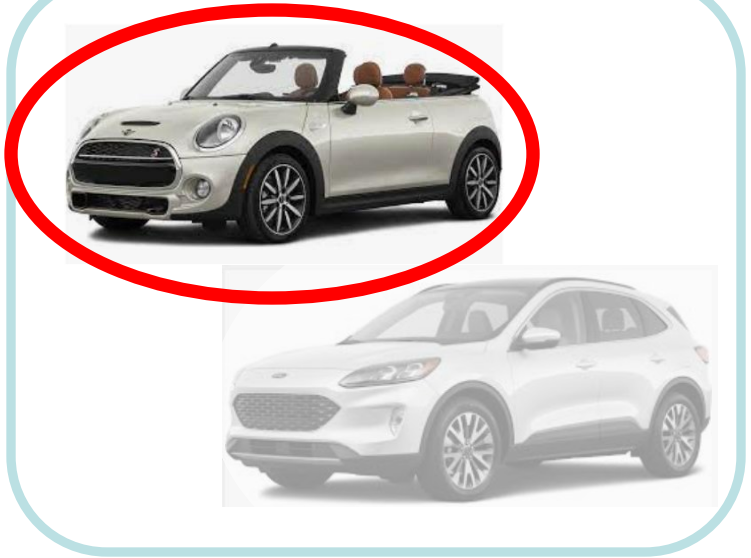
Sources of Context Dependence

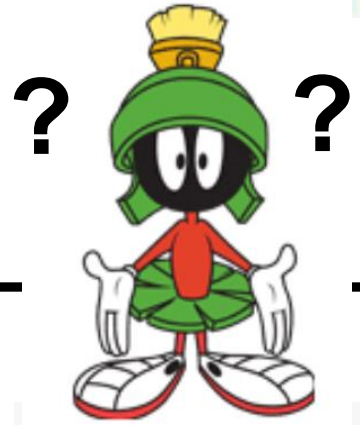


Sources of Context Dependence



Sources of Context Dependence



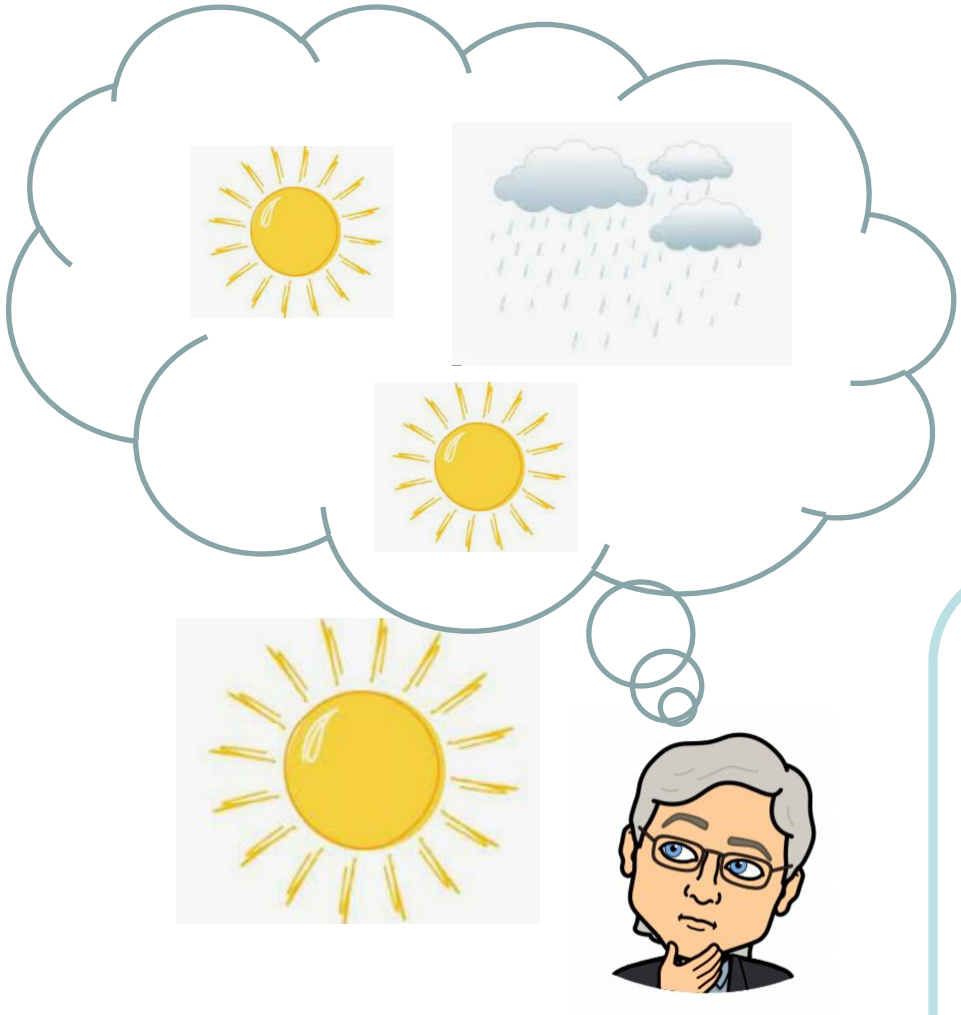


Sources of Context Dependence

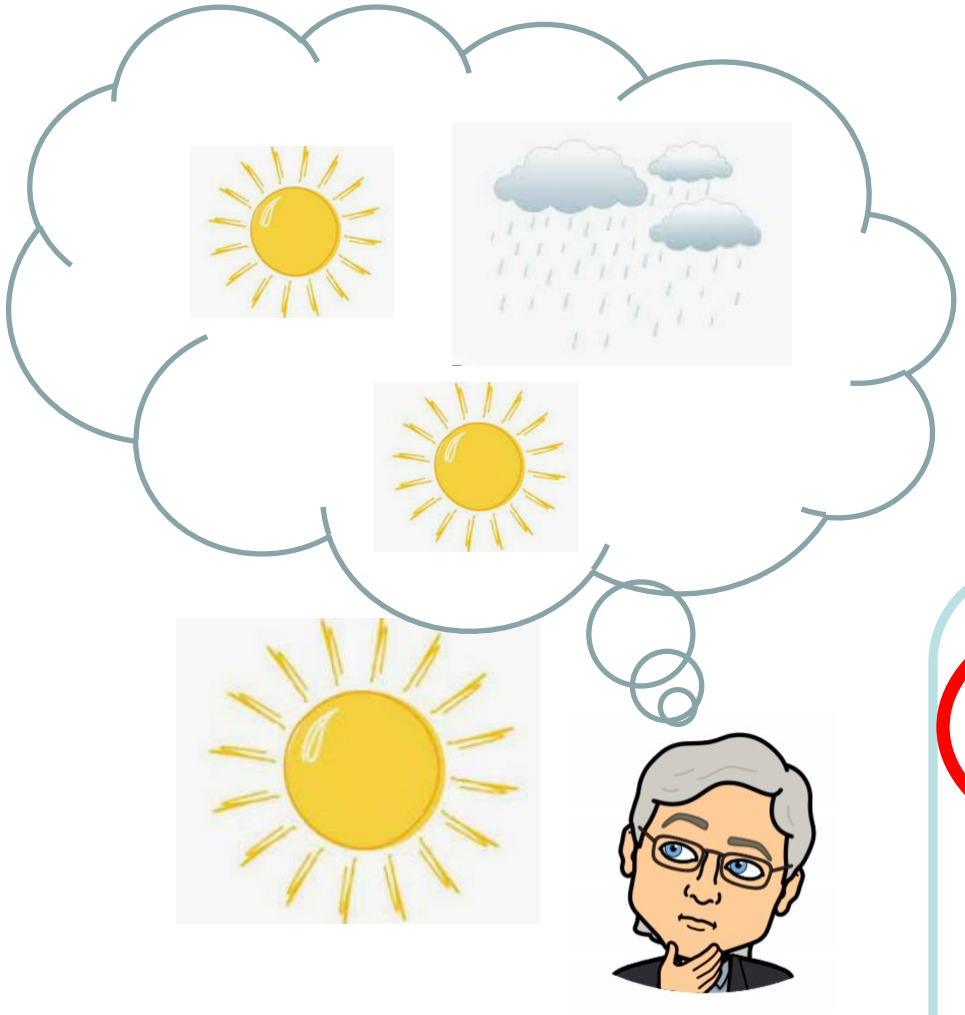
Hypothesis #1: Different contexts trigger different beliefs.

Sources of Context Dependence

Sources of Context Dependence

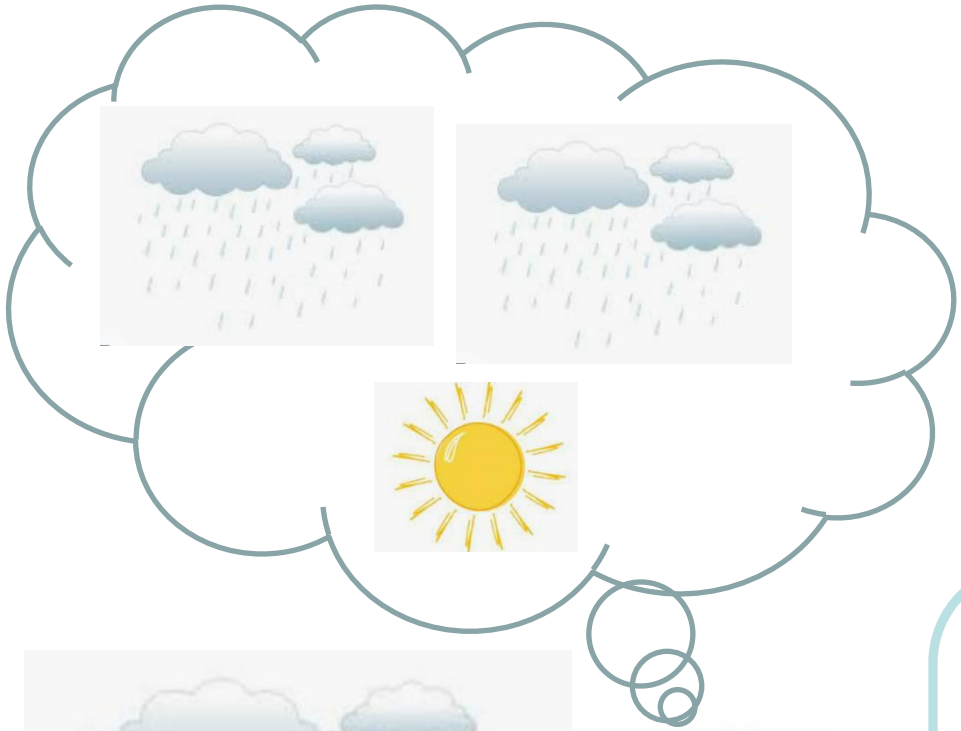


Sources of Context Dependence

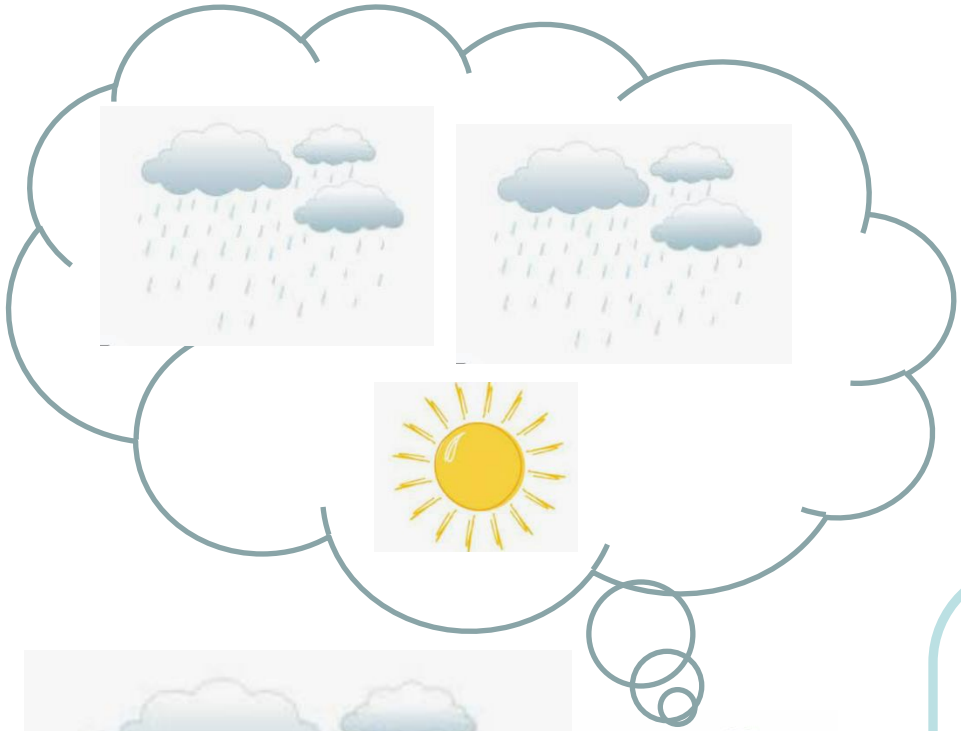


Sources of Context Dependence

Sources of Context Dependence



Sources of Context Dependence



Sources of Context Dependence

Hypothesis #1: Different contexts trigger different beliefs.

