The Challenges of Behavioral Welfare Economics

Prof. B. Douglas Bernheim, Stanford University

IIPF Annual Congress, August 2023

Introduction

- Traditional Public Economics posits neoclassical decision makers
- A rapidly growing branch of the literature asks, what if people…
  - …save to little due to “weakness of will”?
  - …regularly succumb to temptation when consuming harmful products?
  - …lack the skills to manage their portfolios?
  - …misunderstand risks they consider insuring?
  - …attend insufficiently to important information (especially if it’s disturbing)?
  - …etc., etc.
- Are there then additional justifications for “behavioral public policies” such as
  - …mandatory retirement saving
  - …aggressive consumer and investor protection
  - …banning harmful substances
  - …imposing “sin” taxes or “internality-correcting” subsidies?
  - …strategic design of default options
  - …all manner of “nudges”
  - …etc. etc.

Introduction

- There is a tendency to evaluate such policies based on simplistic preconceptions about “good” and “bad” outcomes
  - e.g., “people don’t save enough”
- These preconceptions improperly focus our attention (exclusively) on Average Treatment Effects
  - e.g., “higher default contribution rates for pension plans lead to greater saving, and are therefore ‘good’”
- But how do we determine which outcomes are good, and which are bad?
- This question falls within the domain of ***Welfare Economics*** (aka ***normative analysis***)

Introduction

- 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 introducing new considerations that provide fresh perspectives on the effects of public policies, ***Behavioral Economics*** also challenges the foundations of Standard Welfare Economics. So how do proceed?

- ***Behavioral 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.