Both beliefs can't be right, so context dependence is evidence for an Implementation Critique of Premise #3.

In principle, one can establish which belief is right. The inconsistency is therefore *reducible*.

Sources of Context Dependence

Hypothesis #2: Different contexts trigger different judgments (*constructed preferences*).

Context & Preference Construction

Context & Preference Construction

Dimensions of experience:

- ❑ Fun
- ❑ Cost
- ❑ Appearance
- ❑ Reliability

Context & Preference Construction

How do we aggregate if there are no "true" preferences to access? No "inner rational agent"?

Dimensions of experience:

- ❑ Fun
- ❑ Cost
- ❑ Appearance
- ❑ Reliability

Context & Preference Construction

Context & Preference Construction

Fun

Reliability

Cost

Appearance

Context & Preference Construction

Fun

Reliability

Cost

Appearance

Context & Preference Construction

Context & Preference Construction

Fun

Reliability

Cost

Appearance

Context & Preference Construction

Fun

Reliability

Cost

Appearance

Sources of Context Dependence

Hypothesis #2: Different contexts trigger different judgments (*constructed preferences*).

There is no “ground truth” for preferences. Inconsistency is *irreducible*.

The behavioral critique

Premise 1: *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.* ⇨ ***Coherence Critiques***

Premise 2: *Each individual is the best judge of their own well-being.* ⇨ ***Judgment Critiques***

Premise 3: *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.* ⇨ ***Implementation Critiques***

Premise 4: *The consequences of the planner's actions for a particular individual are reproducible as consequences of actions when that individual is the decision maker.* ⇨ ***Reproducibility Critiques***

The behavioral critique

A Reproducibility Critique:

The act of choosing for oneself trigger welfare-relevant sensations.

As a result, the consequences of the planner's actions are not reproducible in an otherwise identical problem where the individual is the decision maker.

The act of choosing can have welfare consequences

Temptation

Anxiety

Sense of empowerment

Guilt

Shame

Pride

Discomfort with responsibility

Regret

The act of choosing can have welfare consequences

Temptation

Anxiety

Sense of empowerment

Guilt

Shame

Pride

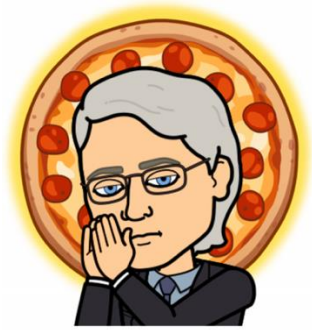
Discomfort with responsibility

Regret

Gives rise to a third source of context dependence

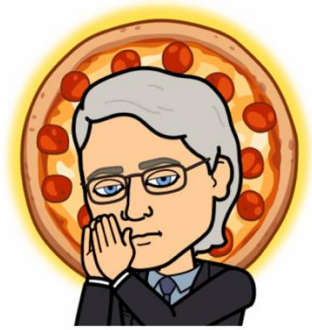
An Issue with Reproducibility

The planner's task



An Issue with Reproducibility

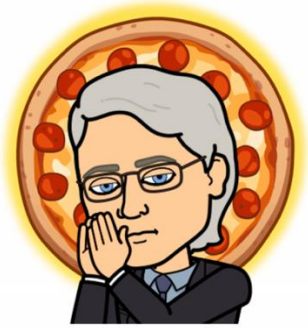
The planner's task



My task

An Issue with Reproducibility

The planner's task



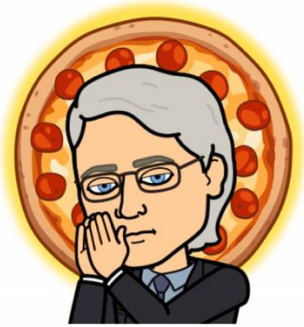
My task

+ temptation

+ guilt

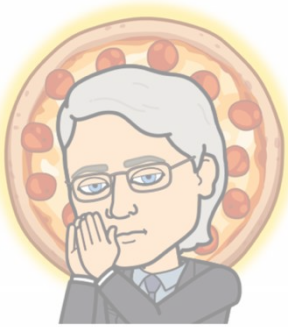
An Issue with Reproducibility

The planner's task



My task

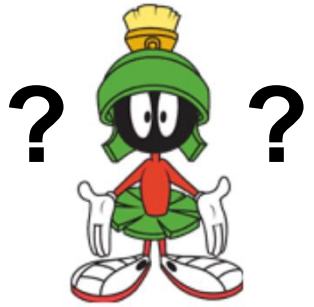
+ temptation

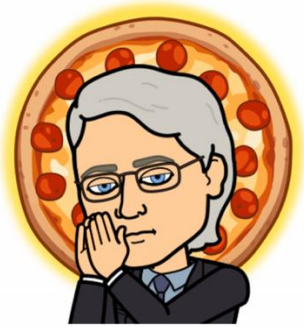


+ guilt

An Issue with Reproducibility

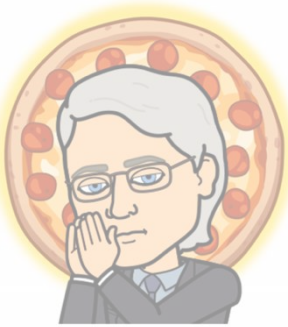
The planner's task





My task

+ temptation



+ guilt

And it gets worse…

- Suppose that, although I will *never* choose pizza for myself (because of guilt), I fervently wish that someone would take the decision out of my hands and order me a pizza (so I can have pizza without feeling guilty about choosing it).
- In that case:
  - A planner who defers to my preference ought to order me a pizza, but
  - *No choice problem can reveal that preference.*

The Non-Comparability Problem

If the experience of choosing falls within the scope of consumers' concerns, then welfare is not recoverable from choice.

Why is the NCP important?

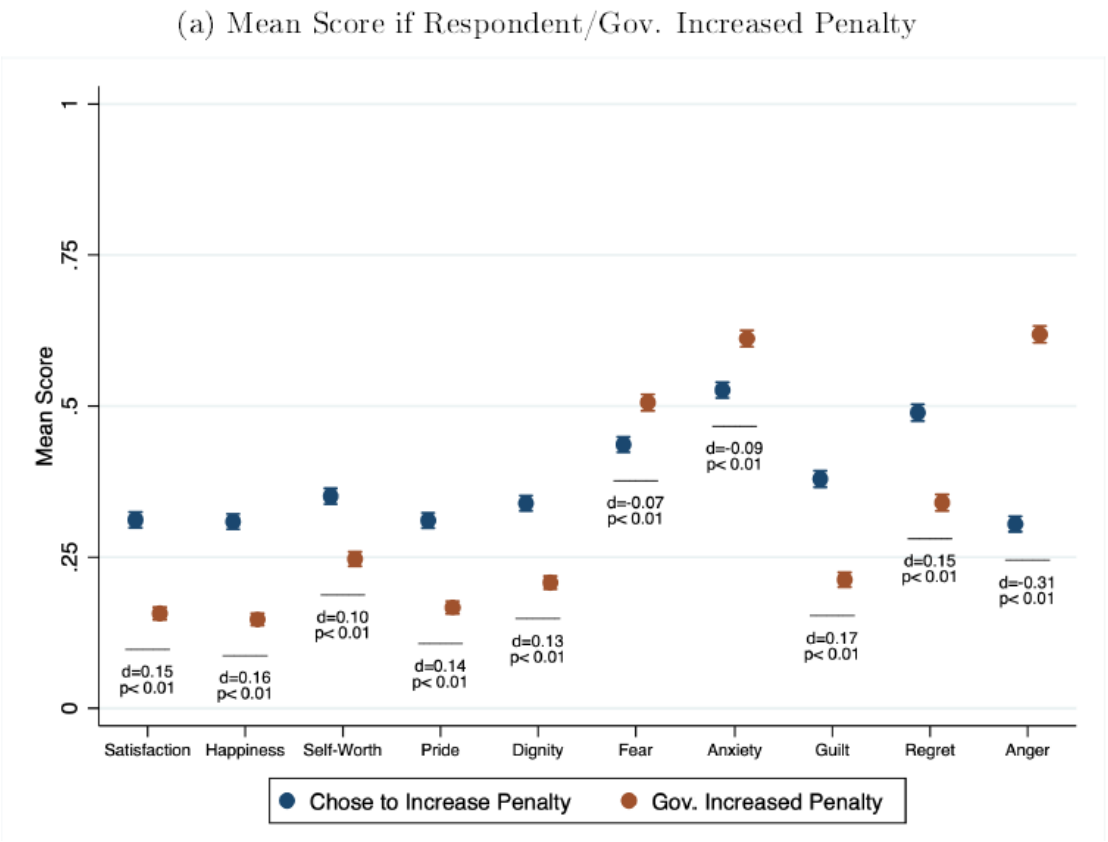
1. *Do we know how to evaluate policies that limit people's opportunity sets?*

Motivating survey evidence

Increase 401(k) early withdrawal penalty from 10% to 30%

% who say increased penalty would improve their well-being:

If chose: 28%
If gov. imposed: 16%



Why is the NCP important?

1. *Do we know how to evaluate policies that limit people's opportunity sets?*

2. *Do we know how to evaluate policies involving risk?*

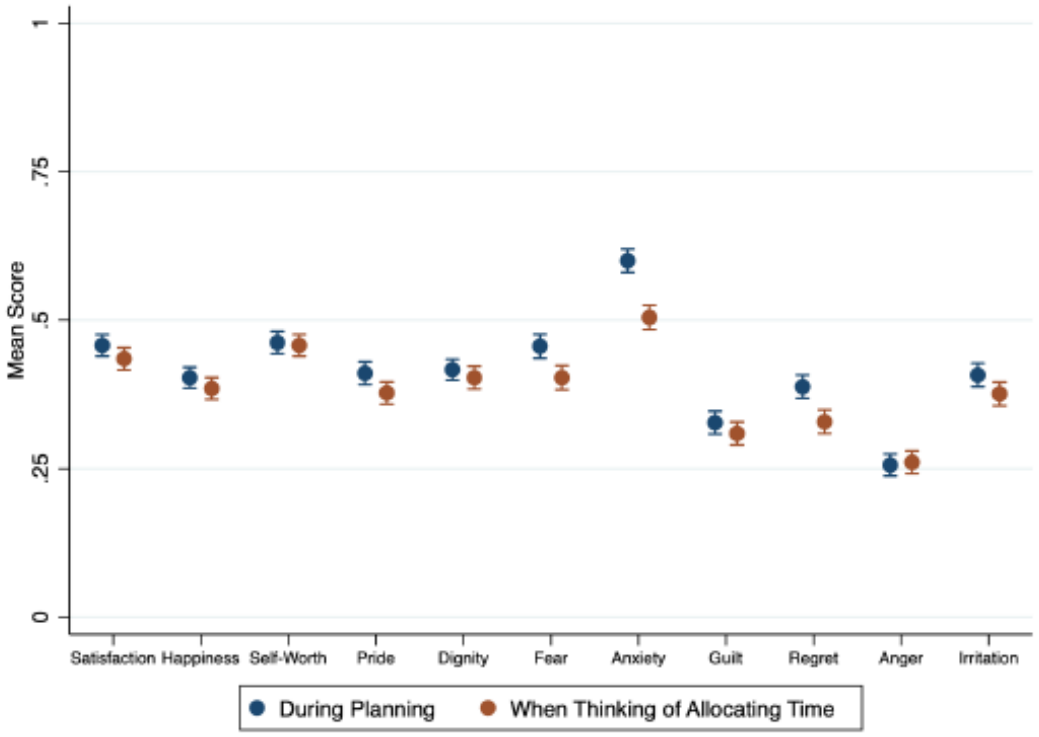
Motivating survey evidence

Avoidance of financial planning

56% of people say they avoid spending time on financial planning due to negative emotions

Feels overwhelming: 44%
Stress, anxiety, fear: 35%
Averse to complexity: 22%

Emotions from choice versus metachoice



Why is the NCP important?