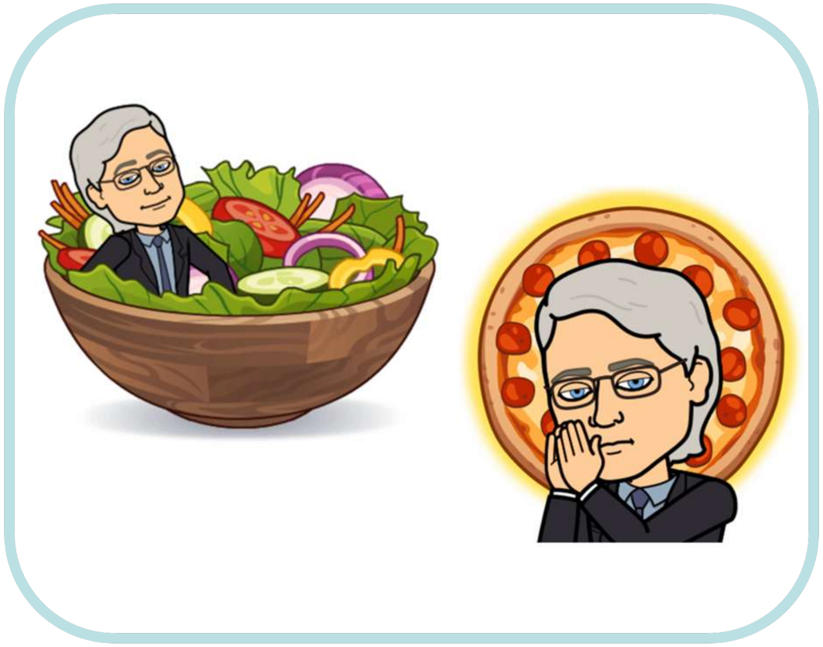
Standard Welfare Economics

Standard Welfare Economics

The planner's task

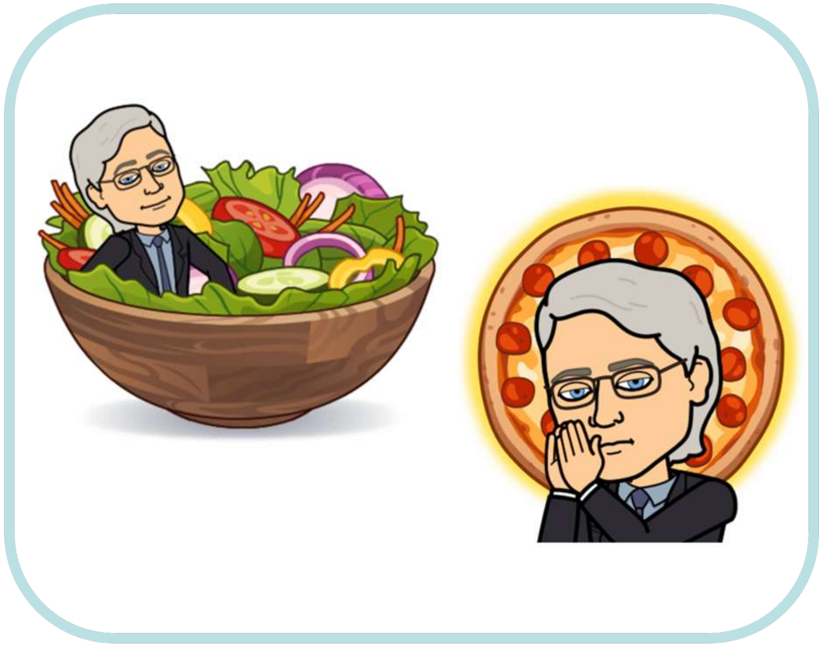
Standard Welfare Economics

The planner's task

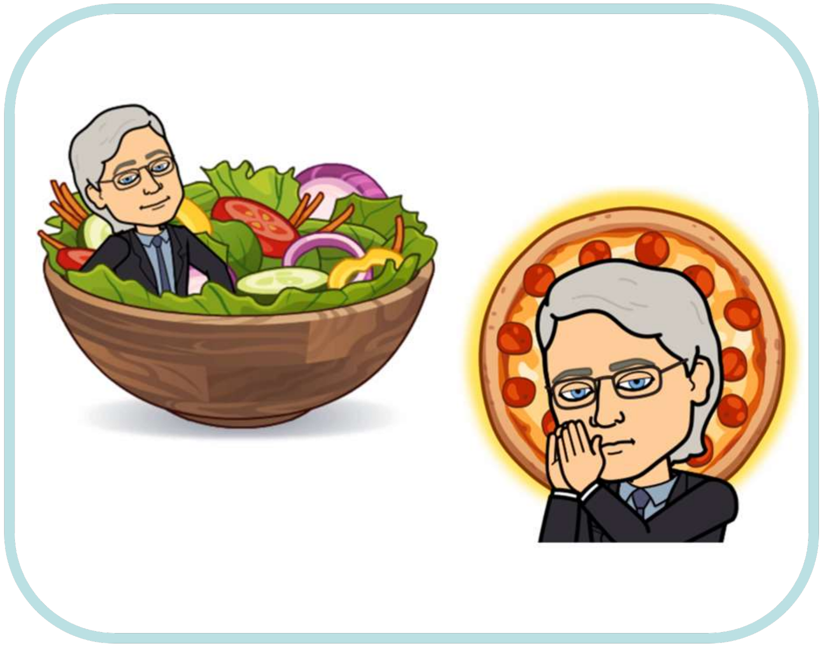


Standard Welfare Economics

The planner's task

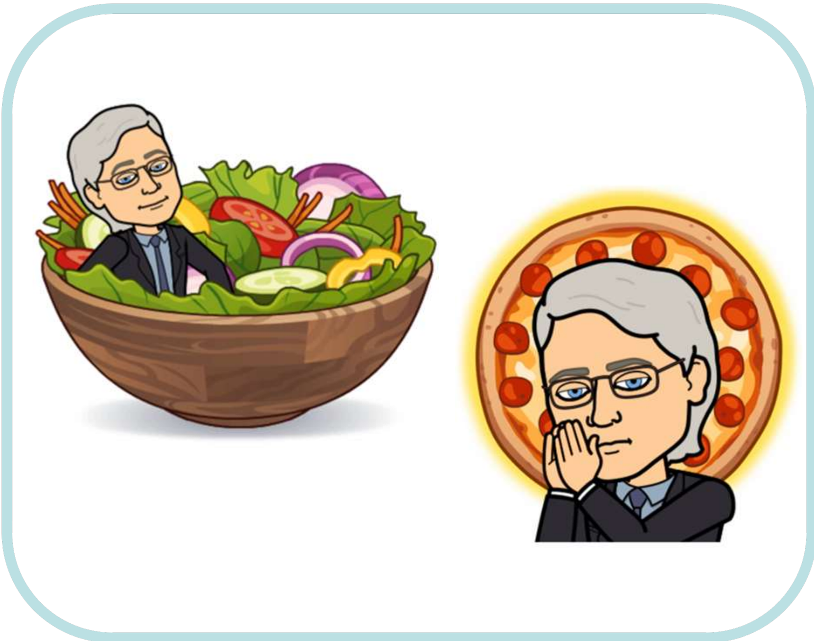


My task

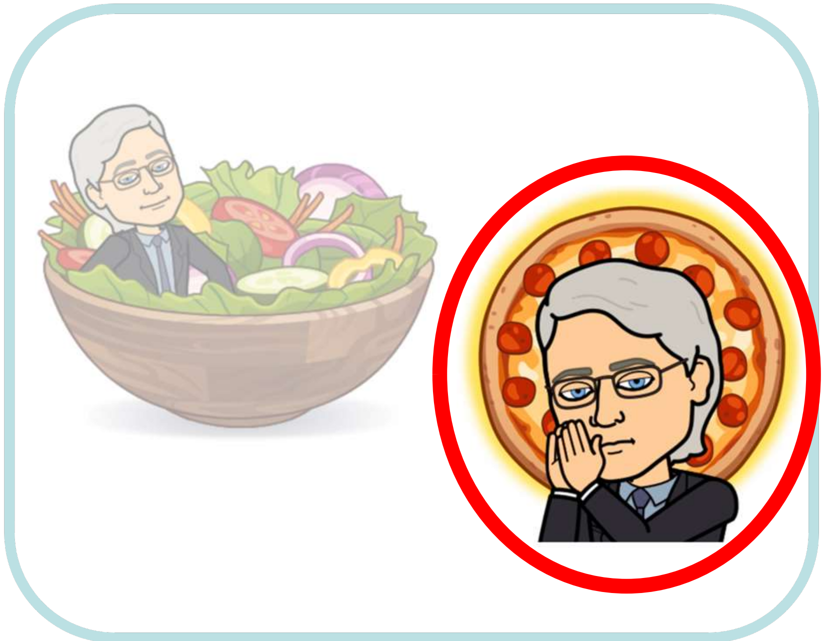


Standard Welfare Economics

The planner's task

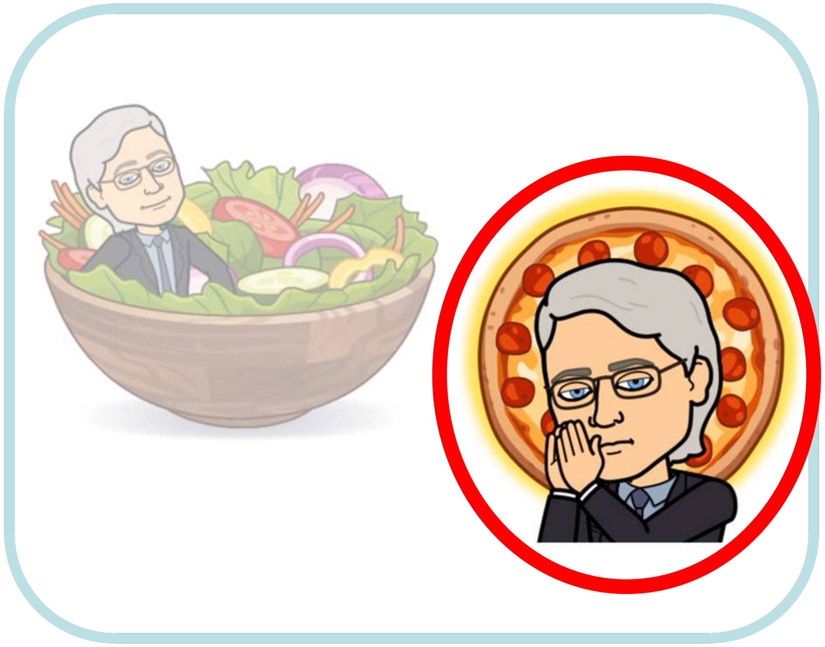


My task

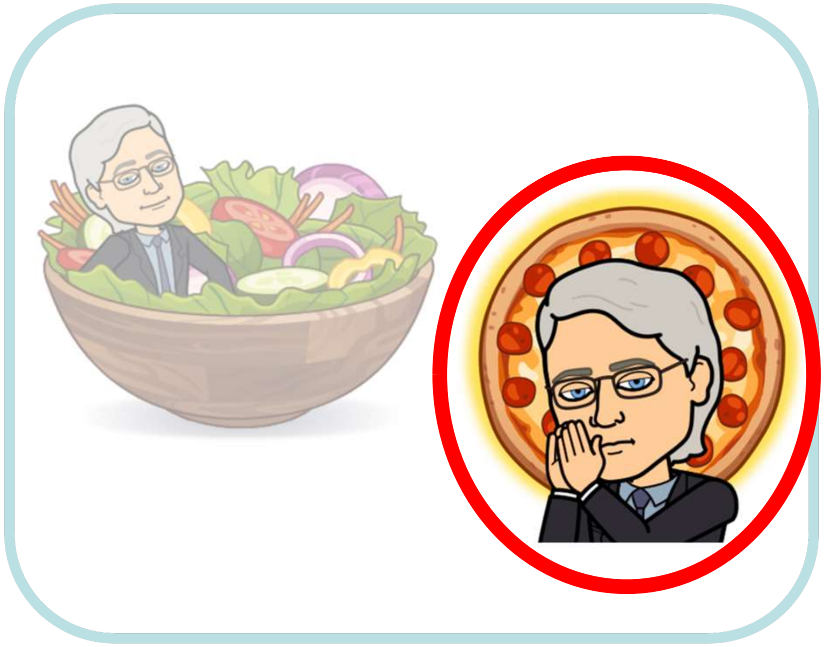


Standard Welfare Economics

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

The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

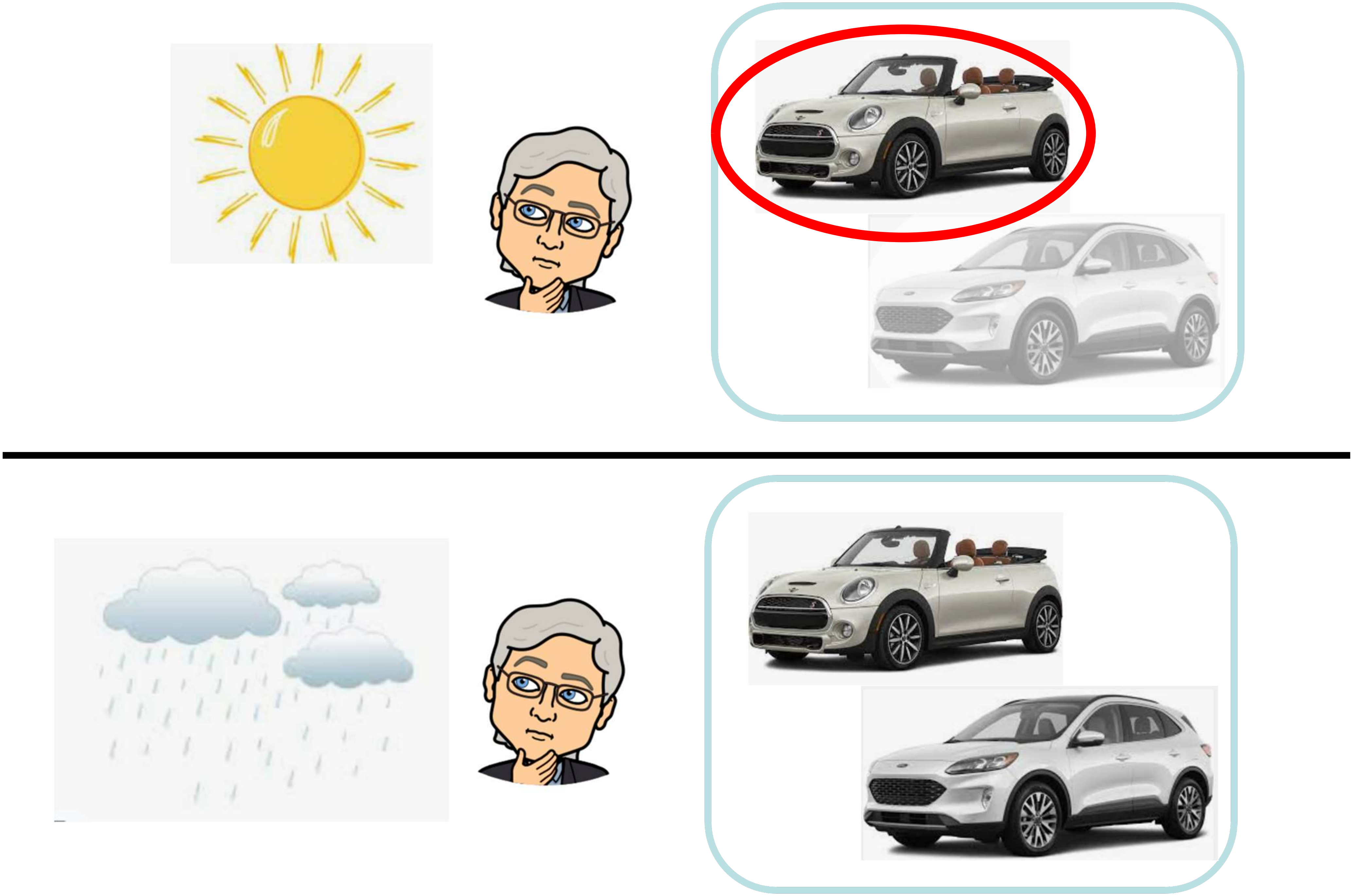
Context Dependence

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

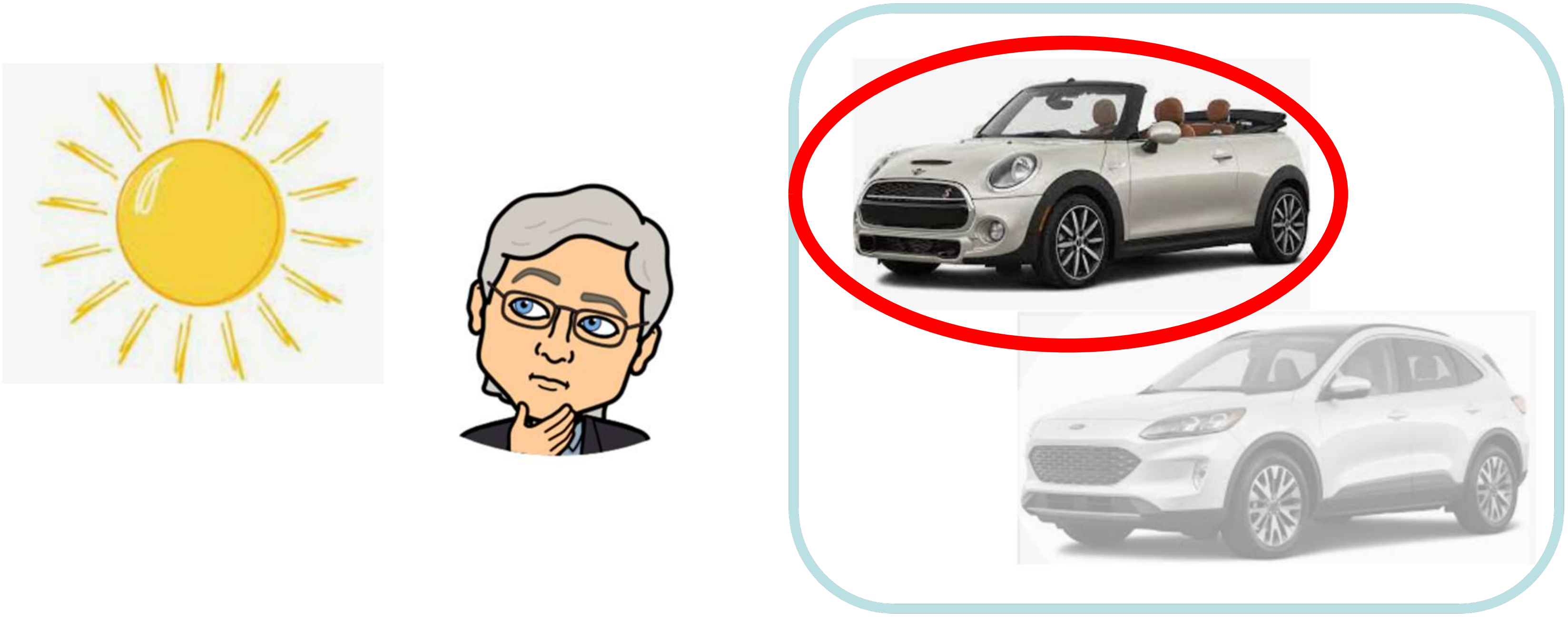
Context Dependence

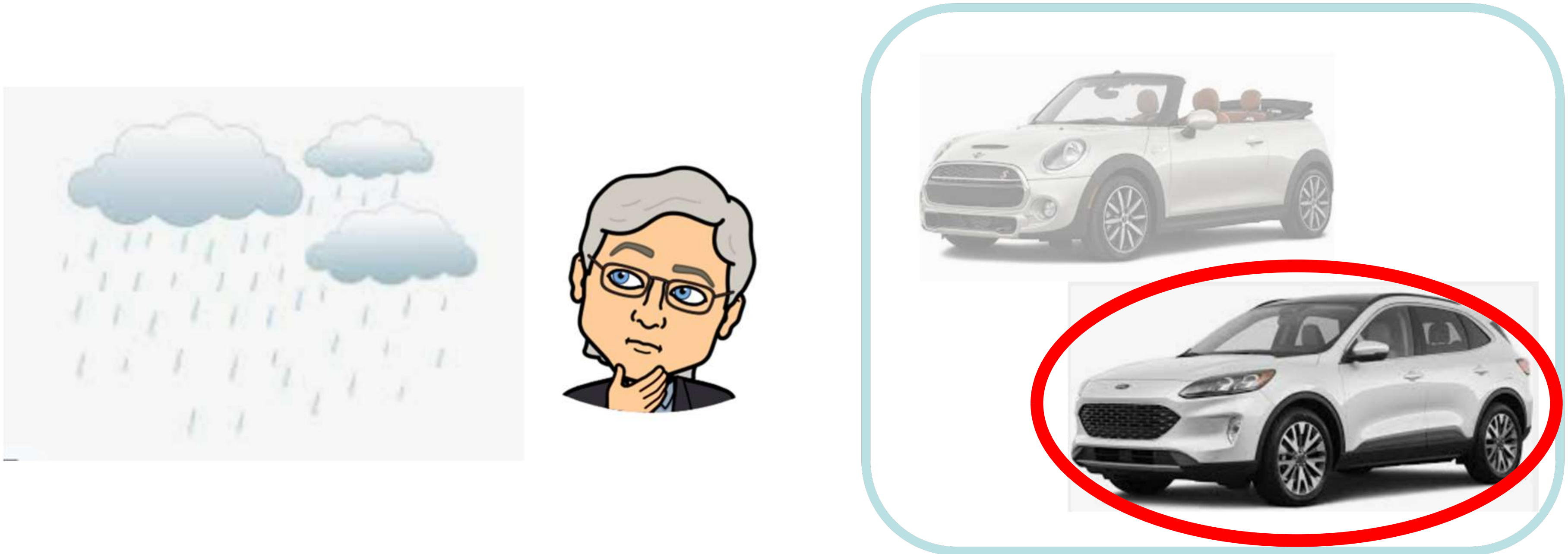
Context Dependence

Context Dependence

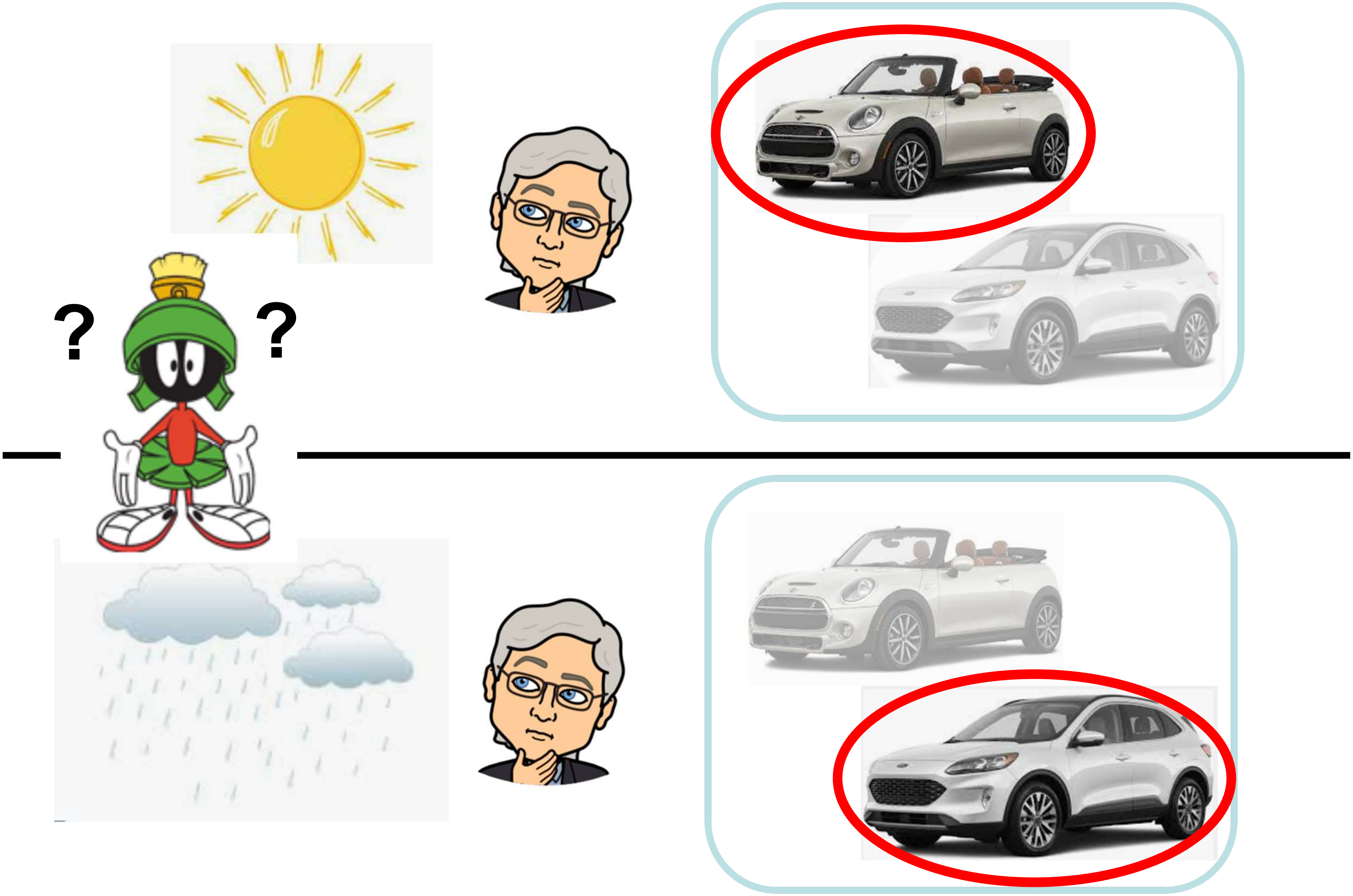


Context Dependence





Context Dependence



The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

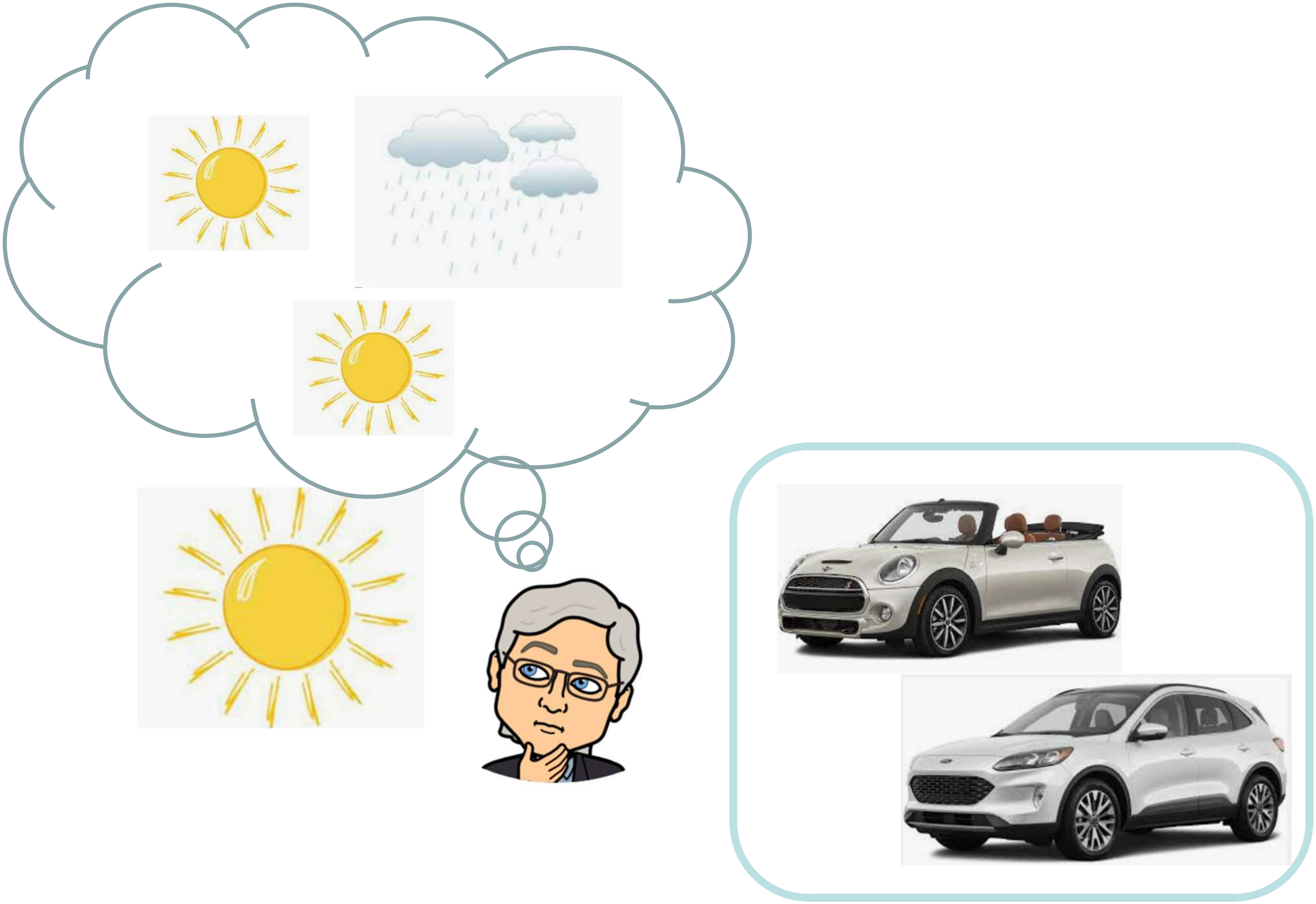
The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