1. *Do we know how to evaluate policies that limit people's opportunity sets?*

2. *Do we know how to evaluate policies involving risk?*

3. *Do we know how to evaluate policies involving false beliefs?*

Red pill or blue pill?

The behavioral critique

Premise 1: *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.* ⇨ ***Coherence Critiques***

Premise 2: *Each individual is the best judge of their own well-being.* ⇨ ***Judgment Critiques***

Premise 3: *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.* ⇨ ***Implementation Critiques***

Premise 4: *The consequences of the planner's actions for a particular individual are reproducible as consequences of actions when that individual is the decision maker.* ⇨ ***Reproducibility Critiques***

The behavioral critique

Does behavioral economics provide a foundation for judgment critiques?

The behavioral critique

Does behavioral economics provide a foundation for judgment critiques?

Some terminology:

- *Intrinsically valued outcomes* are those we care about for their own sake

- *Instrumentally valued outcomes* are those we care about because they leads to intrinsically value outcomes

- Example: I east an apple because the taste and texture produces pleasurable mental states. Eating an apple is an instrumentally valued outcome; the mental states are intrinsically valued outcomes.

- A normative judgment is *direct* if it pertains to intrinsically valued outcomes, and *indirect* if it pertains to instrumentally valued outcomes.

The behavioral critique

Does behavioral economics provide a foundation for judgment critiques?

- Behavioral Economics provides good reasons to question indirect judgments: they be tainted by faulty understanding of consequences (e.g., I may be wrong about the pleasure I'll get from eating an apple). But that's a false belief, which we've already covered under the heading of Implementation Critiques.

- Behavioral economics does not provide a foundation for challenging direct judgments. Such challenges are "differences of opinion."

- So, if we understand Premise 2 as applying to the direct judgments that motivate our indirect judgments, behavioral economics does not provide a basis for challenging it.

- Objections to Premise 2 are, however, found in Philosophy (e.g., objective list theories of well-being)

Paths Forward: Replace the Standard Approach

Paths Forward: Replace the Standard Approach

- One leading possibility: evaluate outcomes based on *self-reported well-being (SRWB)*
- This approach presents its own set of conceptual and practical challenges
- Chief among those challenges: *The Aggregation Problem*
  - The “best case” scenario for SRWB: there is an internal “meter” encompassing our feelings about the present, memories of the past, and expectations of the future; we can “read the meter” when asked
  - But then, how do we aggregate over different meter readings at different dates, and in different states of nature?
  - We can ask people to aggregate over past and expected future meter readings. But then they aren’t “reading a meter.” The principle of aggregation is “linguistic” (i.e., based on our understanding of the question).

Paths Forward: Fix the Standard Approach

Premise 1: *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*

⇨ ***Coherence Critiques***

Premise 2: *Each individual is the best judge of their own well-being.*

Premise 3: *Each individual's preferences determine their choices. When they choose, they seek and achieve the greatest benefit according to their own judgment, subject to their constraints.*

⇨ ***Implementation Critiques***

Premise 4: *The consequences of the planner's actions for a particular individual are reproducible as consequences of actions when that individual is the decision maker.*

⇨ ***Reproducibility Critiques***

A non-solution

- A first instinct for many economists: introduce *metachoices*
  - If someone's choices are context-dependent, ask them to select the context, and respect the preferences those decisions reveal.
  - If the act of choosing engenders welfare-relevant emotions, measure those responses by gauging the extent to which people are attracted/repelled by the decision problem
- This method has gained popularity
  - Dana, Cain, and Dawes (2006) (exit in the dictator game), Lazear, Malmendier, and Weber (2012) (sorting in experiments), DellaVigna, List, and Malmendier (2012) (charitable solicitation), Bartling, Fehr, Herz (2014) (valuing autonomy), Allcott and Kessler (2019) (nudges involving social comparisons), Butera, Metcalfe, and Taubinsky (2022) (social recognition for YMCA attendance)

A non-solution

- Why doesn't the metachoice method work?
  - A metachoice is just another way of structuring a choice. So, any conceptual problem that arises a choice also arises for a metachoice.
- The car purchase problem:
  - To deploy the metachoice method, we would want to know if I prefer to select a car on a sunny day or a rainy day
  - But what if, on sunny (resp. rainy) days, I feel the need to make important decisions on sunny (resp. rainy) days? What if the metachoice framing leads to different (false) beliefs, or triggers a different type of preference construction?
- The lunch purchase problem:
  - To deploy the metachoice method, we would want to know if I prefer to select my own lunch, or delegate to someone who will select Pizza for me
  - But I'll still feel guilty about delegating to someone I know will choose Pizza

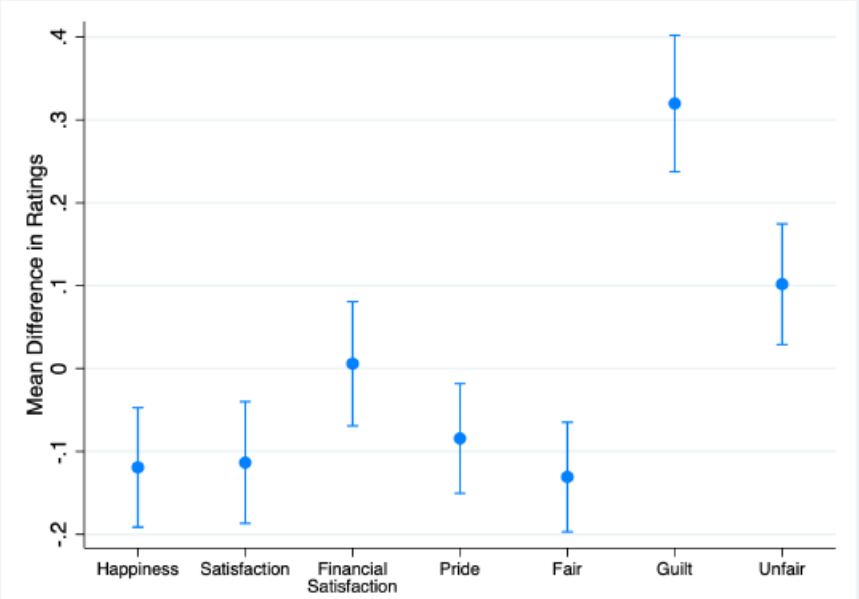
Motivating survey evidence

The charitable solicitation problem (based on DellVigna, List, & Malmendier, 2012)

Person soliciting charitable contributions rings your doorbell (& you identify them through doorbell camera).

Scenario 1: You pretend you're not home (meta-choice).

Scenario 2: You aren't home (decision not to give isn't in your hands)



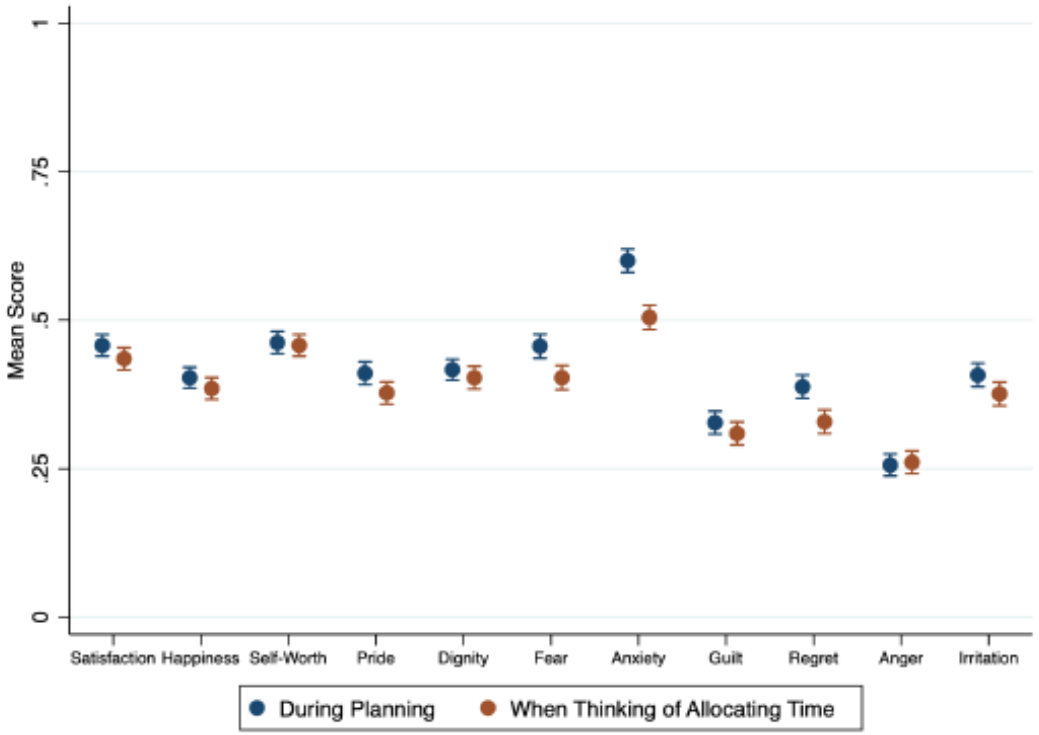
Motivating survey evidence

Avoidance of financial planning

56% of people say they avoid spending time on financial planning due to negative emotions

Feels overwhelming: 44%
Stress, anxiety, fear: 35%
Averse to complexity: 22%

Emotions from choice versus metachoice



Challenge 1: Implementation Critiques

Challenge 1: Implementation Critiques

The method of Behavior Revealed Preference (BRP): Supplement standard models of choice with additional elements representing the "cognitive biases" that purportedly account for imperfections of implementation. Use choices to learn about preferences and biases simultaneously.

- Tries to tackle Implementation Critiques by relaxing Premise 3 while maintaining all the other premises.
- In effect, it assumes all context-dependence involves implementation failures (not preference construction, not act-of-choosing emotions).

Elements of a BRP analysis:

- $U(x, f)$: a "decision utility" function that rationalizes observed choices over options x conditional on a decision frame f.
- $V(x)$: a normative objective function used to evaluate welfare (*true preferences*).

Challenge 1: Implementation Critiques

The usual route to identification of $V(\cdot)$ for BRP:

- We assume that, for certain decision frames f, $U(x, f)$ and $V(x)$ agree (frames that yield "unbiased choices")
- We use the set of "unbiased choices" (the ***Welfare-Relevant Domain***) to recover "true preferences"
- For other decision frames, we allow for the possibility that $U(x, f)$ and $V(x)$ diverge (frames that yield "biased choices")

Challenge 1: Implementation Critiques

How do we define "unbiased choices"?

A common proposal: "Unbiased choices" are those that are consistent with true preferences (V)

This definition is too vague to be of any use:

- Even if true preferences exist, how would we recognize them? How would we figure out which choice is mistaken?
- What makes a preference "true"? What are the defining characteristics of true-ness?

Challenge 1: Implementation Critiques

The Circularity Trap: True preferences are revealed by choices that are not mistakes, and mistakes are choices that are inconsistent with true preferences.

In effect, the BRP approach requires us to know what's inconsistent with true preferences so we can exclude it before trying to recover true preferences.