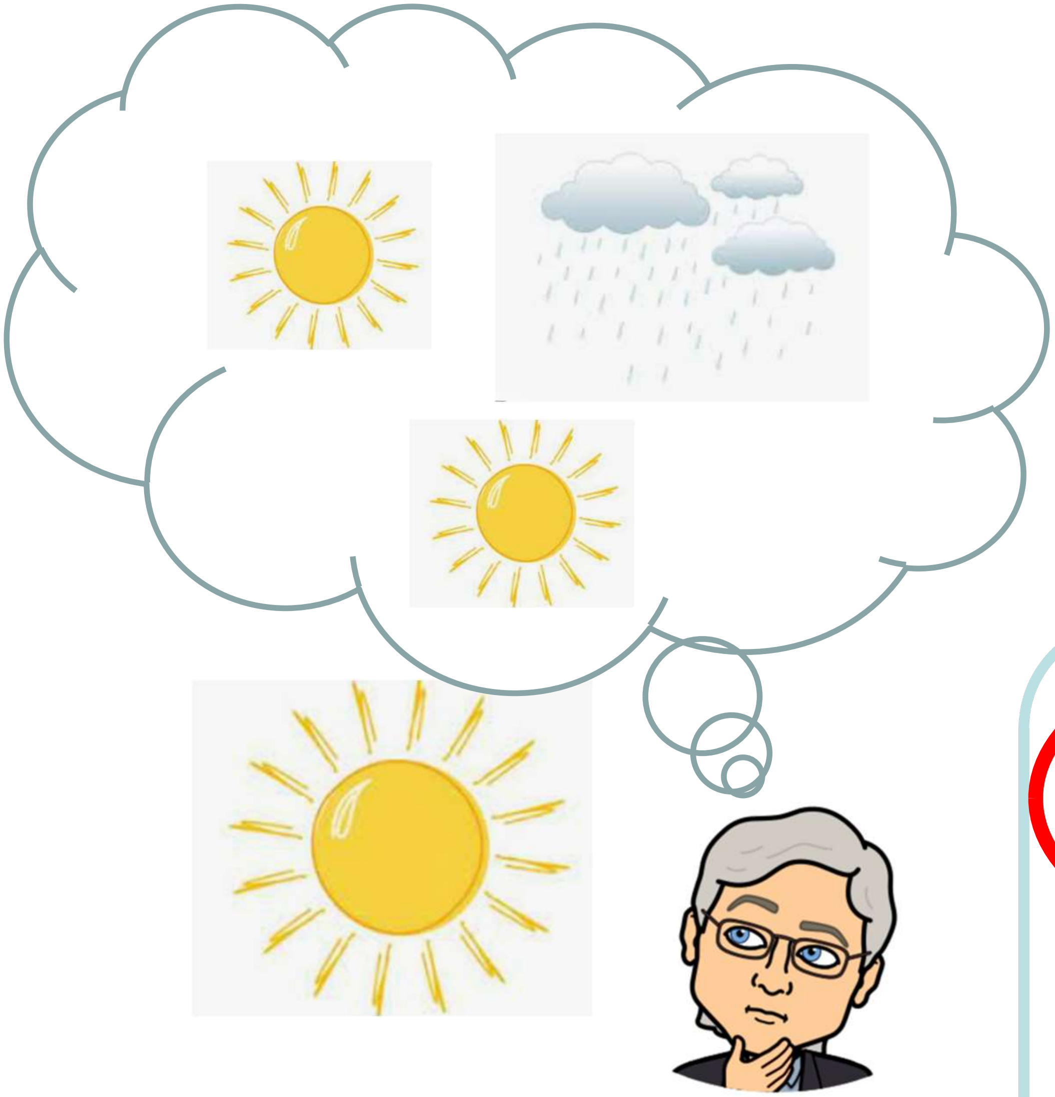
- Type M: certain framings trigger *mistakes*

Context & Mistakes

Context & Mistakes

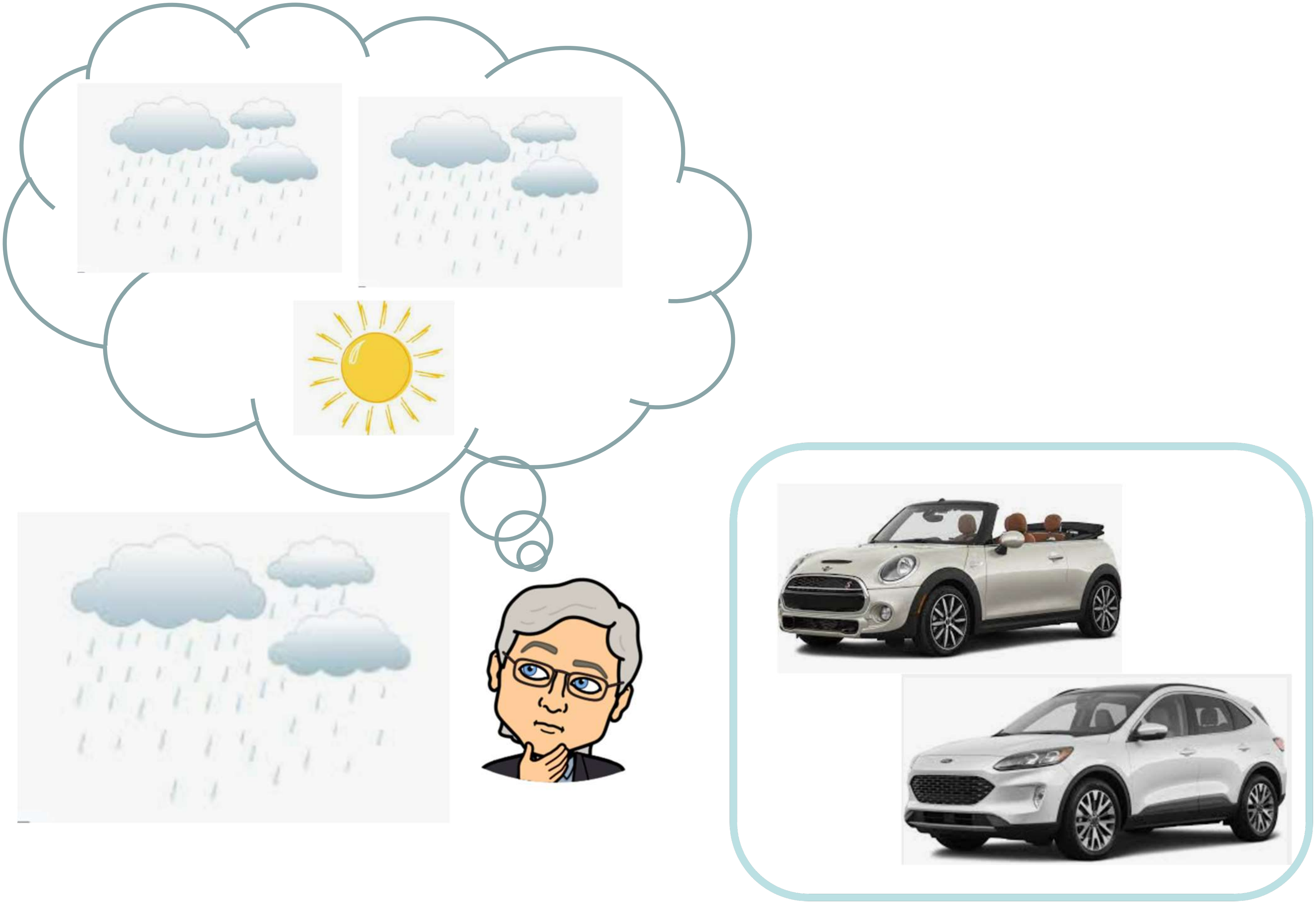


Context & Mistakes

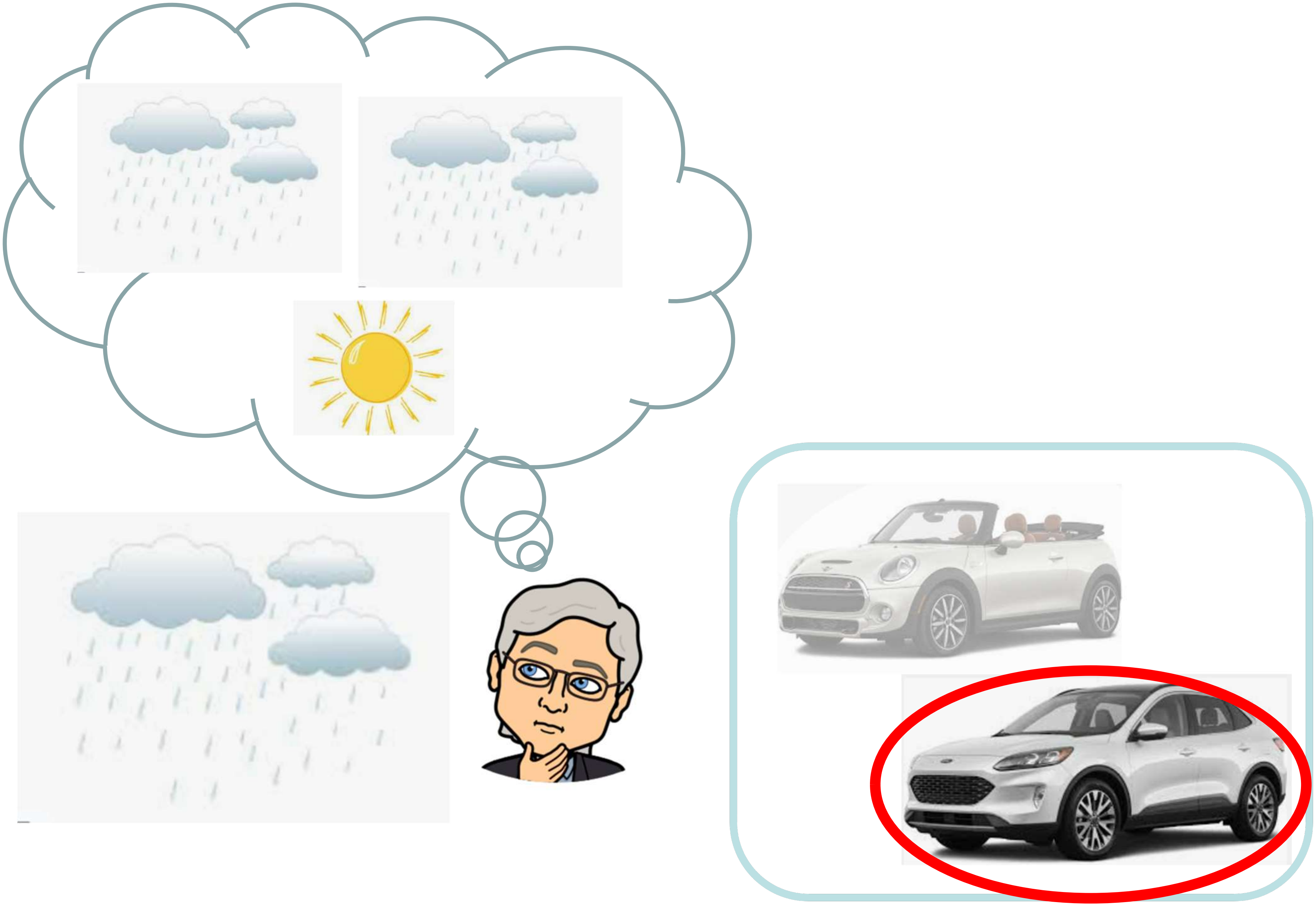


Context & Mistakes

Context & Mistakes



Context & Mistakes



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.*

The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

- Type M: Certain framings trigger *mistakes*

Challenges Premise #3 – “Implementation Critiques”

The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

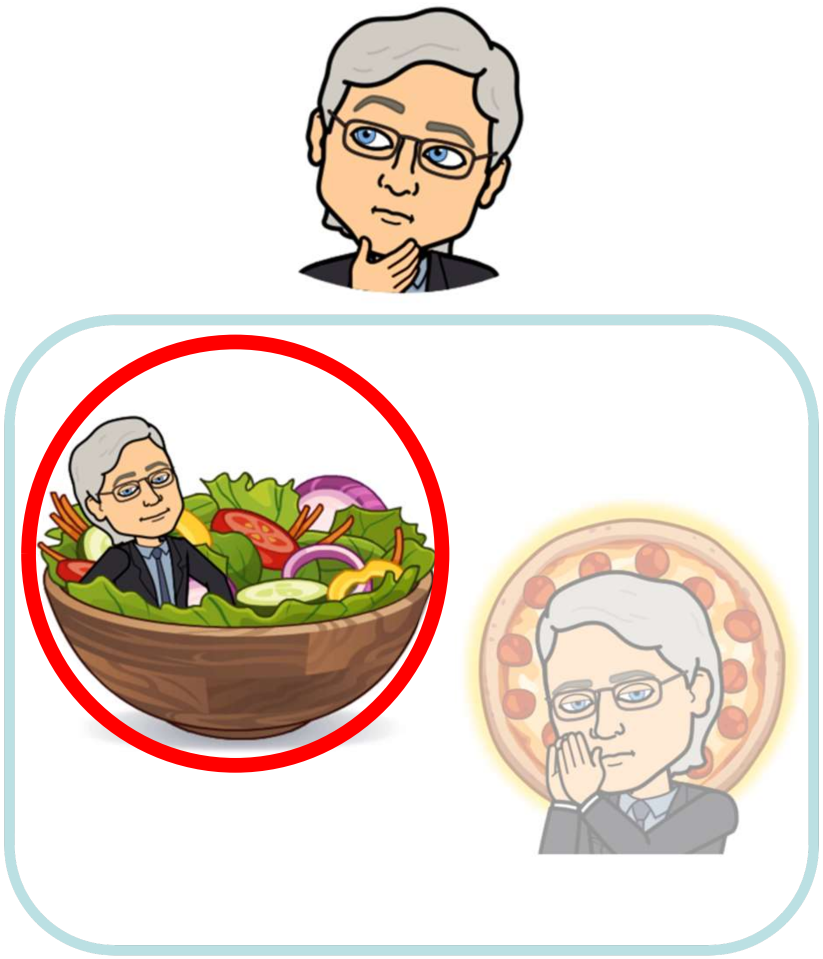
- Type M: Certain framings trigger *mistakes*

Challenges Premise #3 – "Implementation Critiques"

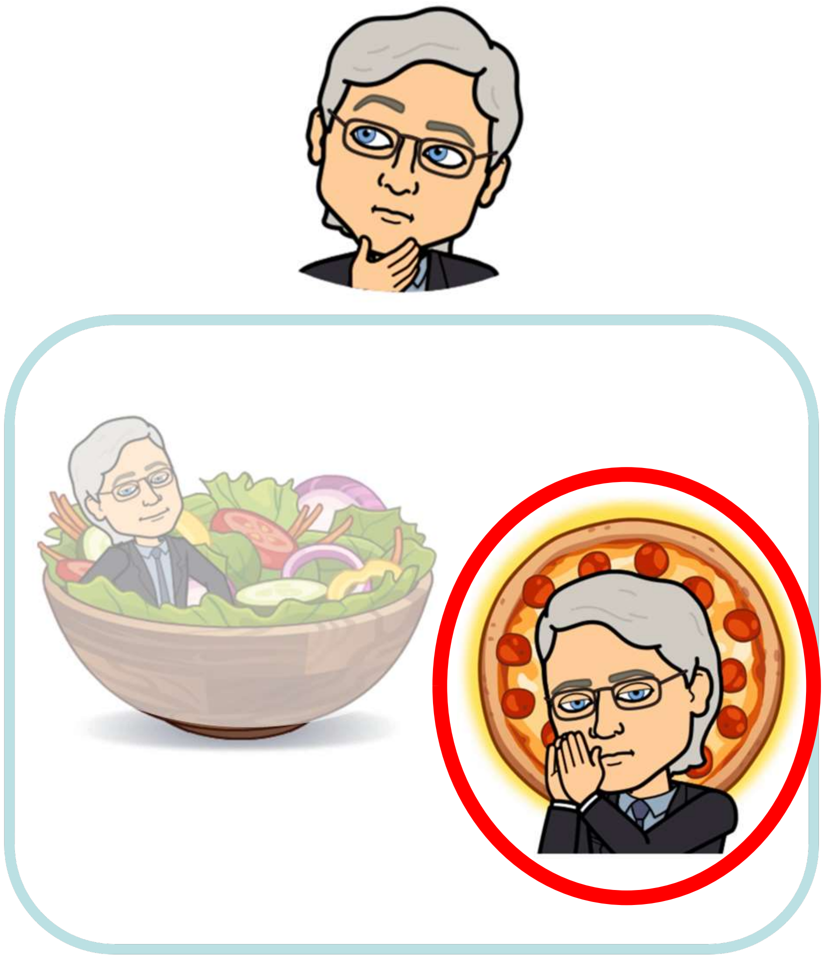
Implementation Critiques do not always reference misunderstandings of decision problems—at least not explicitly

“Weakness of Will”

Intention as of every morning

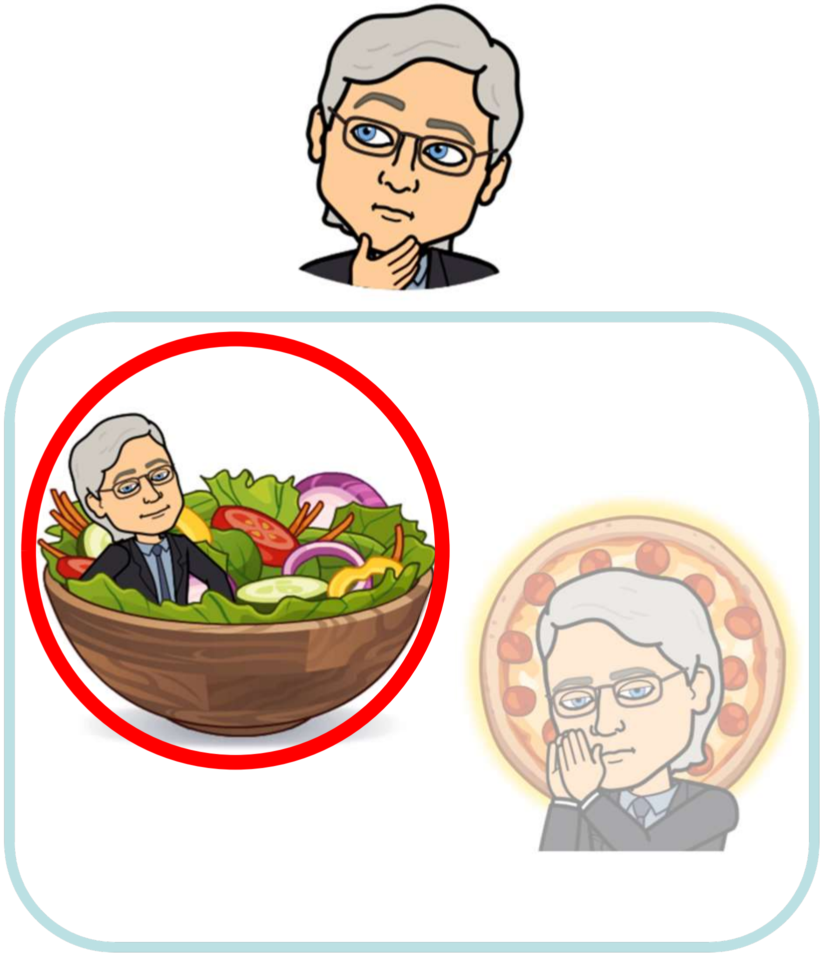


Choice at lunchtime

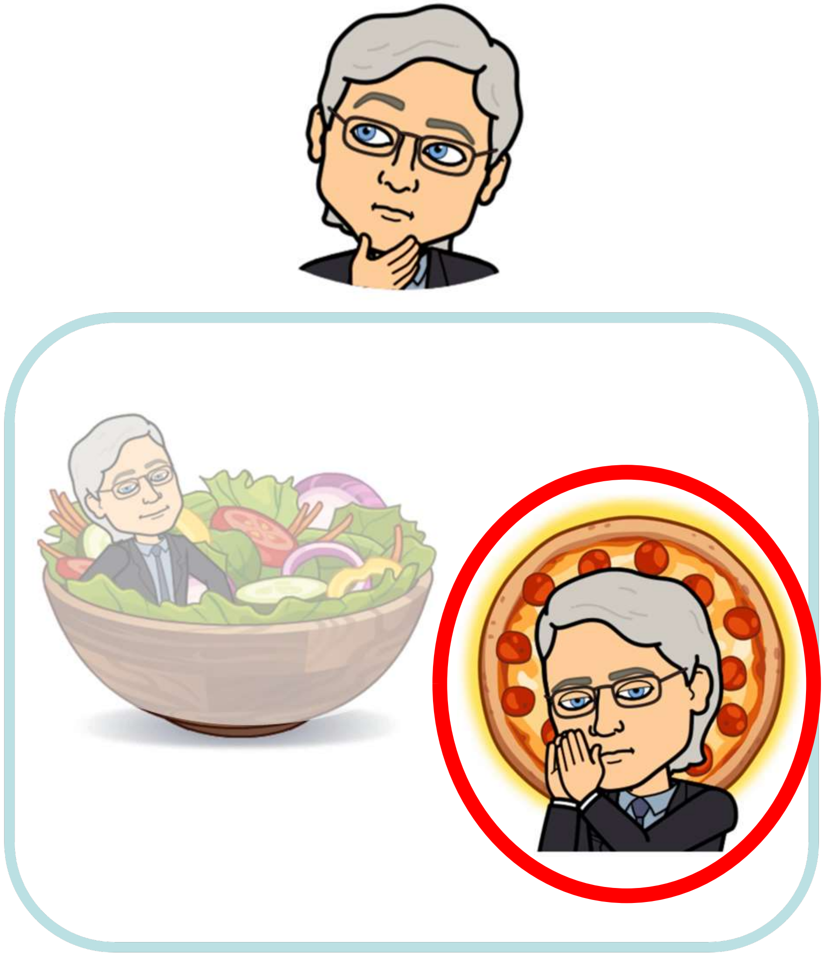


“Weakness of Will”

Intention as of every morning



Choice at lunchtime



But is this a *valid* Implementation Critique?

The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

- Type M: Certain framings trigger *mistakes*

  Challenges Premise #3 – "Implementation Critiques"

- Type C: People don't have preferences they can access – their *judgments are constructed* contextually