Challenge 1: Implementation Critiques

Example: "Present-bias"

- Standard model of "decision utility": $U_t = u_t + \beta(\delta u_{t+1} + \delta^2 u_{t+2} + \ldots)$
- A widespread view of "true preferences": $V_t = u_t + \delta u_{t+1} + \delta^2 u_{t+2} + \ldots$
  - $\beta < 1$ is taken to be a bias (*weakness of will*)
  - Unbiased choices are those that are made in advance, and involve full commitment (the *long-run criterion*)

Challenge 1: Implementation Critiques

Example: "Present-bias"

- Standard model of "decision utility": $U_t = u_t + \beta(\delta u_{t+1} + \delta^2 u_{t+2} + \ldots)$
- A widespread view of "true preferences": $V_t = u_t + \delta u_{t+1} + \delta^2 u_{t+2} + \ldots$
  - $\beta < 1$ is taken to be a bias (*weakness of will*)
  - Unbiased choices are those that are made in advance, and involve full commitment (the *long-run criterion*)
- What principles and/or evidence support this perspective? Consider:
  - Pejorative views of present-focus are not universal
  - Deathbed regrets favor present-focus
  - Is the long-run criterion a reflection of "Type A paternalism"?

Challenge 1: Implementation Critiques

Avoiding the Circularity Trap (Bernheim and Rangel, 2004, Bernheim 2025)

- We need to define a mistake without referring to "true preferences" (V)
- Decisions are logically separable into three components
  - *Characterization:* what options are available, and how do they map to intrinsically valued consequences?
  - *Judgment:* is one bundle of intrinsically valued consequences better or worse than another?
  - *Optimization:* among the available options, find the one that is best given the
- Because Premise 2 precludes us from challenging (direct) judgment, an implementation failure must entail a *Characterization Failure* or an *Optimization Failure*
  - These failures are not necessarily "mistakes," in that they may be optimal responses to complexity. But their existence still implies *improvability.*

Challenge 1: Implementation Critiques

Identifying Implementation Failures

- See Bernheim and Taubinsky (2018) for a review of methods, or Ambuehl, Bernheim, and Lusardi (2022) for a detailed application.

The empirical identification of Characterization Failures through *direct evidence* is not especially difficult:

- Document incorrect beliefs about the consequences of actions
- Document lack of awareness of alternatives

The empirical identification of Optimization Failures through *direct evidence* is also possible but more challenging:

- Document reliance on shortcuts

Challenge 1: Implementation Critiques

The problem with direct proof is that it's always limited to the specific failures one looks for.

An alternative is to relay on *indirect evidence*, which includes sensitivity of choices to:

- Opacity or complexity of the decision problem
- Poor comprehension of principles governing consequences
- Cognitive limitations affecting attention, memory, forecasting

Challenge 1: Implementation Critiques

Some additional possibilities for *indirect evidence* include:

- Parallelism of behavioral patterns between the setting of interest and a "mirror" setting in which those patterns are definitely mistakes (Oprea, 2023)

- Self-reported lack of confidence in decisions (Enke, Graeber, and Oprea, 2024)

Challenge 1: Implementation Critiques

Research in progress (Bernheim, Lucia, Nielsen, & Sprenger)

A concern about the parallelism method:

- Similarity between the primary setting and the “mirror” setting may lead to confusion that causes the mistakes in the mirror setting
- If people find tasks in the mirror setting more difficult, they may use their criteria from the primary setting as heuristics

A concern about the stated confidence method: people may express low confidence for a variety of irrelevant reasons.

Challenge 1: Implementation Critiques

This escape route from the Circularity Trap is unworkable within the BRP framework

- If people construct their preferences contextually (irreducible inconsistency), then conflicts will exist within the WRD among choices that involve no Characterization or Optimization Failure according to any appropriate objective criterion.
- In such cases, the BRP approach requires us to invent additional reasons for declaring that some of the conflicting choices are mistakes.
- The BRP paradigm therefore *stands in the way* of developing general objective principles for classifying choices as mistakes: it consigns us to ad hoc judgments (to resolve context dependence arising from preference construction).

Challenge 1: Implementation Critiques

Example: Suppose we find that automobile purchases depend on the current weather, but pertinent beliefs (e.g., about future weather) do not. Can we say whether sun or rain makes people irrational?

BRP forces us to invent a reason for officiating

Challenge 1: Implementation Critiques

Example: Suppose we find that automobile purchases depend on the current weather, but pertinent beliefs (e.g., about future weather) do not. Can we say whether sun or rain makes people irrational?

BRP forces us to invent a reason for officiating

Conclusion: To overcome Challenge 1 (mistakes), we first have to address Challenge 2 (irreducible inconsistency).

- If we can figure out how to accommodate inconsistent choices, we won't need ad hoc criteria for identifying mistakes. Instead, we'll be free to use general objective criteria.

Challenge 1: Implementation Critiques

Circling back to "weakness of will"

It could be Characterization Failure: In the moment, we may blind ourselves to future consequences in order to justify indulgence.

It could be contextual preference construction: We may place different weight on the dimensions of our experience in advance and in the moment.

- In that case, using the phrase "weakness of will" is simply a way of expressing disagreement with the choice, and rationalizing the superimposition of the analyst's judgment

In the absence of evidence, we have no business assuming the first explanation is the correct one.

Challenge 2: Coherence Critiques

Challenge 2: Coherence Critiques

Welfare analysis at the crossroads...

- Is our commitment to Premise 2 (deference to the individual's judgments) conditional on Premise 1 (consistency of those judgments)?

- My answer (based on the justifications for Premise 2 given earlier) is that it's not conditional.

- Analogy: a panel of experts merits deference, even if the experts do not agree on every point.

  – The expertise concerning my well-being lies within me, even if I take different views of my well-being under different conditions.

Challenge 2: Coherence Critiques

The proposal (Bernheim & Rangel, QJE, 2009)

- Evaluate welfare according to the following criterion:

> *The Unambiguous Choice Relation:* *Option x is better than option y if there is a decision problem in the WRD for which x is chosen when y is available, but there is no decision problem in the WRD for which y is chosen when x is available.*

- This is a binary relation, written xP^*y
- Generalizes the standard notion of revealed preference
- Admits the possibility that welfare is ambiguous (because choice is not entirely consistent within the WRD)

Challenge 2: Coherence Critiques

Why this particular criterion?

- It is the only criterion satisfying a small collection of attractive properties.
  - Coherence of the welfare criterion (acyclicity)
  - Responsiveness to choice
  - Consistency with the WRD

Challenge 2: Coherence Critiques

Where does this criterion lead?

- Substituting this welfare criterion for the standard revealed preference criterion in Step 2, we can accommodate *irreducible inconsistency*, as well as *partial purification*. We can therefore accommodate any definition of mistakes, including the one proposed earlier (characterization failure)

- This framework yields counterparts for all the standard of tools of welfare analysis (consumer surplus, equivalent and compensating variations, Pareto optimality…)

  - See Bernheim, Fradkin, & Popov (*AER*, 2015) for foundations of aggregate versions of equivalent and compensating variation.

- The solution requires us to live with a degree of ambiguity.

Challenge 2: Coherence Critiques

A conceptual example

- Depending on framing, I always choose a coffee mug over $4, and always choose $5 over a mug, but my decision is frame-dependent in between $4 and $5
- In that case, we can say that the equivalent variation associated with having the mug is the range $4 to $5.

A practical application: What is the optimal default contribution rate for employee-directed pension plans?

- Default options may matter for psychological reasons (procrastination, inattention, anchoring…) that create normative ambiguity.

- And yet, the ambiguity turns out to be smaller than expected, and has no impact on the optimal policy (Bernheim, Fradkin, & Popov, *AER*, 2015, Bernheim and Mueller-Gastell, WP, 2022)

Challenge 3: Reproducibility Critiques

Challenge 3: Reproducibility Critiques

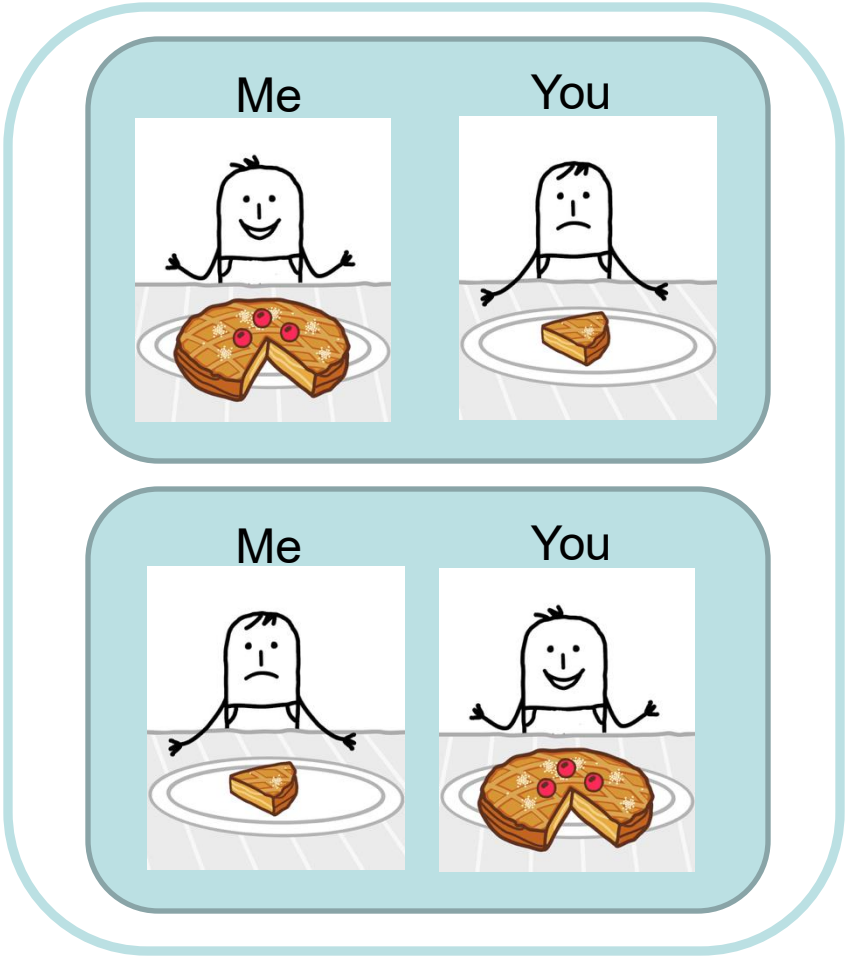
Synopsis of a proposed solution (Bernheim, Kim, and Taubinsky, 2024)

- Adopt the philosophical position that welfare consists of having one's desires satisfied (*Desire Satisfaction Theory*)