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

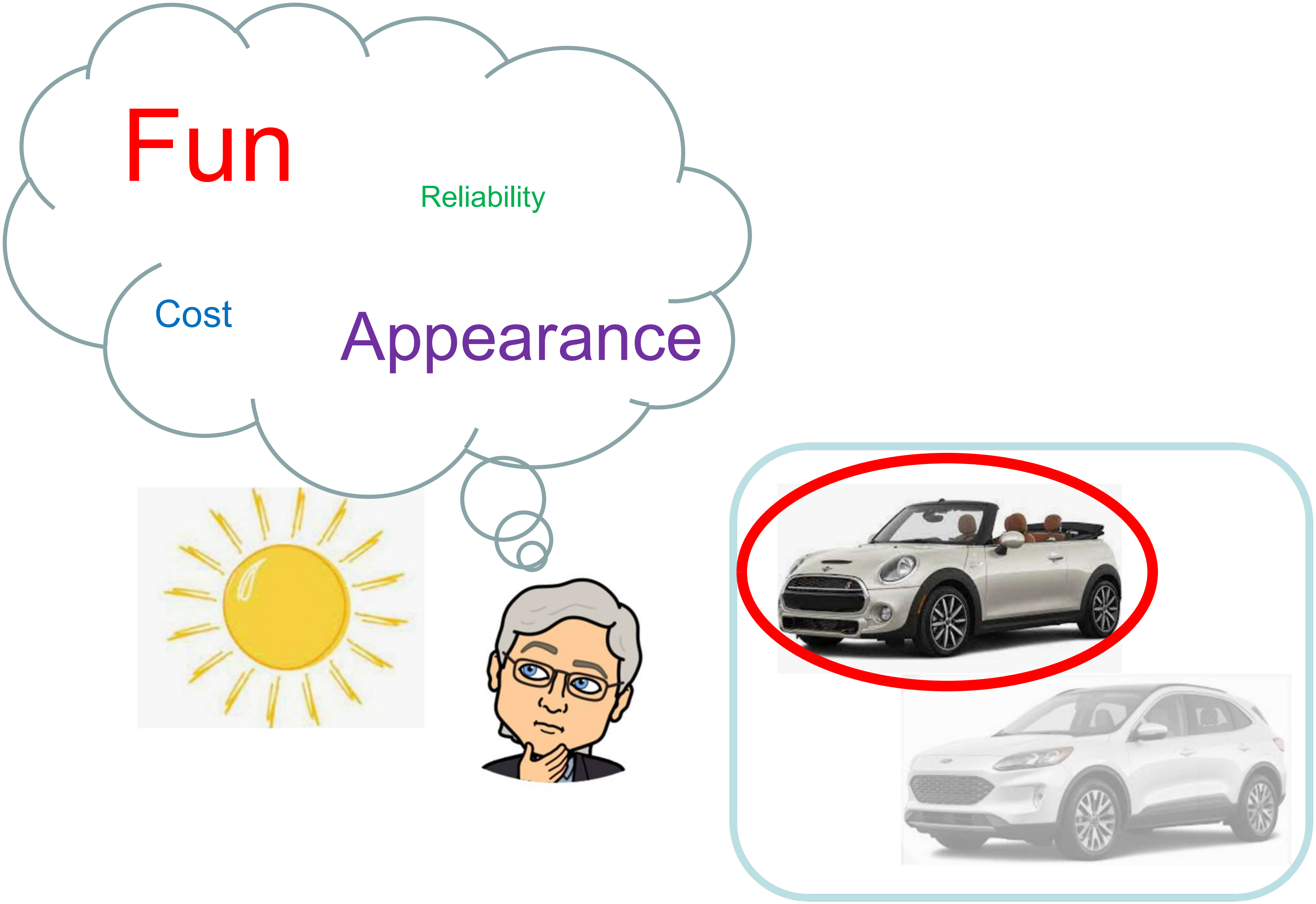
Fun

Reliability

Cost

Appearance

Context & Preference Construction



Fun

Reliability

Cost

Appearance

Context & Preference Construction

Context & Preference Construction

Fun

Reliability

Cost

Appearance

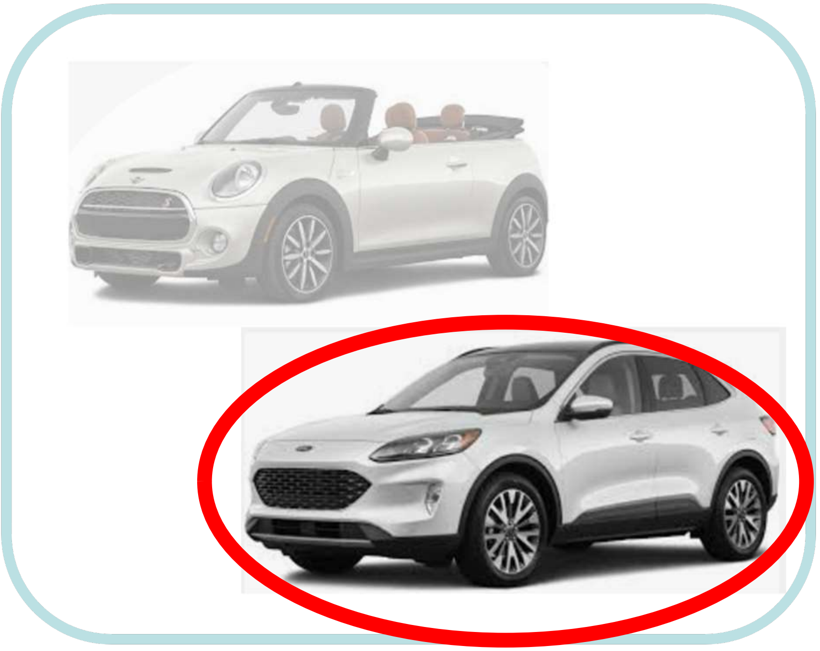
Context & Preference Construction

Fun

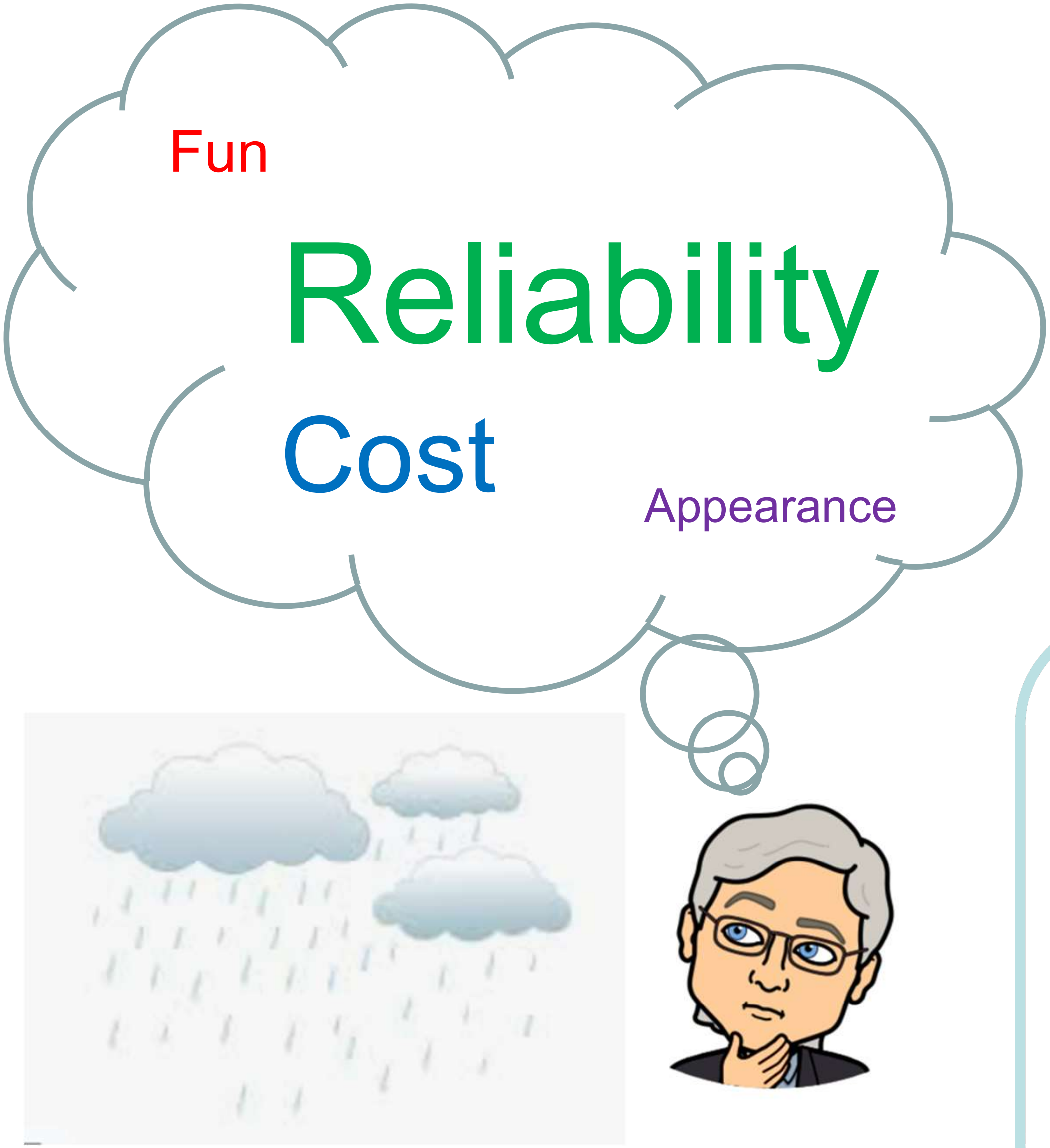
Reliability

Cost

Appearance



Context & Preference Construction



Fun

Reliability

Cost

Appearance

Irreducible inconsistency

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.*

The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

- Type M: Certain framings trigger *mistakes*

  Challenges Premise #3 – “Implementation Critiques”

- Type C: People don’t have preferences they can access – their *judgments are constructed* contextually

  Challenges Premise #1 – “Coherence Critiques”

The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

- Type M: Certain framings trigger *mistakes*

  Challenges Premise #3 – "Implementation Critiques"

- Type C: People don't have preferences they can access – their *judgments are constructed* contextually

  Challenges Premise #1 – "Coherence Critiques"

Behavioral Principle #2: The act of choosing for oneself alters the available options

The Premises Standard Welfare Economics

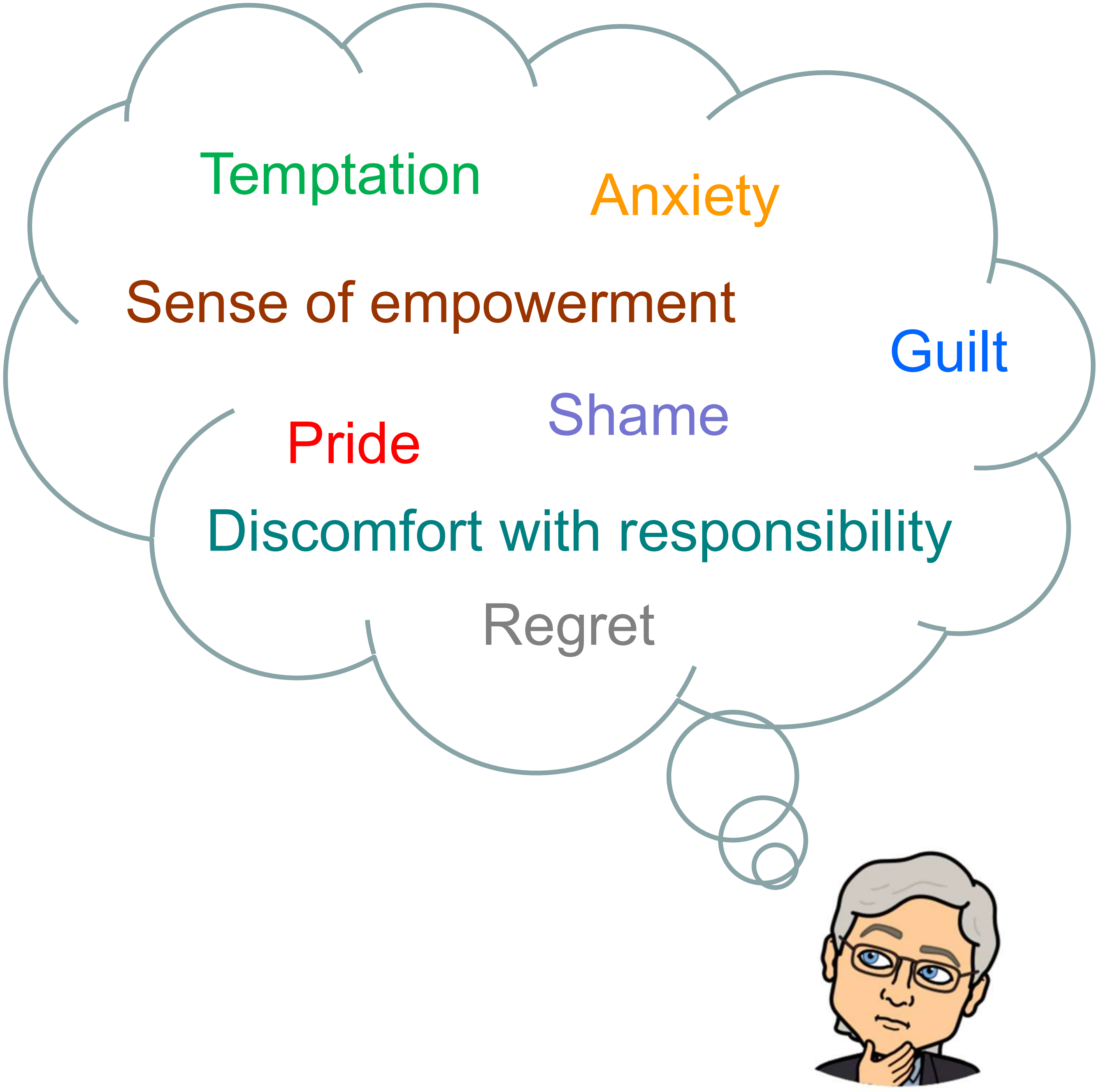
- ***Premise 1:*** *Coherent preferences, $\succsim$, govern each individual's judgments about their own well-being.*