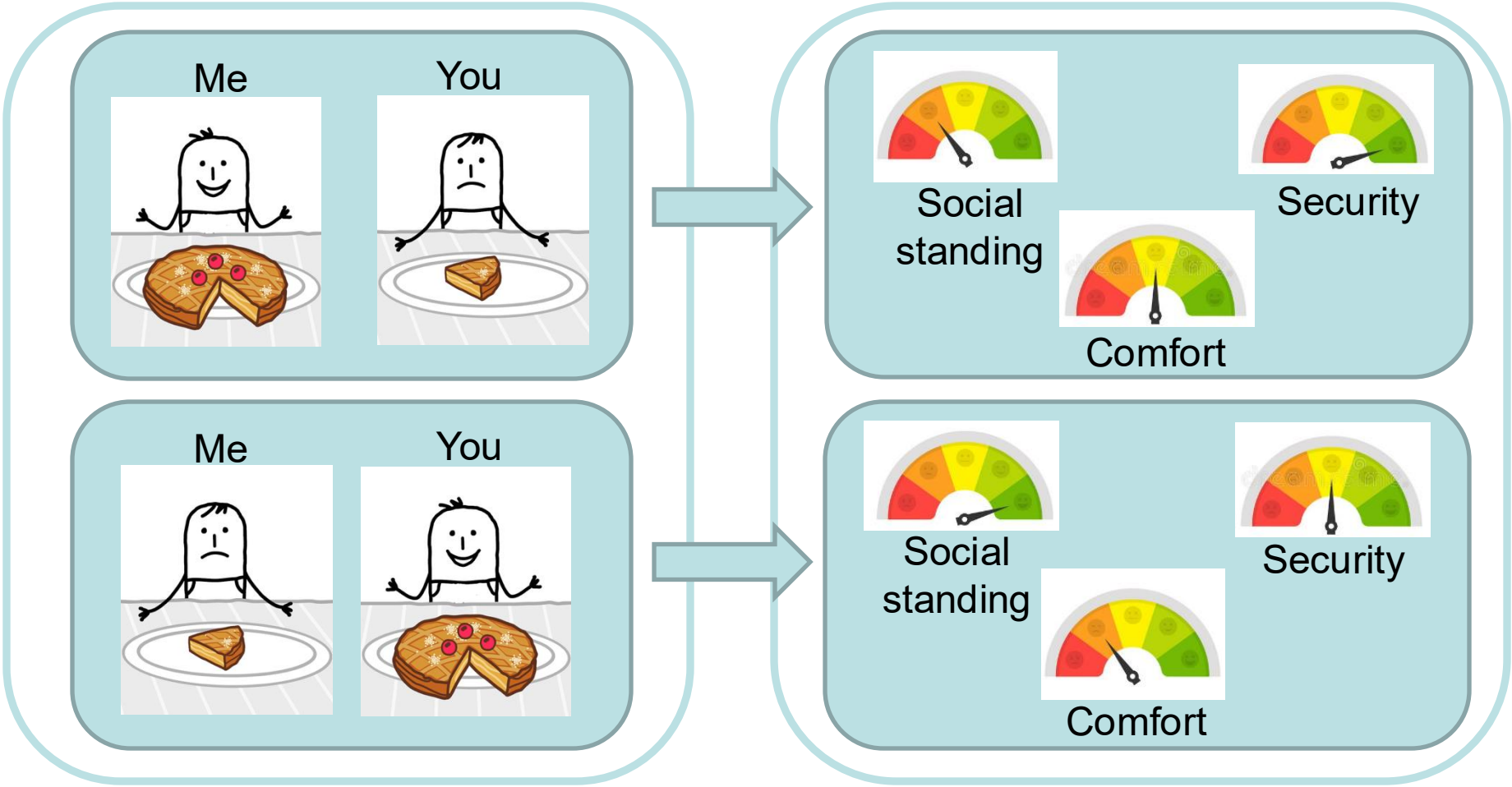
- Assume that people are *mental statists*: they care only about their mental states (meaning that they have preferences, $\succsim$, over mental state bundles, z)
  - Connection to Lancaster (1966): mental states are the "characteristics" of goods

- Assume that, to make a decision, they evaluate the mental state bundles they expect to follow from each option and then pick their favorite option from the resulting menu of mental state bundles.