- ***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.*

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

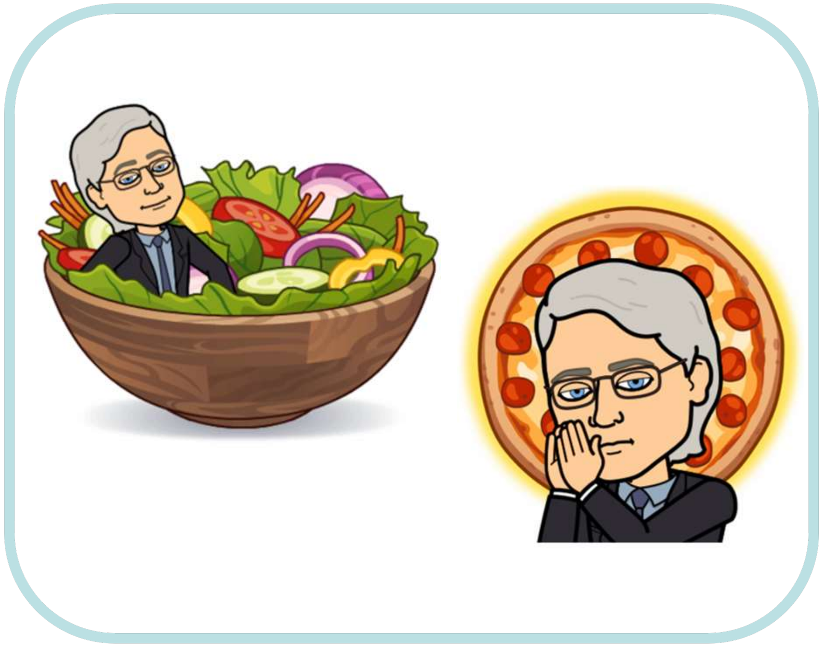
Discomfort with responsibility

Regret

Gives rise to a third type of context dependence – Type P

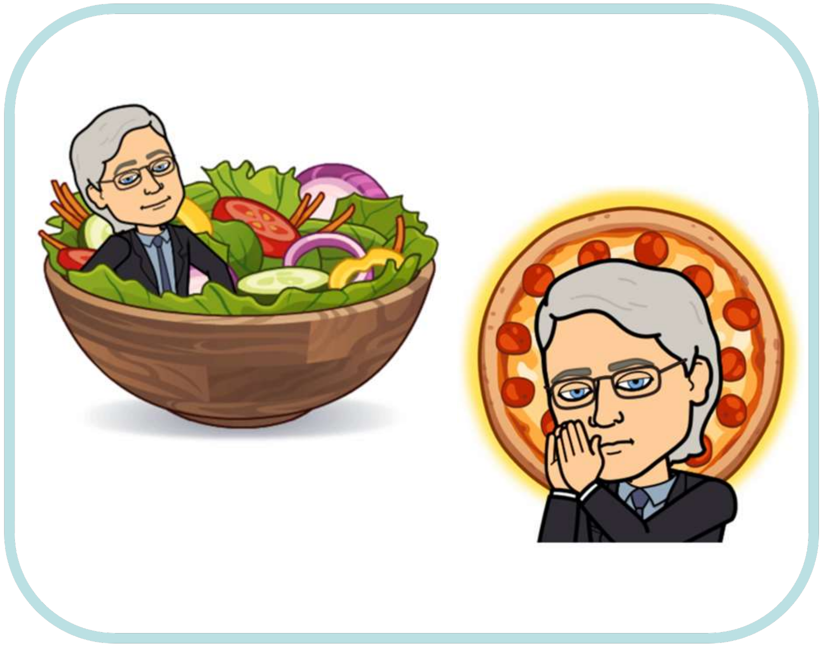
An Issue with Reproducibility

The planner's task

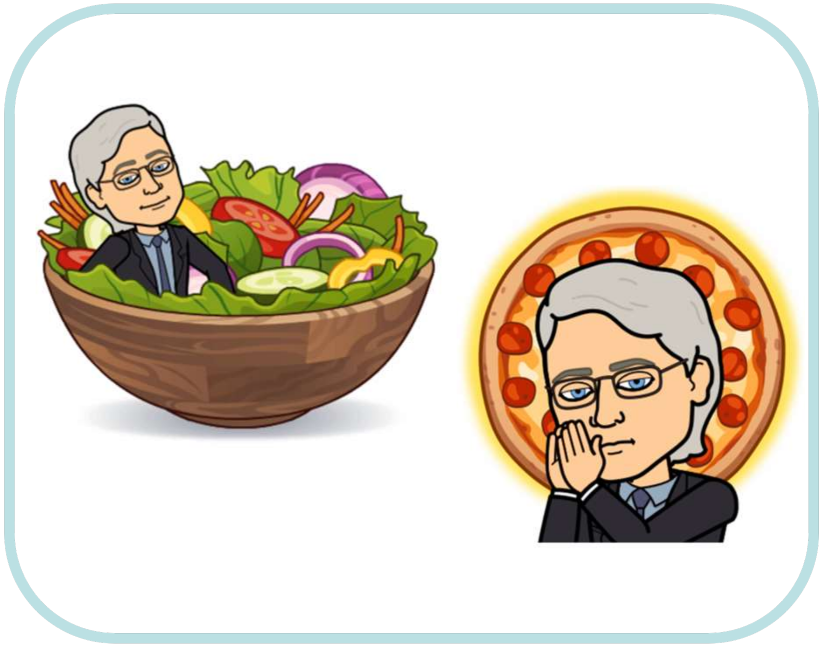


An Issue with Reproducibility

The planner's task



My task



An Issue with Reproducibility

The planner's task

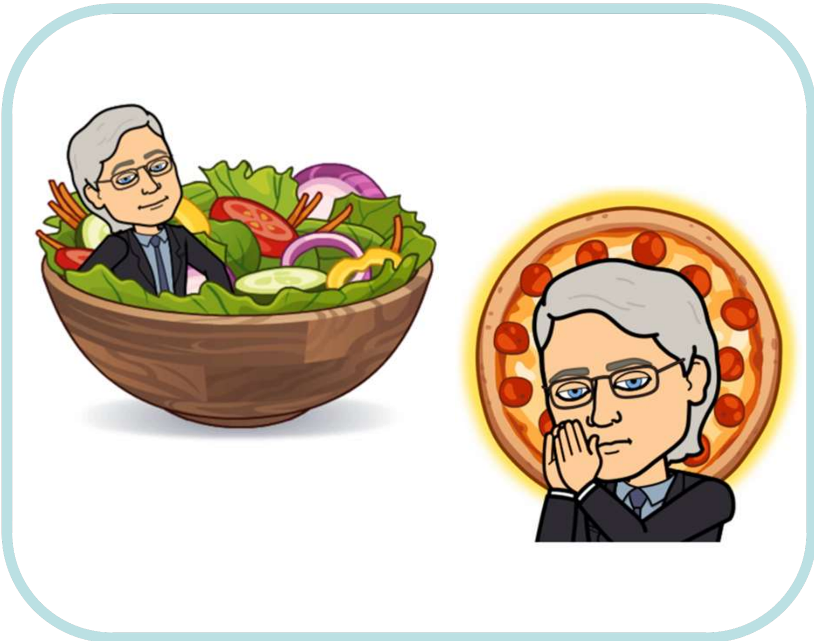
My task

+ temptation

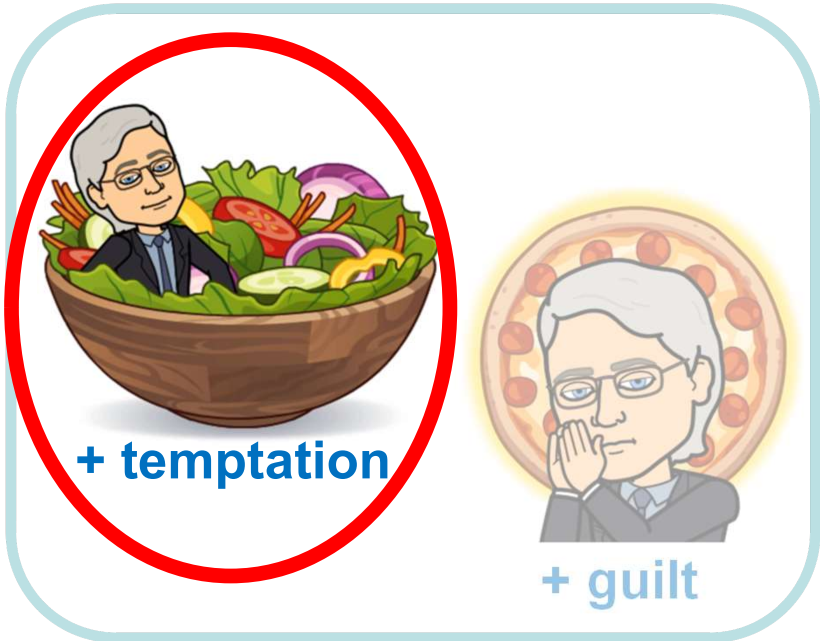
+ guilt

An Issue with Reproducibility

The planner's task

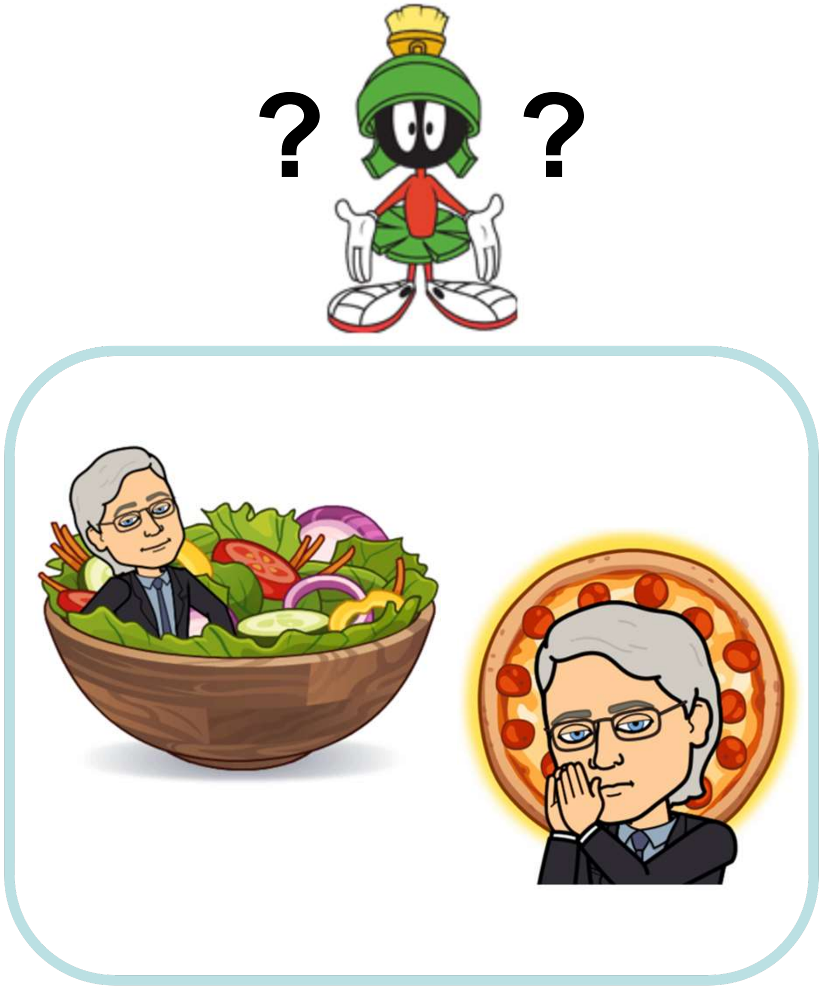


My task

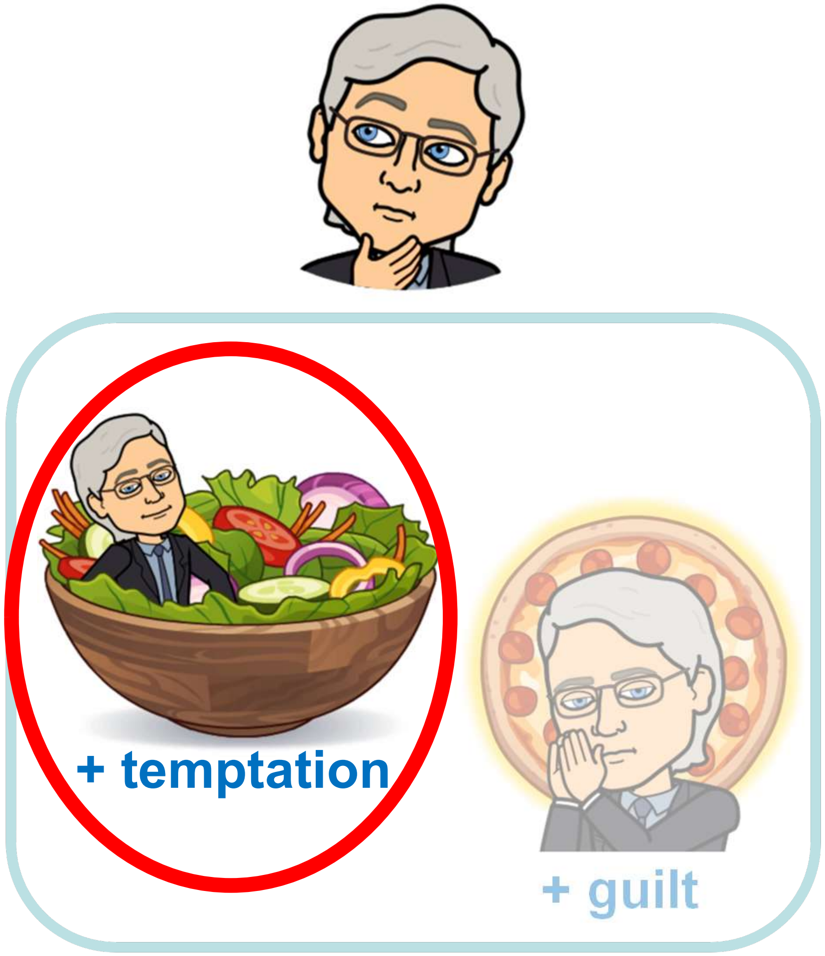


An Issue with Reproducibility

The planner's task



My task



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 standard 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 standard choices.

Another Reproducibility Issue

What are the welfare effects of false beliefs?

The inherent problem: A planner can choose to create conditions that lead someone to hold false beliefs, but can a consumer consciously choose for herself to hold a false belief?

The usual work-around: Assume that beliefs only matter for instrumental reasons.

But this perspective has troubling implications. For example, it means that the government can improve welfare by tricking people.

- Policies that make taxes less salient reduce deadweight losses.

If people prefer to live in a world where their beliefs are accurate, even when the accuracy has no instrumental consequences, then planners' decisions that impact the accuracy of beliefs are irreproducible.

A Third Reproducibility Issue

What are the welfare effects of intertemporal tradeoffs?

The inherent problem: We can’t make choices about our past.

Illustration of why it matters: Let’s say the planner has to make a choice involving a tradeoff between periods t and $t+1$. She might want to consider the possibility that the individual feels differently about that tradeoff at different points in time.

The planner’s choice is reproducible as a choice for the individual in periods 0 through t, so she can take all those perspectives into account.

The planner’s problem is not reproducible for any later period.

Shouldn’t the individual’s preferences about that tradeoff in subsequent periods matter to the planner?

The behavioral critique

Behavioral Principle #1: Behavior depends on the *context of choice*

- Type M: Certain framings trigger *mistakes*

  Challenges Premise #3 – "Implementation Critiques"

- Type C: People don't have preferences they can access – their *judgments are constructed* contextually

  Challenges Premise #1 – "Coherence Critiques"

Behavioral Principle #2: The act of choosing for oneself alters the available options.

Challenges Premise #4 – "Reproducibility Critiques"

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.*

“Judgment Critiques” of Premise 2

- Does behavioral economics challenge Premise 2?
  - Distinguish between *direct judgments* (opinions that pertain to outcomes we care about for their own sake), and *indirect judgments* (alternatives that lead to those outcomes)
  - Behavioral economics does not provide a foundation for challenging direct judgments. Such challenges are “differences of opinion.”
  - In contrast, if an indirect judgment is tainted by a faulty understanding of consequences, we can legitimately question its normative relevance. But that’s an Implementation Critique, not a Judgment Critique.
  - 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: Fix the Standard Approach

Paths Forward: Fix the Standard Approach

- ***Challenge 1:*** How do we accommodate the possibility that people make mistakes (Implementation Critiques)?

- ***Challenge 2:*** How do we accommodate the possibility that people may not have coherent preferences (Coherence Critiques)?

- ***Challenge 3:*** How do we overcome the inherent differences between choices by planners that affect individuals, and choices by those individuals (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

Challenge 1: Mistakes

- Much of the literature tries to tackle Implementation Critiques (mistakes) in isolation.
  - Relax Premise 3 (Implementation) while retaining all the other premises
  - With Premise 1 (Coherent Judgments) retained, we can treat the judgment relation, $\succcurlyeq$, as *true preferences*, and judge well-being accordingly
- The problem with this approach:
  - If we make Premise 1 (Coherent Judgments) the cornerstone of our approach to handling Implementation Critiques (mistakes), then we can't deal with Coherence Critiques (irreducible inconsistency) without abandoning our solution to Implementation Critiques.
  - In effect, the approach requires us to assume that all context dependence is Type M, not Types C or P, whether or not we know this to be true

Challenge 1: Mistakes

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.

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: Mistakes

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")
- For other decision frames, we allow for the possibility that $U(x, f)$ and $V(x)$ diverge (frames that yield "biased choices")

Challenge 1: Mistakes

Nest BRP within a more general recipe:

- **Step 1:** Starting with the choice correspondence $C(X, f)$, identify the *Welfare-Relevant Domain* (WRD) of choice problems by excluding the ones for which choices are mistakes

- **Step 2:** Construct a welfare criterion based on choices within the WRD.

- **Step 3:** Apply the welfare criterion to evaluate the decisions of interest.

For BRP, we use the standard revealed preference relation in Step 2, on the assumption that choices in the WRD are governed by some well-behaved objective function $V(x)$.

Challenge 1: Mistakes

How do we define mistakes, so we can identify them in Step 1?

A common proposal: mistakes are choices that conflict with true preferences (V)

Challenge 1: Mistakes

Problems with the common proposal (mistakes are choices that conflict with “true preferences”):

- 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: Mistakes

Problems with the common proposal (mistakes are choices that conflict with “true preferences”):

- 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?

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

- In the three-step recipe, we don’t recover V (true preferences) until Step 2, so we can’t use it to define or identify mistakes in Step 1.

Challenge 1: Mistakes

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: Mistakes

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”?

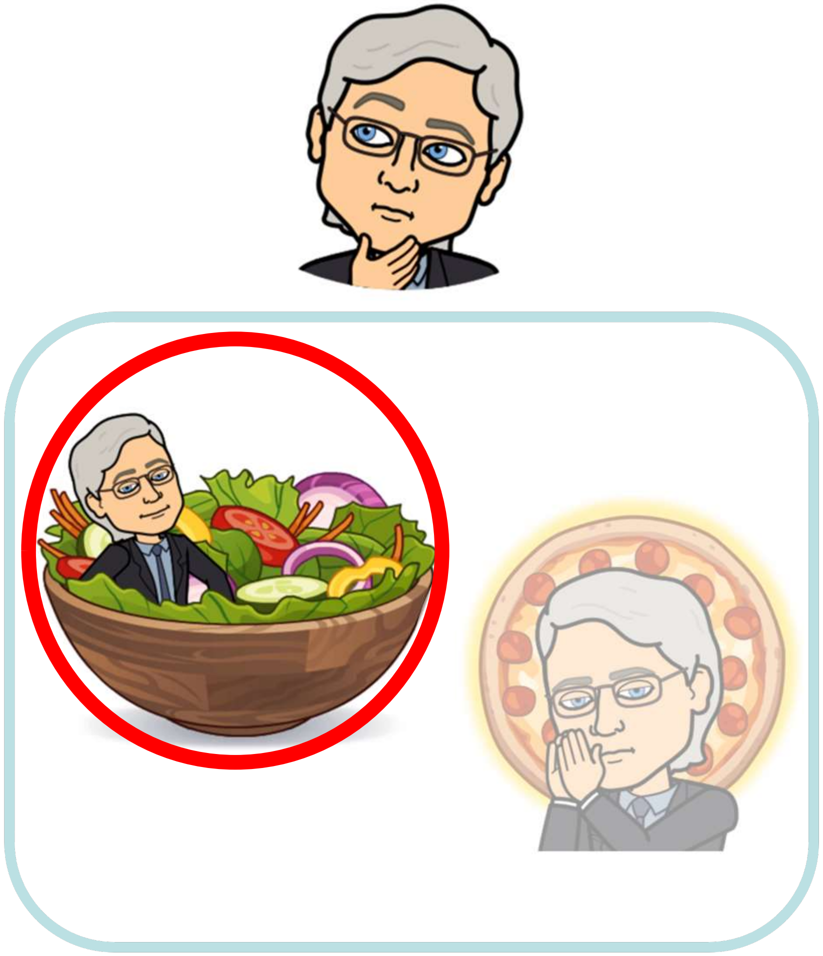
Challenge 1: Mistakes

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

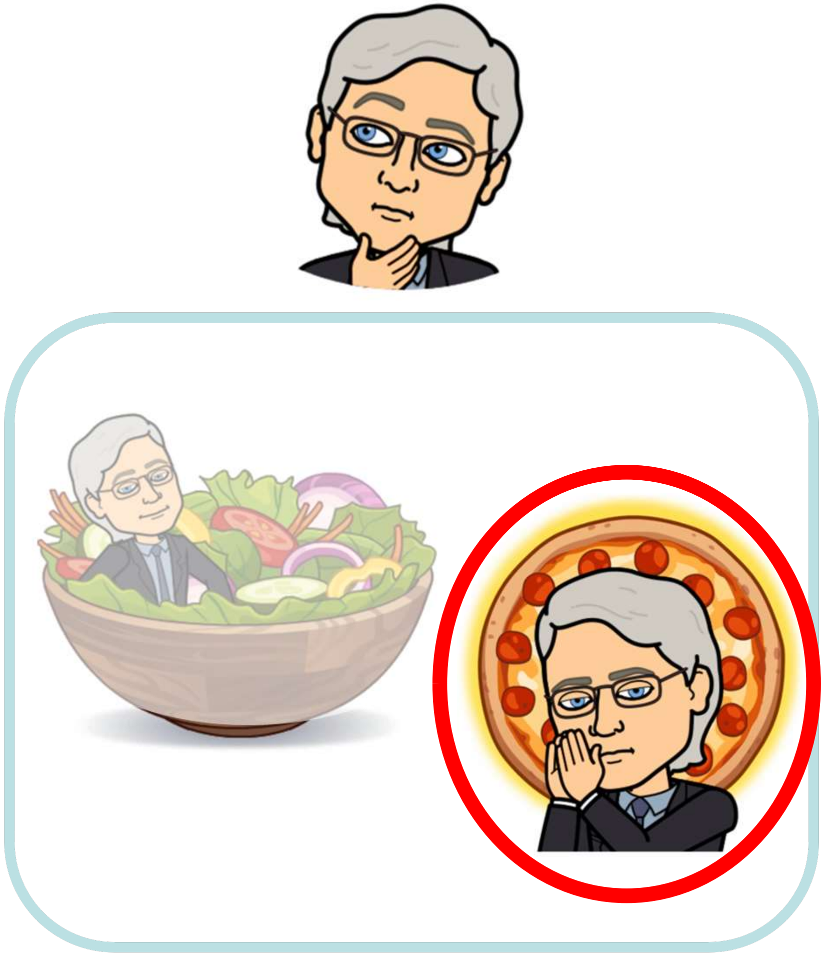
- We need to define a mistake without referring to “true preferences” (V)
- Decisions are logically separable into two components
  - *Characterization:* what options are available, and how do they map to consequences?
  - *Judgment:* which bundle of consequences is better?
- Because Premise 2 precludes us from challenging (direct) judgment, a “mistake” must entail a *Characterization Failure*

How do we interpret “Weakness of Will”?