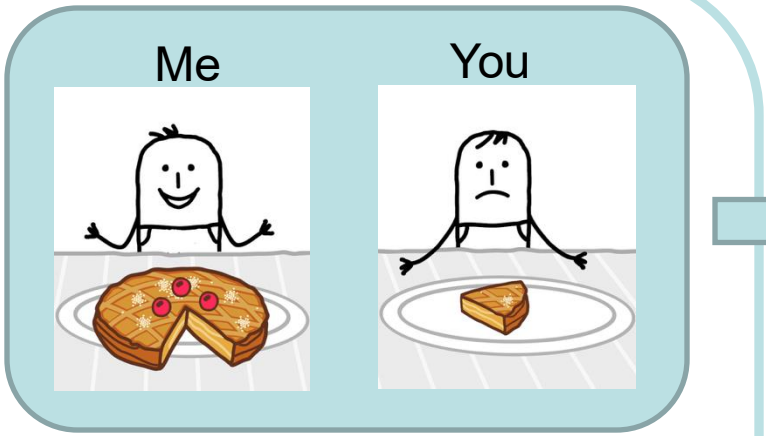
Challenge 3: Reproducibility Critiques

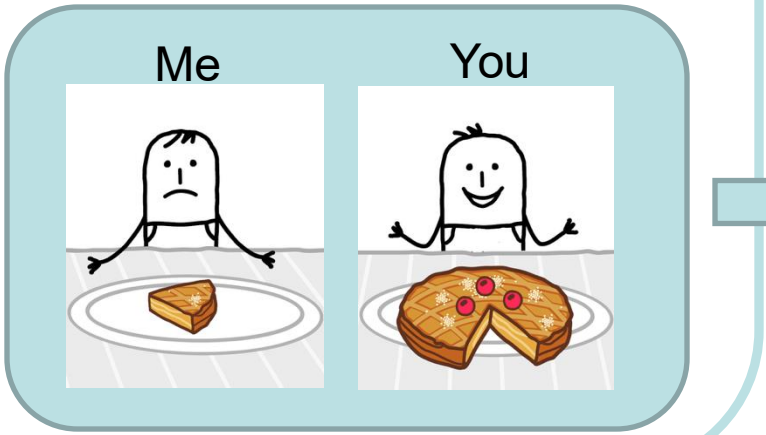


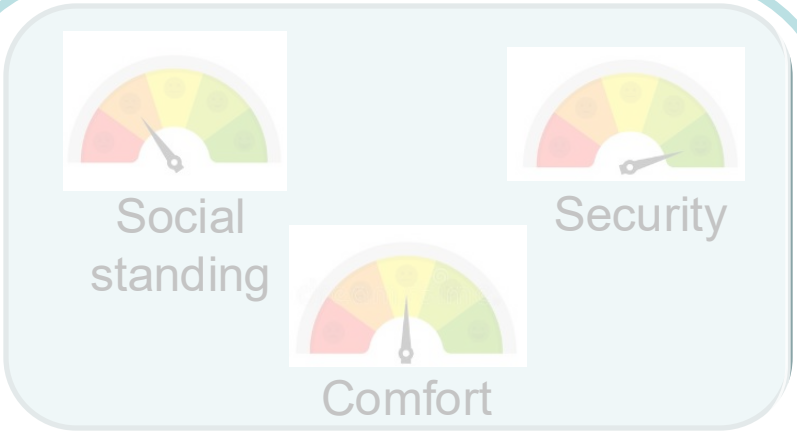
Challenge 3: Reproducibility Critiques



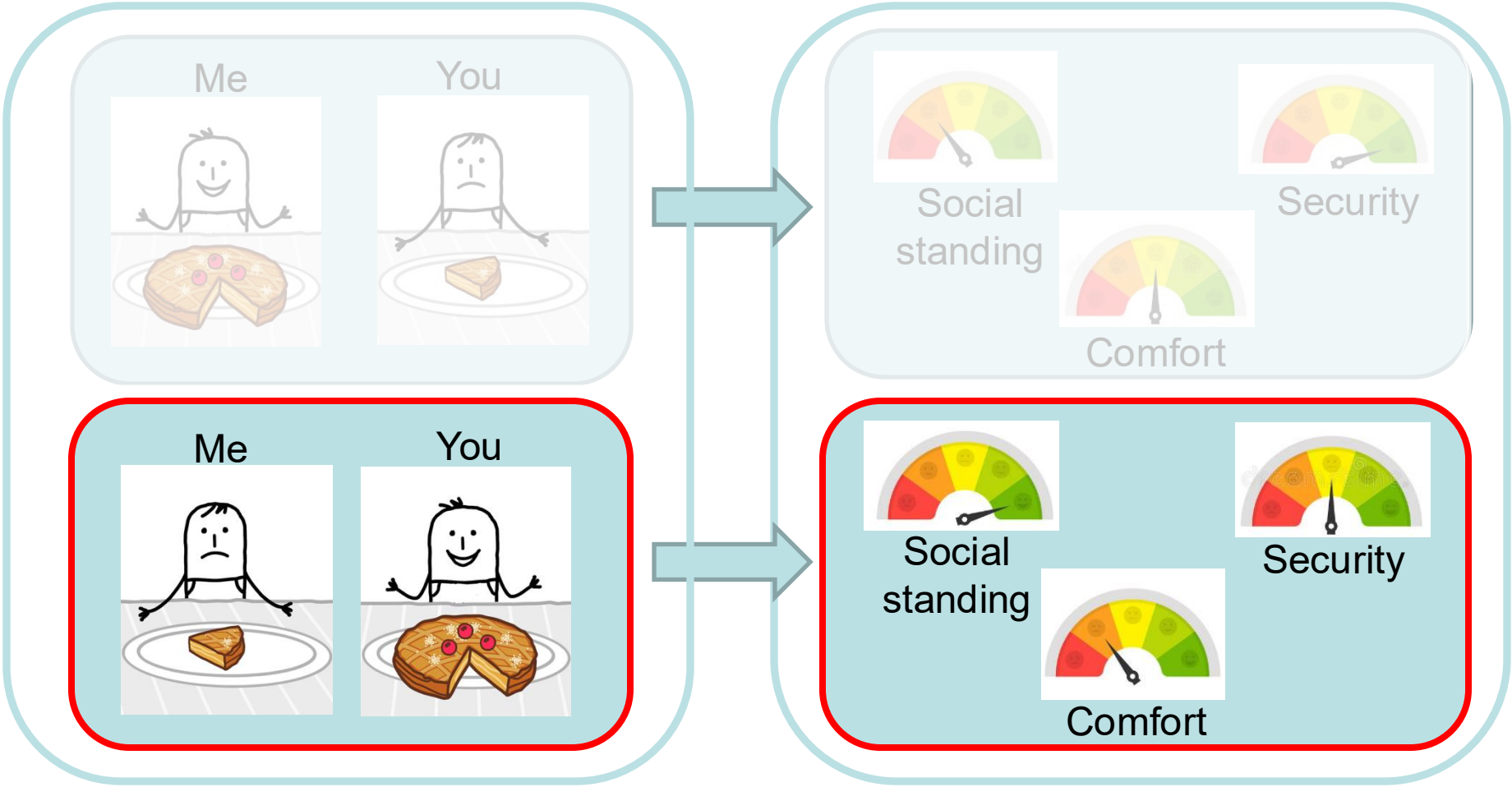
Challenge 3: Reproducibility Critiques



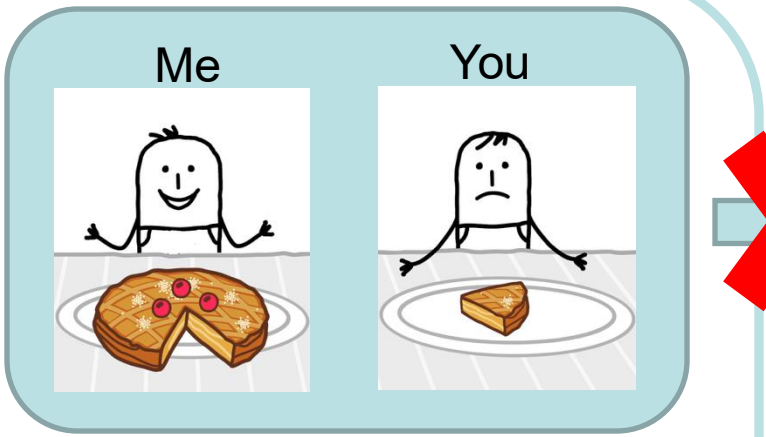


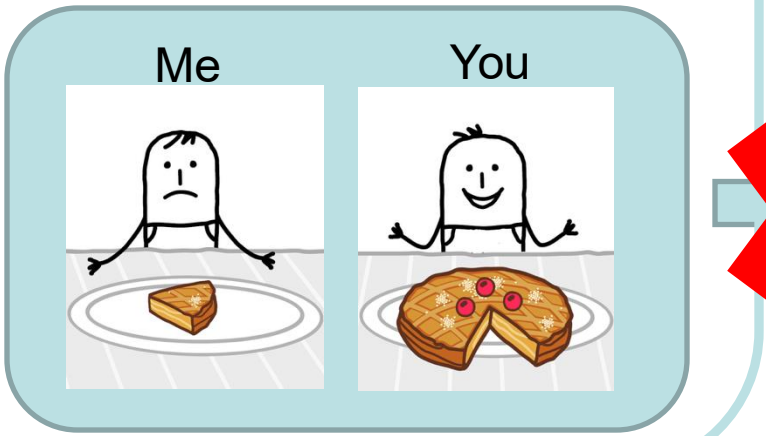


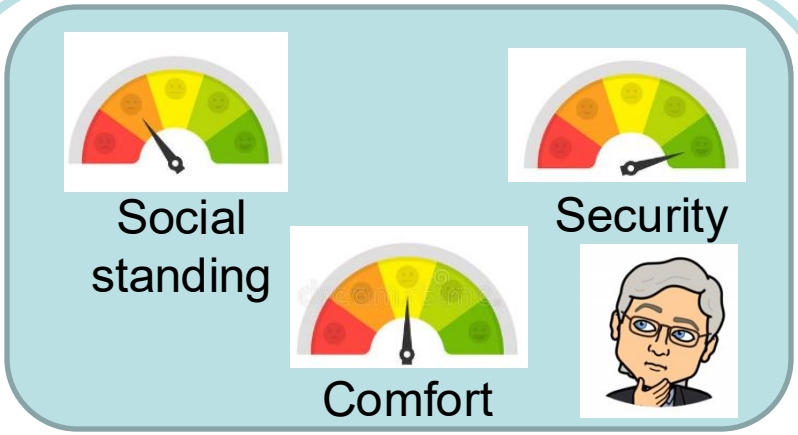
Challenge 3: Reproducibility Critiques

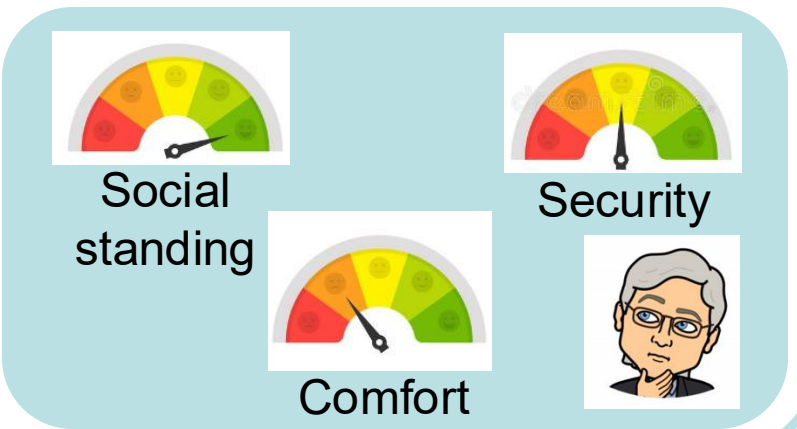


Challenge 3: Reproducibility Critiques

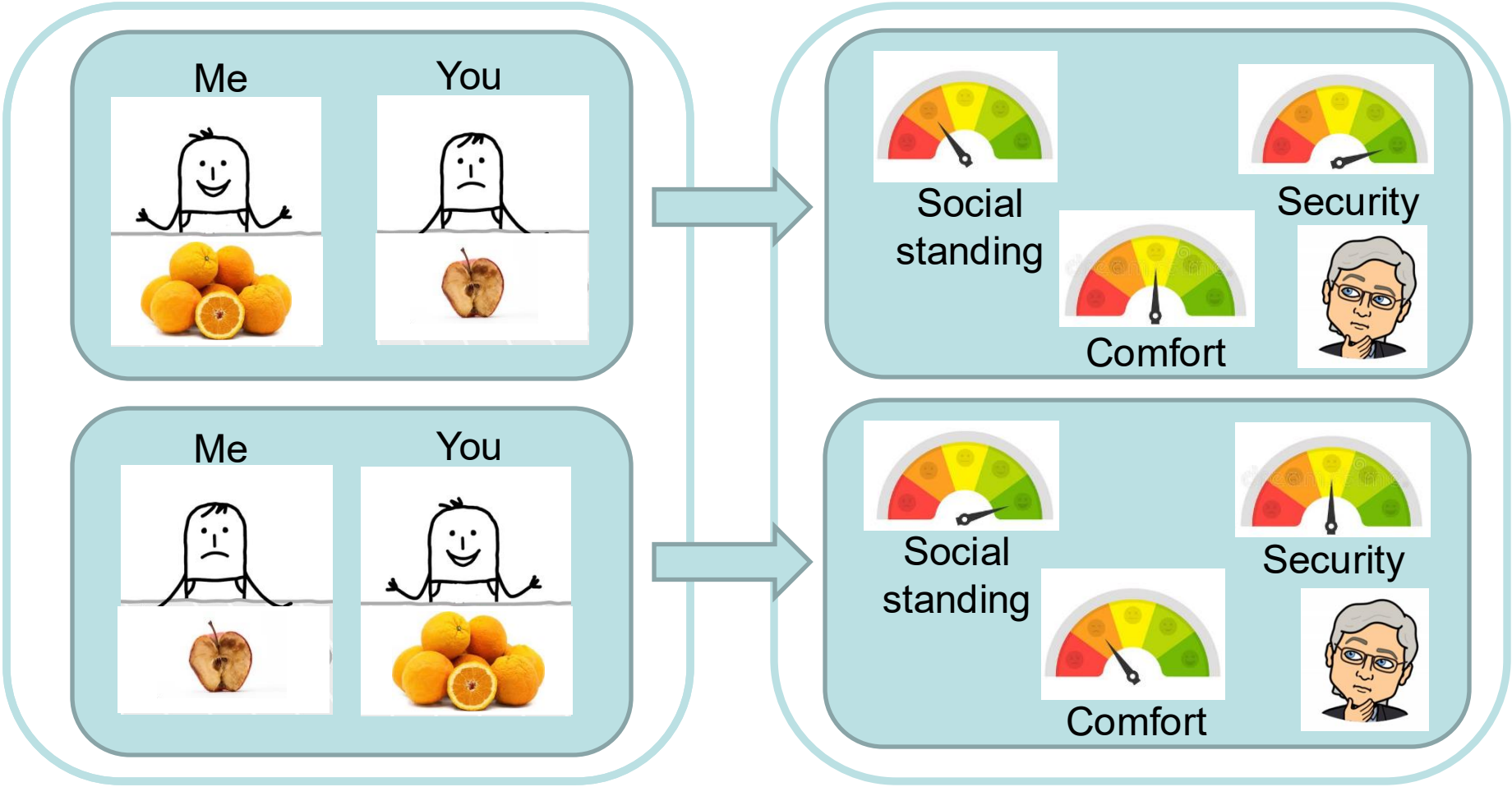




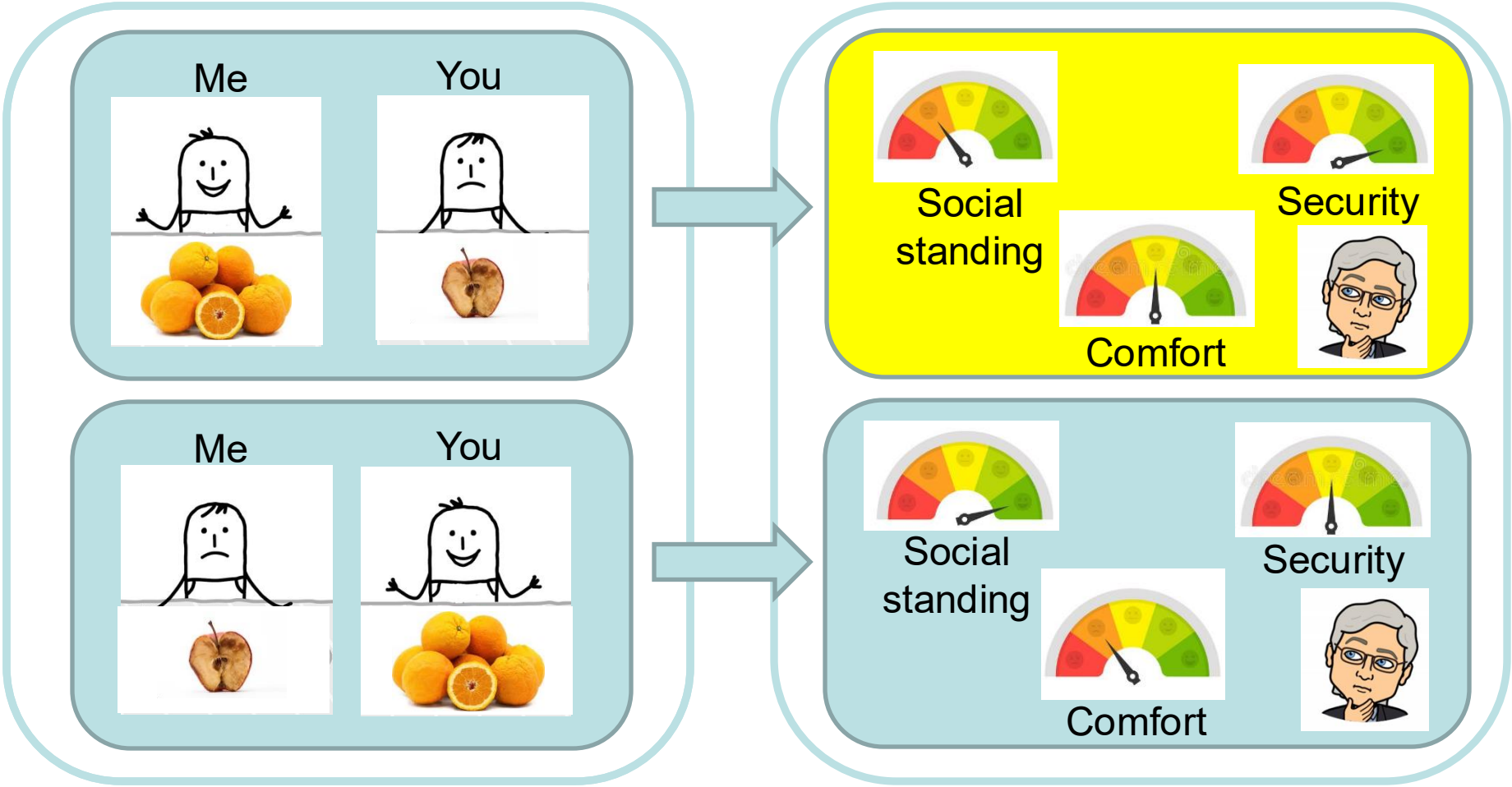




Challenge 3: Reproducibility Critiques

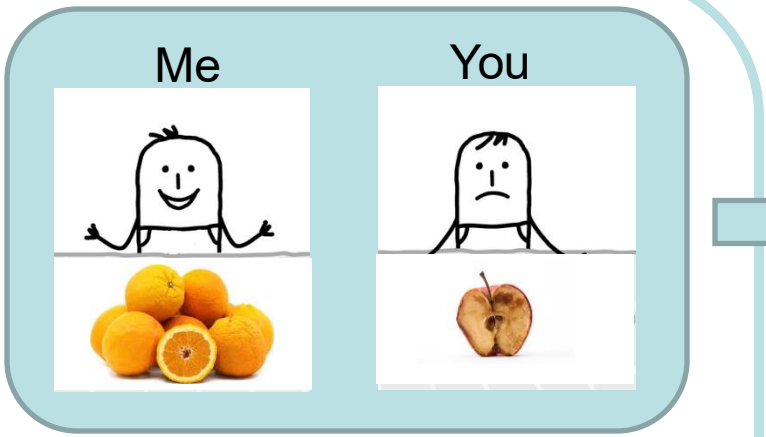


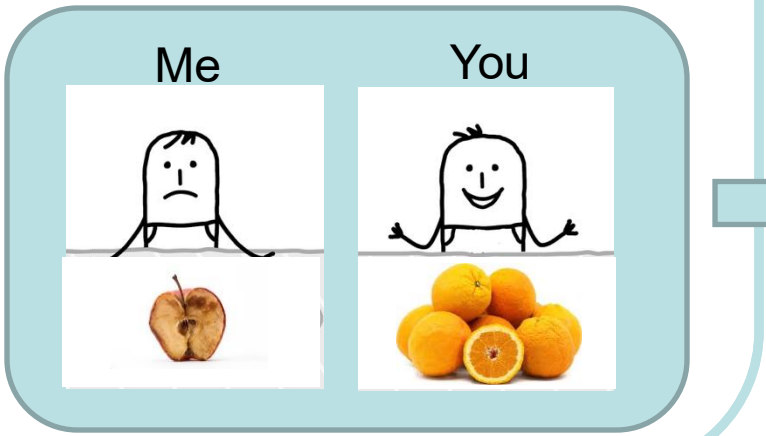
Challenge 3: Reproducibility Critiques

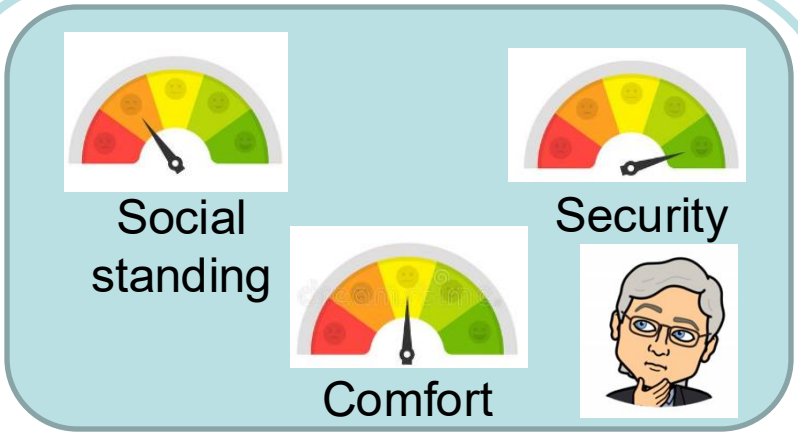


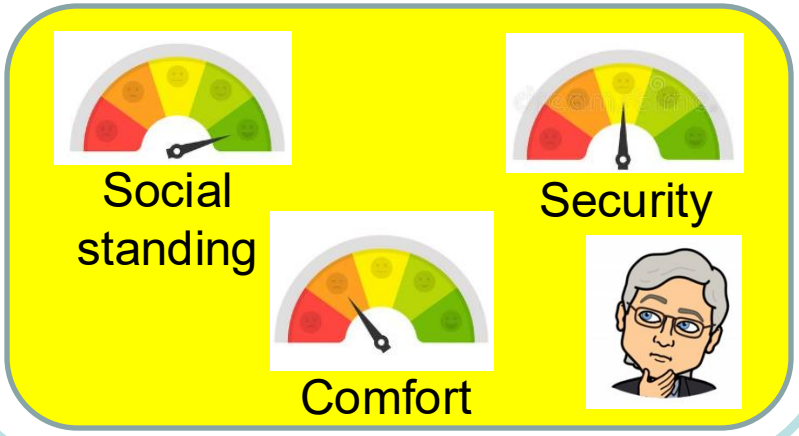
Same as selfish option when I divide the pie myself

Challenge 3: Reproducibility Critiques



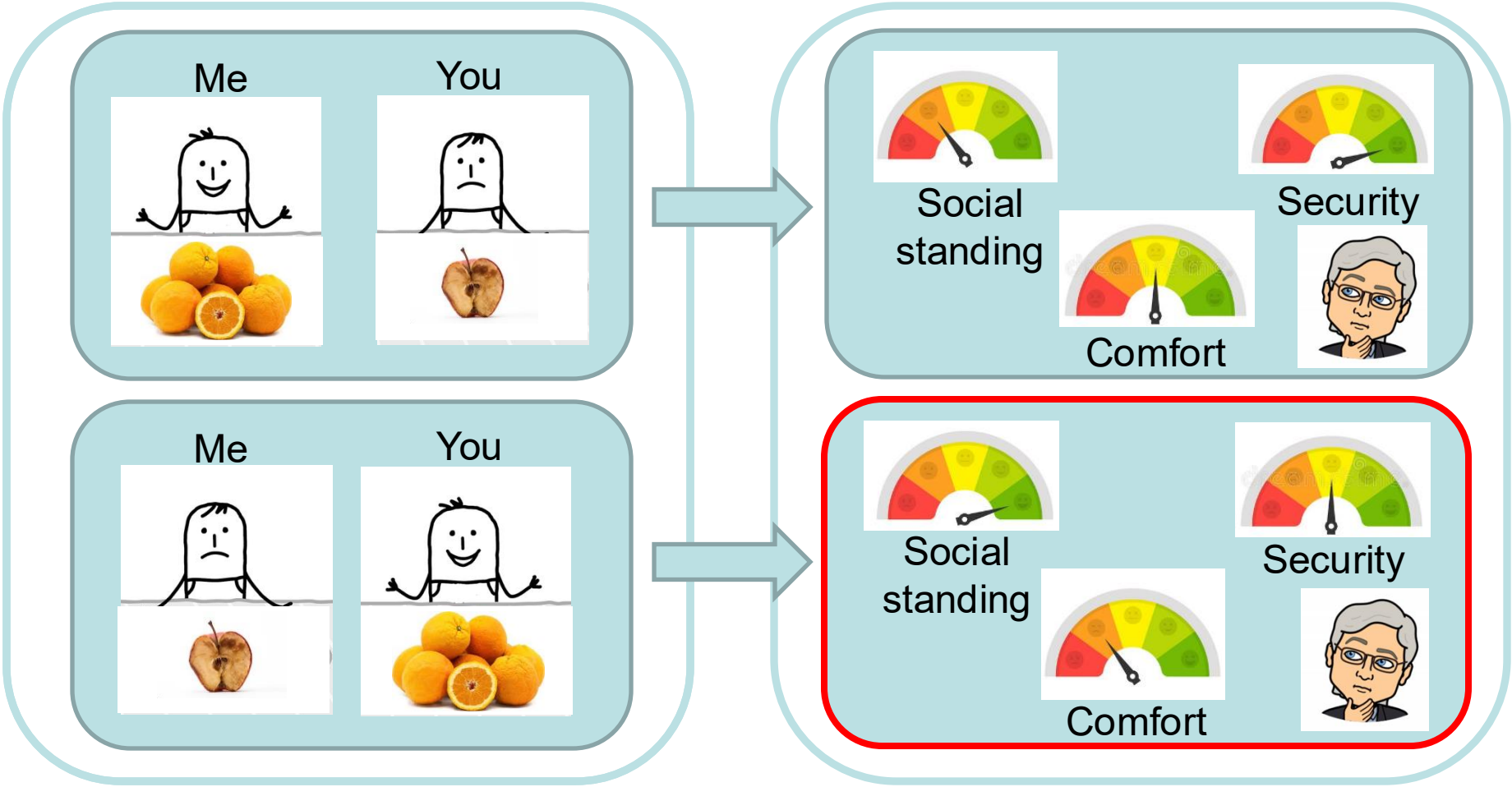




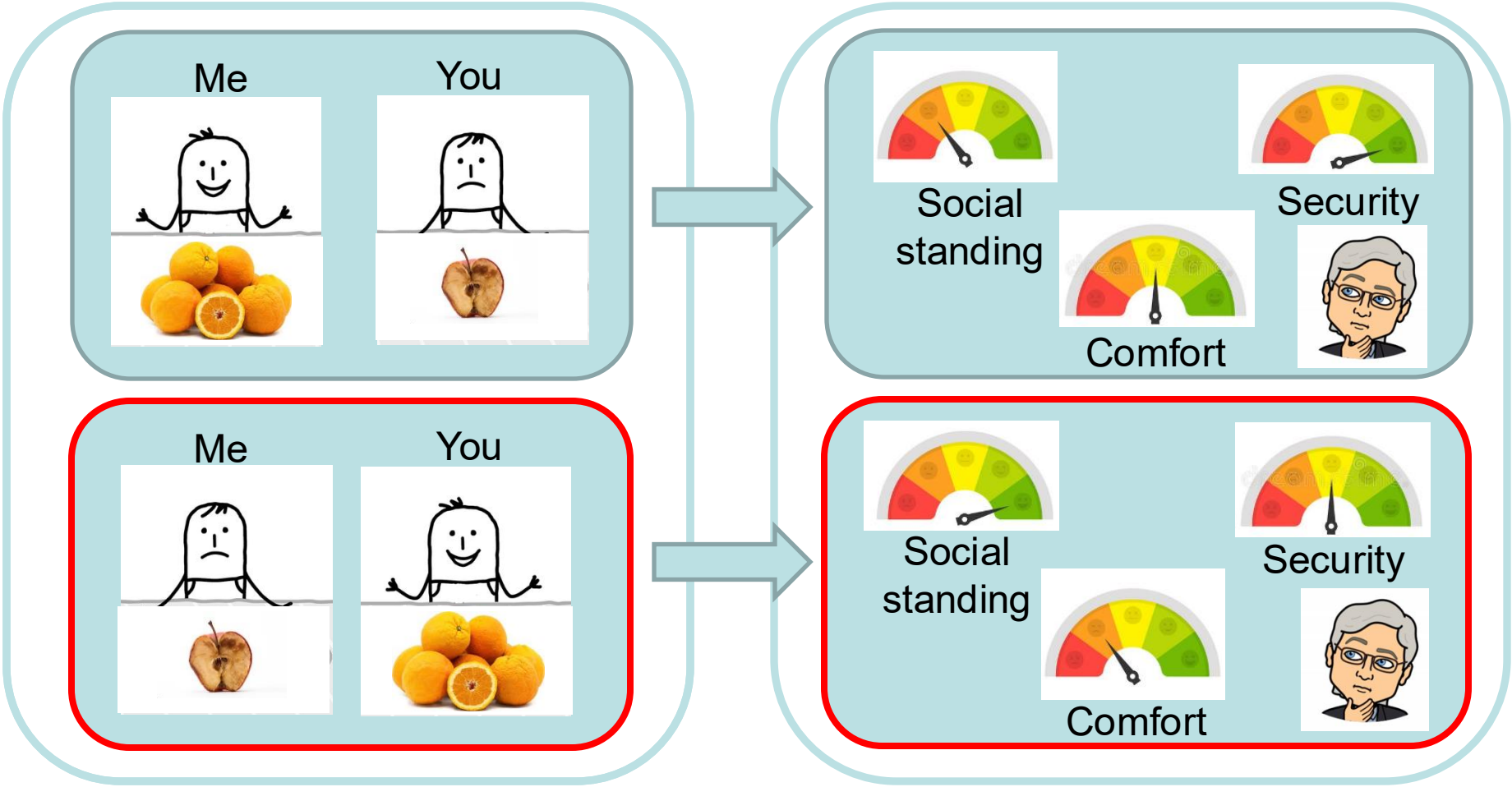


Same as generous option when I divide the pie myself

Challenge 3: Reproducibility Critiques



Challenge 3: Reproducibility Critiques



Challenge 3: Reproducibility Critiques

Method:

1. Assess (proxies for) the mental state bundles the individual expects to follow from each option in a collection of choice problems
   - For the lunch problem, choices would involve food items
2. Estimate preferences, $\succcurlyeq$, over mental state bundles using standard techniques (building on Benjamin et al., 2012, 2014).
3. To determine which of a Planner's options is better for the individual, assess the mental states the Planner's options induce and then determine the best one according to $\succcurlyeq$.