Intention as of every morning



Choice at lunchtime



Challenge 1: Mistakes

Two possible interpretations of “weakness of will”:

- **Hypothesis #1**: We simply place different weight on the dimensions of our experience in advance and in the moment
  - In that case, what’s the objective foundation for declaring one set of weights right and the other wrong?
  - Using the phrase “weakness of will” is then simply a way of expressing disagreement with the choice, and rationalizing the superimposition of the analyst’s values
- **Hypothesis #2**: In the moment, we blind ourselves to consequences in order to justify indulgence
  - In that case, “weakness of will” describes a form of Characterization Failure (a failure to face facts in the moment)

Challenge 1: Mistakes

How do we identify instances of Characterization Failure? (Bernheim and Taubinsky, 2018)

1. ***Transparency & opaqueness***
   - Are the options and consequences stated directly, or do they require inference?
   - If inference is required, is it complex or subtle?

2. ***Comprehension***
   - Do people actually understand the principles required to infer opaquely stated options and consequences?
   - Are people invoking the principles and deploying the information required to infer opaquely stated options and consequences?
   - Do people consider all available options?
   - Are people making correct inferences?

3. ***Cognitive processes***
   - Are cognitive processes involving attention, memory, forecasting, etc. consistent with the ability to make correct inferences?

Challenge 1: Mistakes

But this escape route from the Circularity Trap is unworkable within the BRP framework due to:

The Goldilocks Problem

After applying objective, evidence-based criteria for identifying mistakes, we may be left with a Welfare-Relevant Domain that is…

1. **Too small:** not enough left in the WRD to recover preferences

2. **Too big:** inconsistencies among choices in the WRD remain, making preference recovery impossible

Challenge 1: Mistakes

Why might the ***too small*** problem arise (not enough left in the WRD to recover preferences)?

- Cognitive limitations may infect most naturally occurring decisions

Challenge 1: Mistakes

Why might the ***too small*** problem arise (not enough left in the WRD to recover preferences)?

- Cognitive limitations may infect most naturally occurring decisions

What can we do about it? (Bernheim and Taubinsky, 2018)

- *Method #1*: Create and implement appropriately reframed decision problems experimentally
- *Method #2*: Extrapolate the missing welfare-relevant choices from other types of decisions using structural models
- *Method #3*: Extrapolate the missing welfare-relevant choices from the decisions of similar individuals who ostensibly avoid characterization failure (the “rational consumer benchmark”)
- *Method #4*: Extrapolate the missing welfare-relevant choices from non-choice data

Challenge 1: Mistakes

Why might the ***too big*** problem arise (inconsistencies among choices in the WRD remain, making preference recovery impossible)?

- **Possibility #1**: True preference may exist, but we may not know enough to identify all the mistakes in Step 1 (*partial purification*)

- **Possibility #2:** Preferences may be constructed contextually, in which case *choice inconsistencies are irreducible*: there is no legitimate way to arrive at a WRD within which choices are entirely consistent.

Challenge 1: Mistakes

BRP leaves no room for resolving the "too big" problem

- To deploy BRP, which tries to recover a well-behaved normative objective function V in Step 2, we can't have inconsistencies.
- In such cases, BRP *requires* us to declare that certain choices are mistakes, regardless of whether we have an objective foundation for doing so.
- If the preference construction view is correct, some of those declarations are necessarily wrong (treating type C context dependence as type M), and therefore incompatible with appropriate general objective principles.
- 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.

Challenge 1: Mistakes

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: Mistakes

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 in Step 2, we won’t have any need for ad hoc criteria in Step 1. Instead, we’ll be free to use general objective criteria for identifying mistakes.

Challenge 2: Irreducible Inconsistency

Challenge 2: Irreducible Inconsistency

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: Irreducible Inconsistency

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: Irreducible Inconsistency

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: Irreducible Inconsistency

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: Irreducible Inconsistency

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: Irreproducibility

Challenge 3: Irreproducibility

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)

Challenge 3: Irreproducibility

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, 2022)

Challenge 3: Irreproducibility

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, 2022)

3. Redefine the consumption bundle in terms of the *mental states* it induces

Challenge 3: The Non-Comparability Problem

- Illustrate strategy #3 by focusing on irreproducibility problems arising from the welfare consequences of choosing (the Non-Comparability Problem)

- A possible solution: make welfare evaluations based on measures of Self-Reported Well-Being (SRWB) – reports of happiness/satisfaction
  - Solution proposed by Koszegi & Rabin (2008)

- Unfortunately, SRWB raises other conceptual issues, including the *Aggregation Problem*.
  - When we answer a question about "happiness" or "satisfaction," we are not "reading a meter." We have to aggregate over various meter readings associated with various dimensions of experience (e.g., time, states of nature). Aggregation is then based our operational understanding of the word's definition, rather than according to a normative principle. ("Linguistic aggregation")

Challenge 3: The Non-Comparability Problem

	Choice-Based Methods	*SRWB Methods*
Non-Comparability Problem	YES	NO
Aggregation Problem	NO	YES

Challenge 3: The Non-Comparability Problem

	Choice-Based Methods	*SRWB Methods*
Non-Comparability Problem	YES	NO
Aggregation Problem	NO	YES

A hybrid method: Use choice-based methods to overcome the Aggregation Problem, while using SRWB methods to overcome the Non-Comparability Problem (Bernheim, Kim, Taubinsky, 2023)

A Hybrid Method

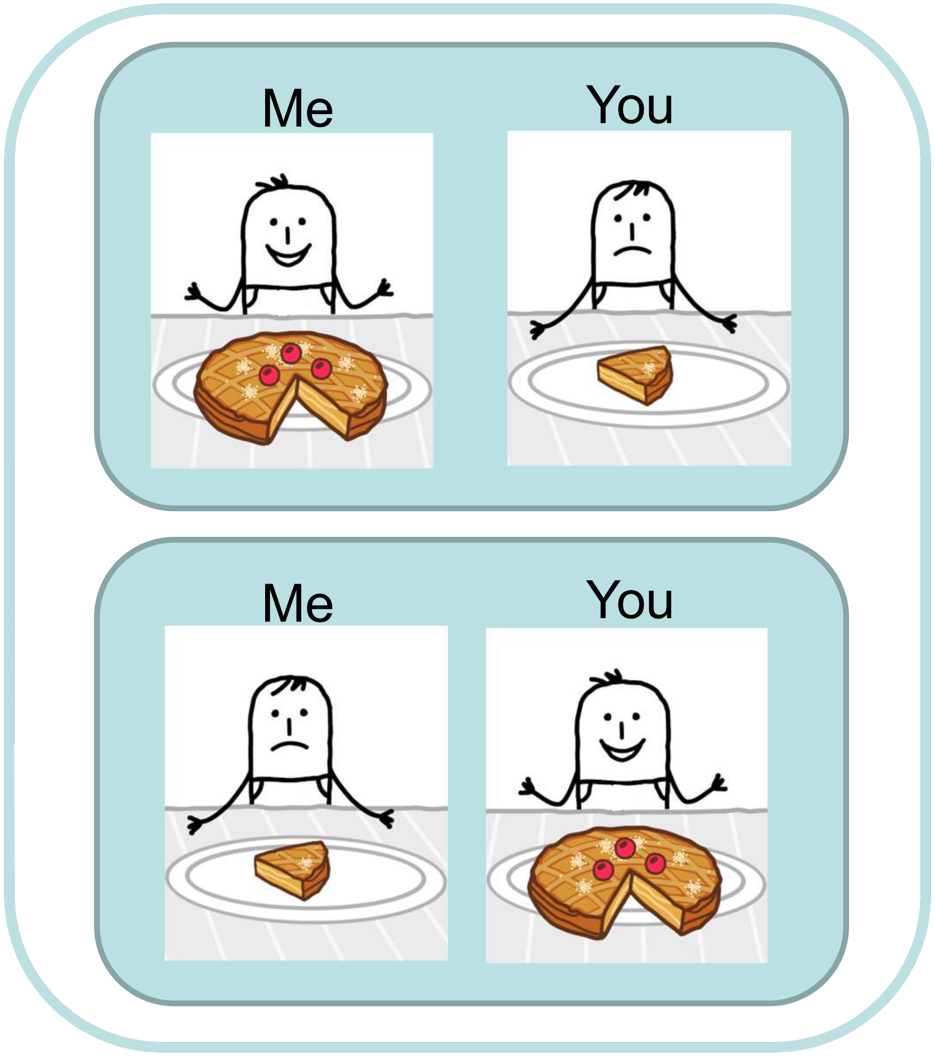
- **Premise:** people value their options based on the mental states the options induce

- The Non-Comparability Problem then arises because the nature of the alternatives to an option can change the mental states that option induces

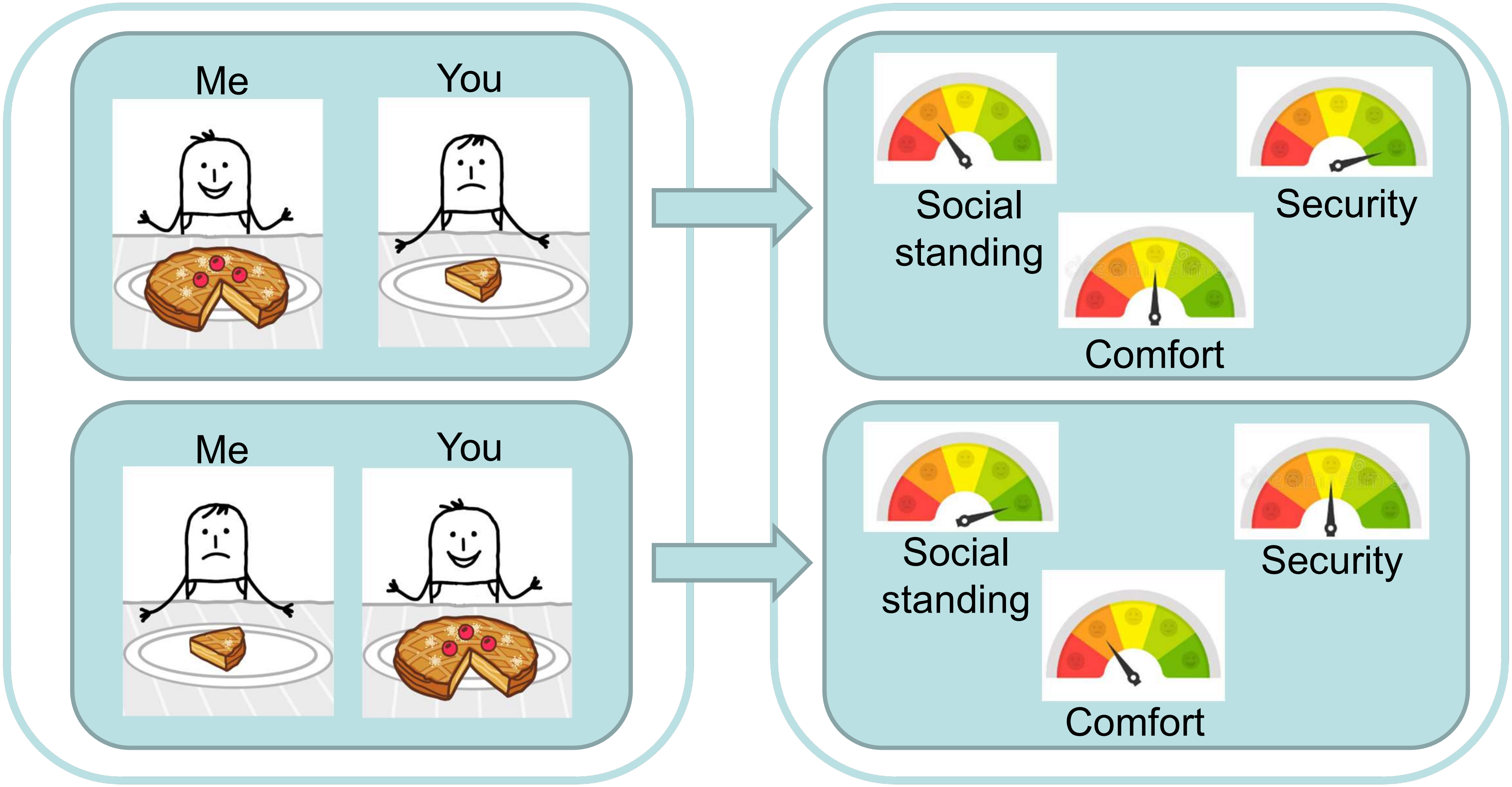
  – Example: Choosing an option that benefits me alone may induce feelings of guilt, but only if there are alternatives that benefit others.

- If we have good proxies for the mental states that options induce in different contexts, then we can use choices to recover preferences over mental state bundles, and then use those preferences to guide the planner's choices.