   - For the lunch problem, one might find that, although I always choose Salad over Pizza, I prefer the mental state bundle I associate with being assigned Pizza to the one I associate with being assigned Salad.

Challenge 3: Reproducibility Critiques

Addressing a possible concern:

- What's to prevent us from running into the same problem – i.e., menu-dependence in choices over mental state bundles? (i.e., preferences defined over objects of the form (z, Z))
- Under the mental statist premise, the set of mental-state bundles from which one can choose can only matter if it affects mental states, in which case z already incorporates it.
- In other words, while mental statism allows for the possibility that the mental state bundle z associated with any given option depends on the menu hierarchy in arbitrarily complex ways, it ensures that menus can *only* affect well-being through the mental states, z.
- Therefore, if we measure z for all the options in each choice problem, we've already encompassed all potential menu dependence.

Challenge 3: Reproducibility Critiques

Key findings from experimental proof of concept:

1. In Dictator Game (DG) settings, having an alternative changes the utility derived from an option. With either option, people are better off if someone else chooses it for them.
2. In DG settings, menu-dependence misleads the planner into paying too much to replace the payout-maximizing option with the pro-social option. (In other words, the NCP is empirically important).
3. Having an opt-out option (weakly) reduces the utility associated with the available options, and the effect differs across the options.
4. Opt-out choices provide misleading measures of the value derived from being assigned to the DGs. (In other words, meta-choices do not properly resolve the NCP.)

Challenge 3: Reproducibility Critiques

But what about the NCP associated with false beliefs?

If we place intrinsic value on the correctness of our beliefs, we can't solve that problem using the mental statist approach.

Requires other strategies…

Challenge 3: Reproducibility Critiques

Other strategies (Arrieta, Bernheim, & Bolte, in progress):

1. Use *surrogate choices*

   – In some cases, it's possible for people to make choices for others that they can't make for themselves (e.g., they can induce false beliefs)

   – *False consensus bias* helps to ensure that people ask, "what would I want someone to do for me?" (Ambuehl, Bernheim, & Ockenfels, AER 2021)

2. Use *stated preferences* (or *hypothetical choices*)

   – We can state preferences over options even when we can't choose among them

   – While stated preferences are susceptible to a variety of biases, it may be possible to use subjective evaluations to predict choice accurately (Bernheim, Bjorkegren, Naecker, & Pollmann, 2024)

Other Applications of the Mental Statist Approach

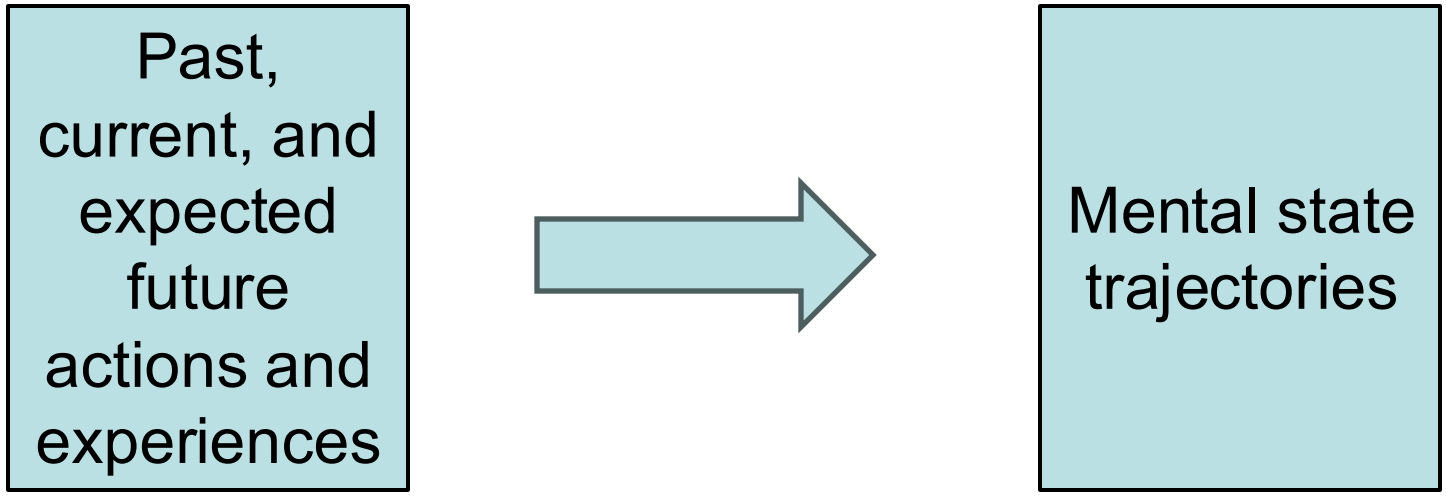
Other Applications of the Mental Statist Approach

The mental statist approach also potentially allow us to resolve other challenges.

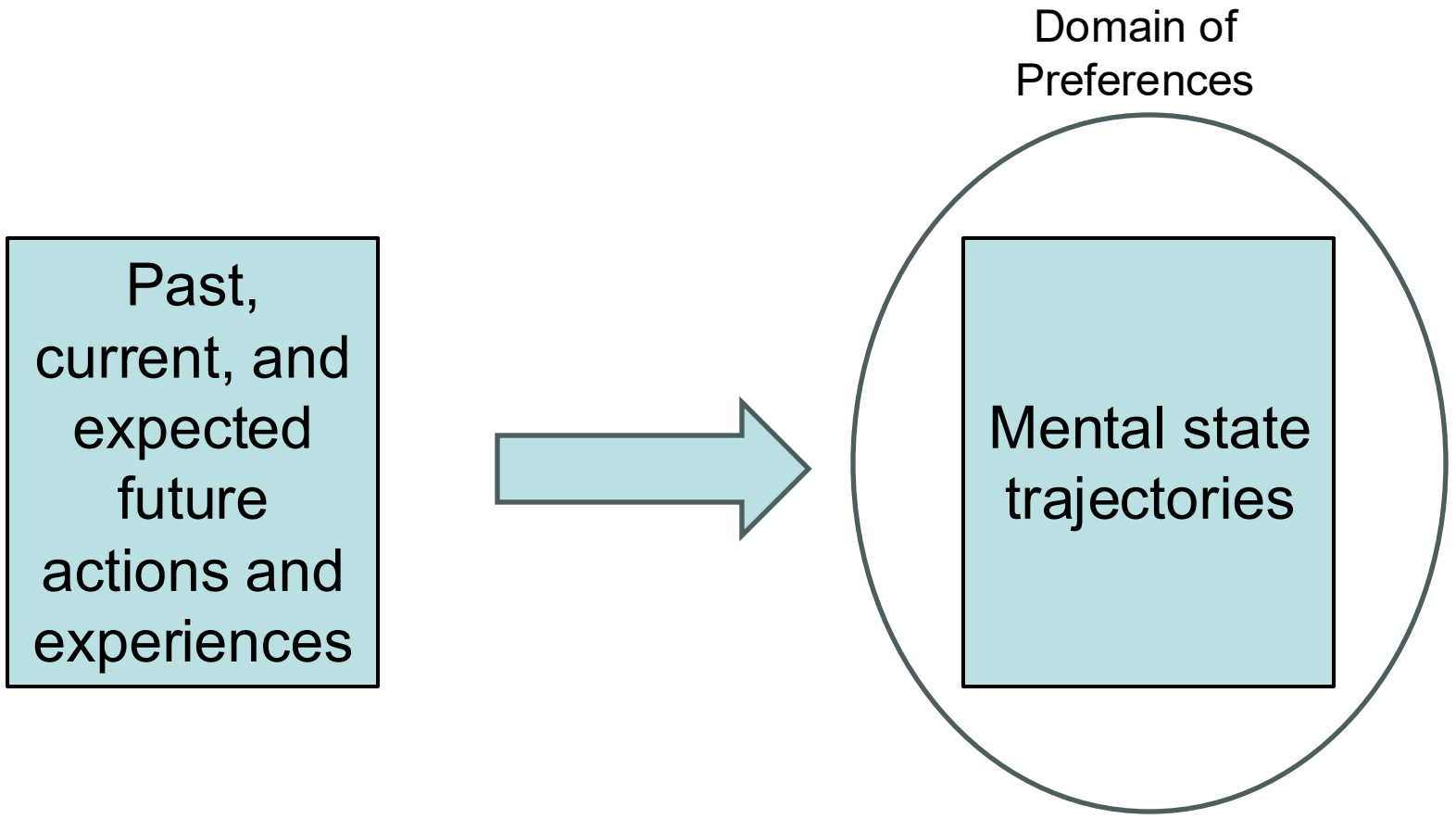
Example: Endogenous Preferences (technically, back to Challenge #2)

(Based on Bernheim, Bolte, Nagel, & Ray, in progress)

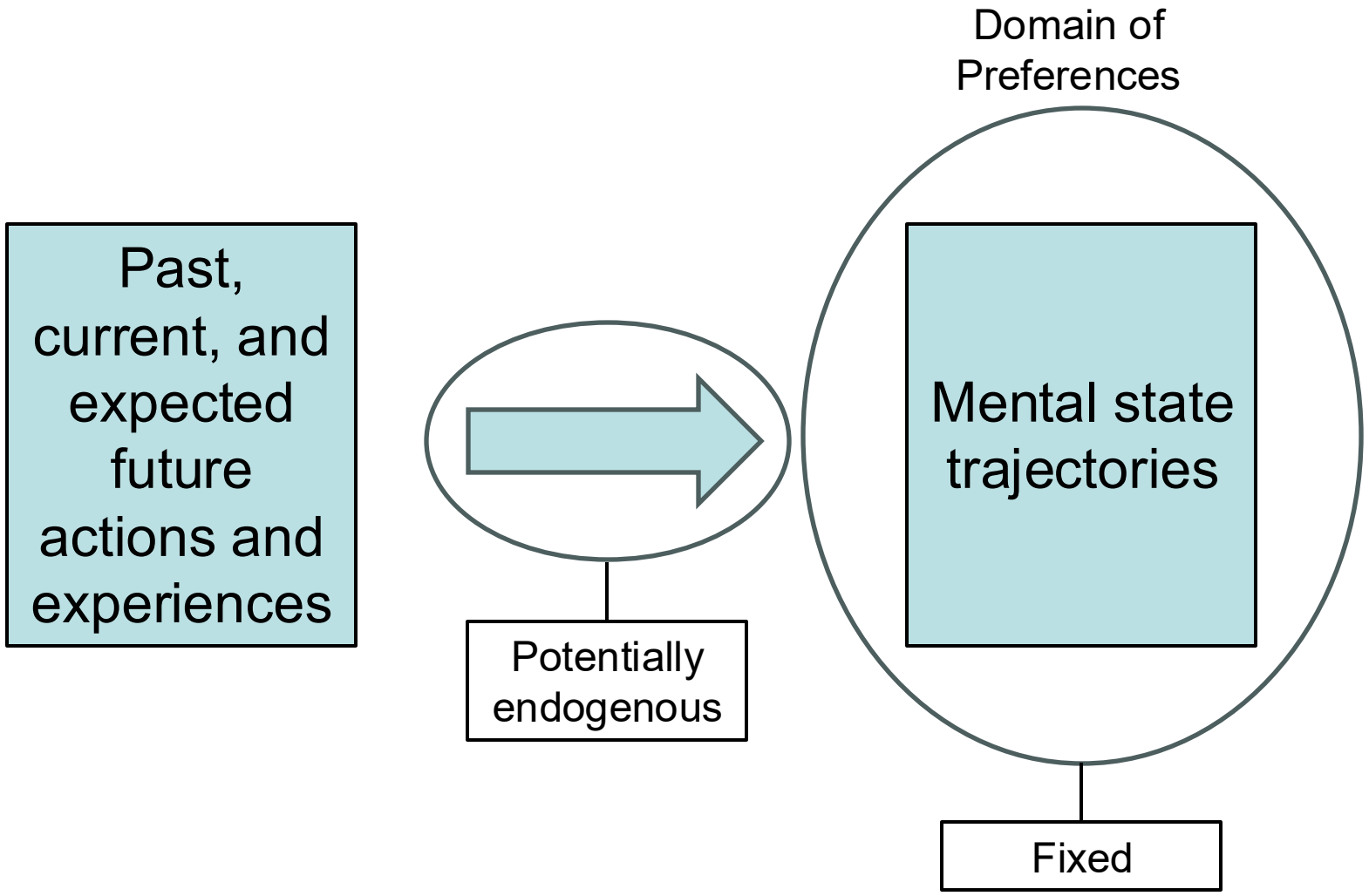
Other Applications of the Mental Statist Approach



Other Applications of the Mental Statist Approach



Other Applications of the Mental Statist Approach



Concluding Remarks

- **Challenge 1:** *Implementation critiques*
- **Challenge 2:** *Coherenece critiques*
- **Challenge 3:** *Reproducibility critiques*

Concluding Remarks

- **Challenge 1:** *Implementation critiques*
- **Challenge 2:** *Coherence critiques*

- Identify characterization & optimization failures
- Apply the unambiguous choice criterion

Concluding Remarks

- **Challenge 1:** *Implementation critiques*
- **Challenge 2:** *Coherence critiques*

- Identify characterization & optimization failures
- Apply the unambiguous choice criterion

- **Challenge 3:** *Reproducibility critiques*

- Recover preferences over mental states
- Surrogate choices
- Hypothetical resposnes