A Hybrid Method



A Hybrid Method



Me

You

Social standing

Security

Comfort

Me

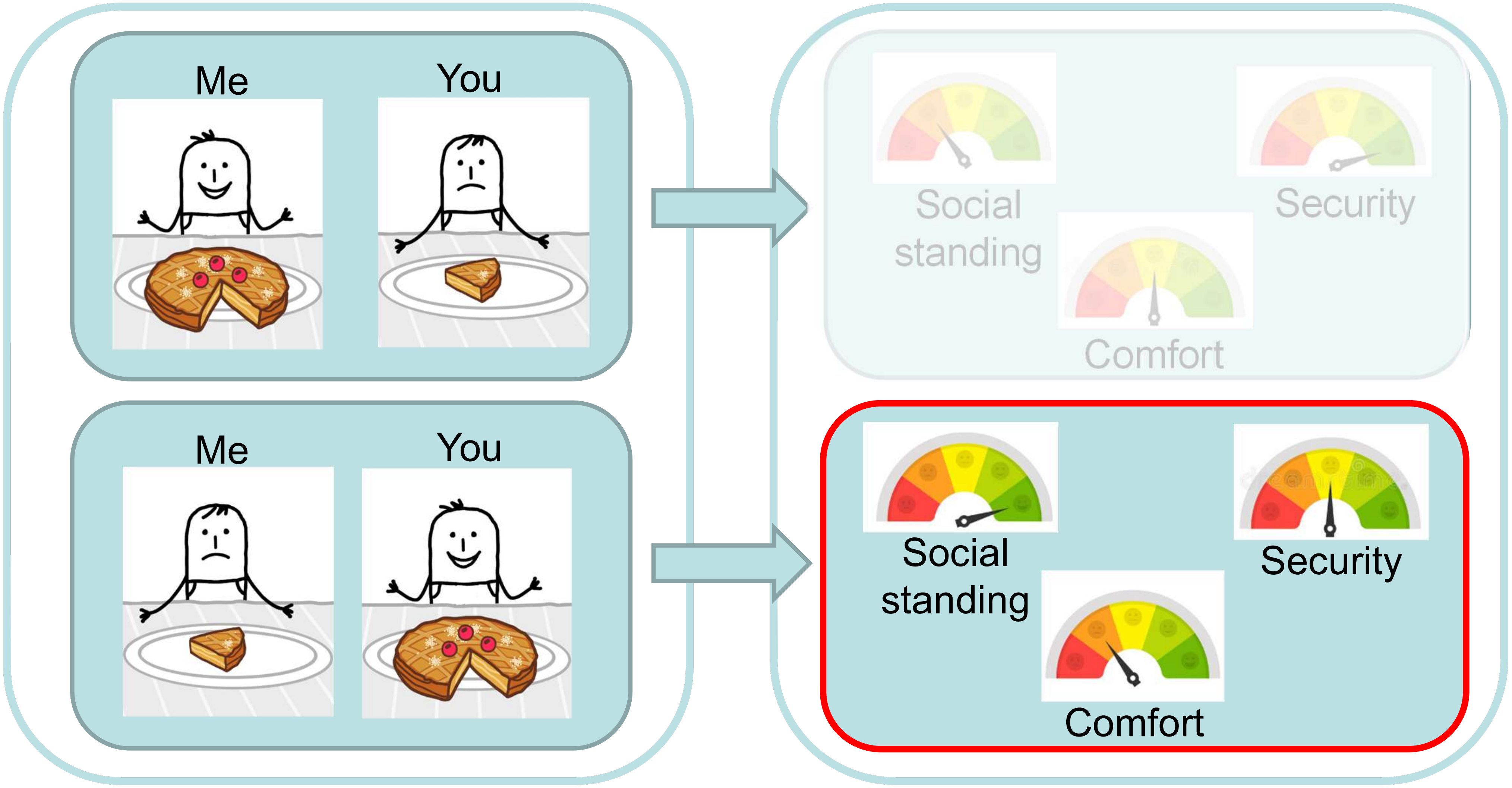
You

Social standing

Security

Comfort

A Hybrid Method



Me

You

Social standing

Security

Comfort

Me

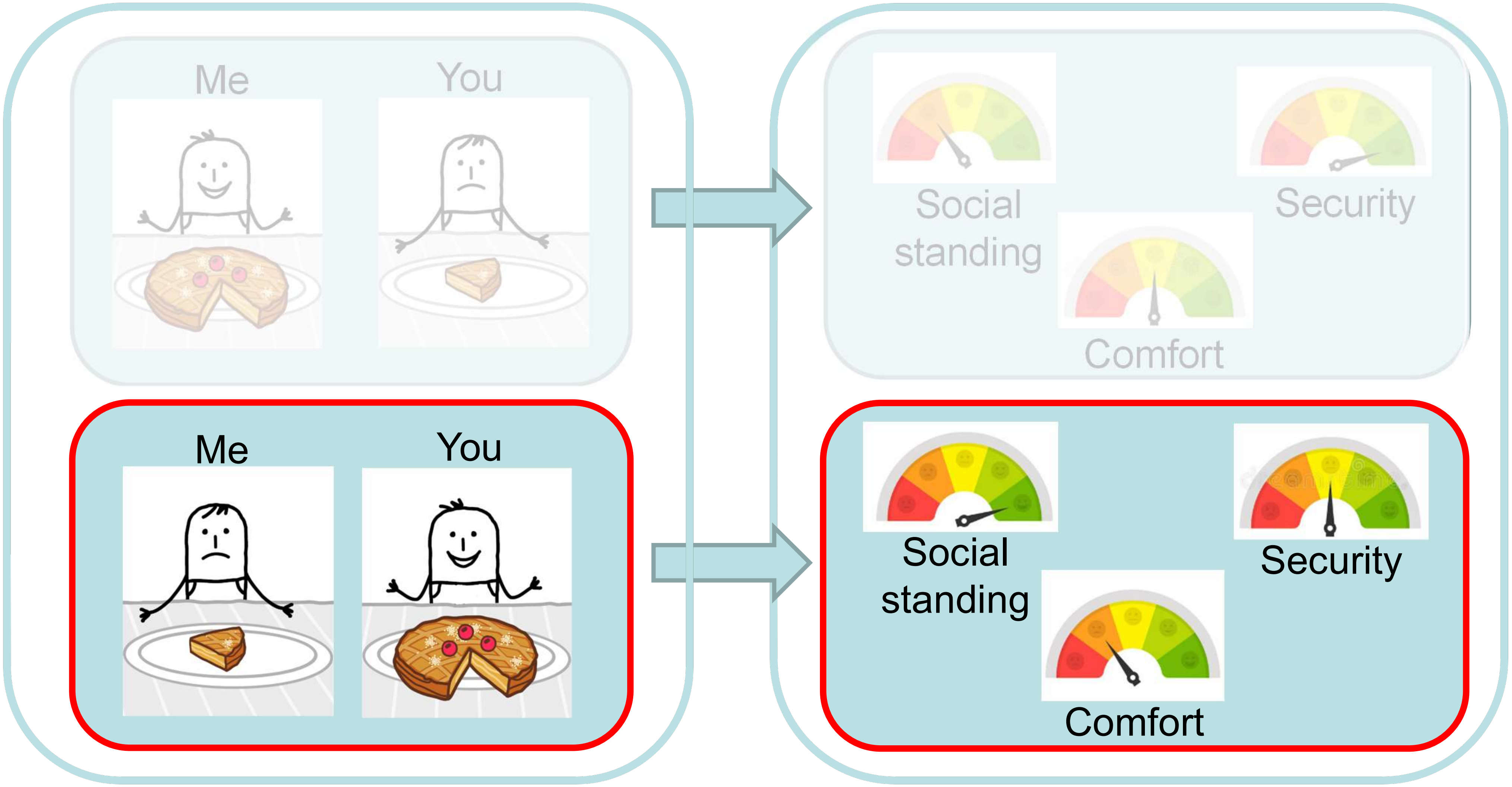
You

Social standing

Security

Comfort

A Hybrid Method



Me

You

Social standing

Security

Comfort

Me

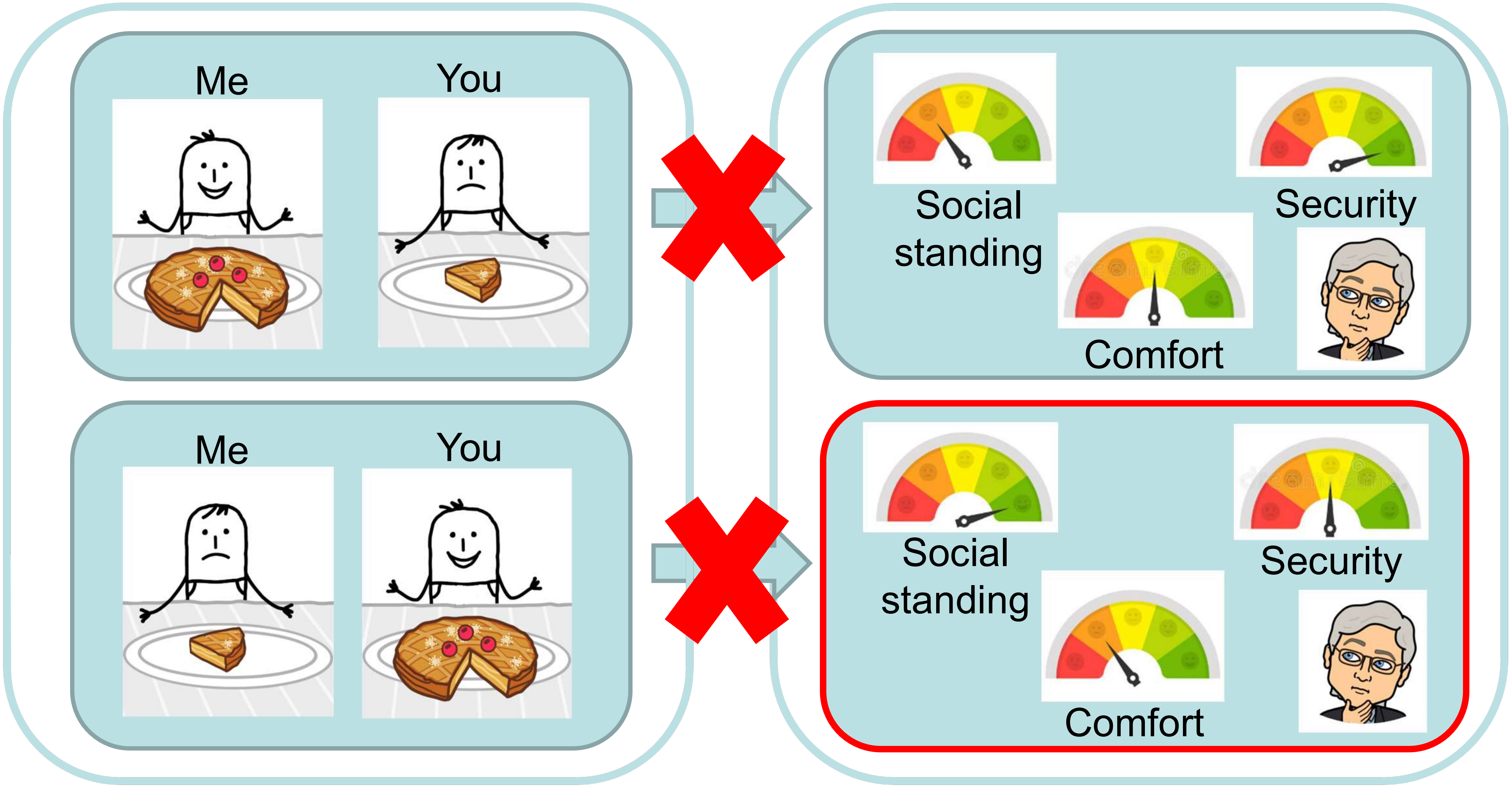
You

Social standing

Security

Comfort

A Hybrid Method



Me

You

Social standing

Security

Comfort

Me

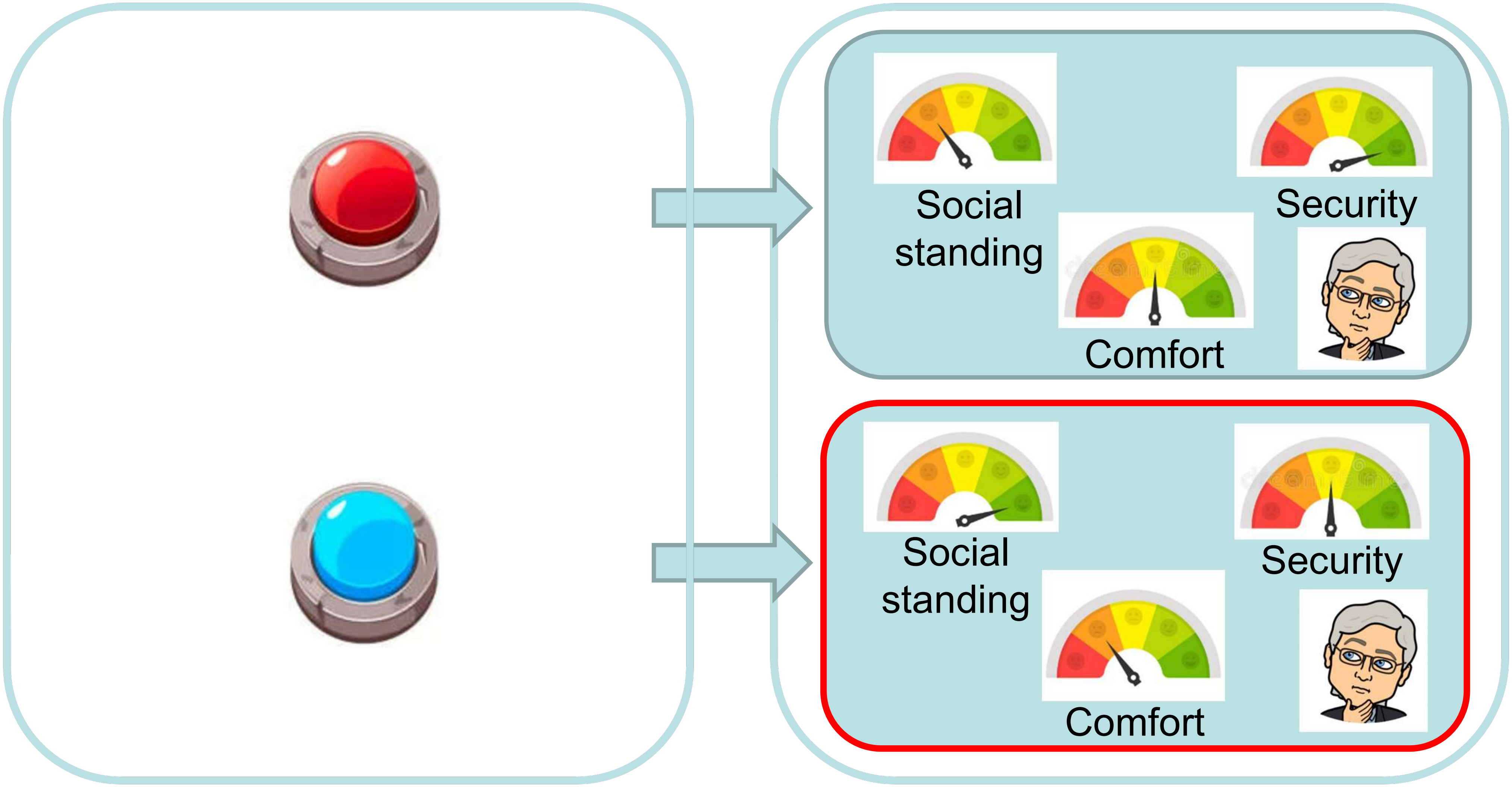
You

Social standing

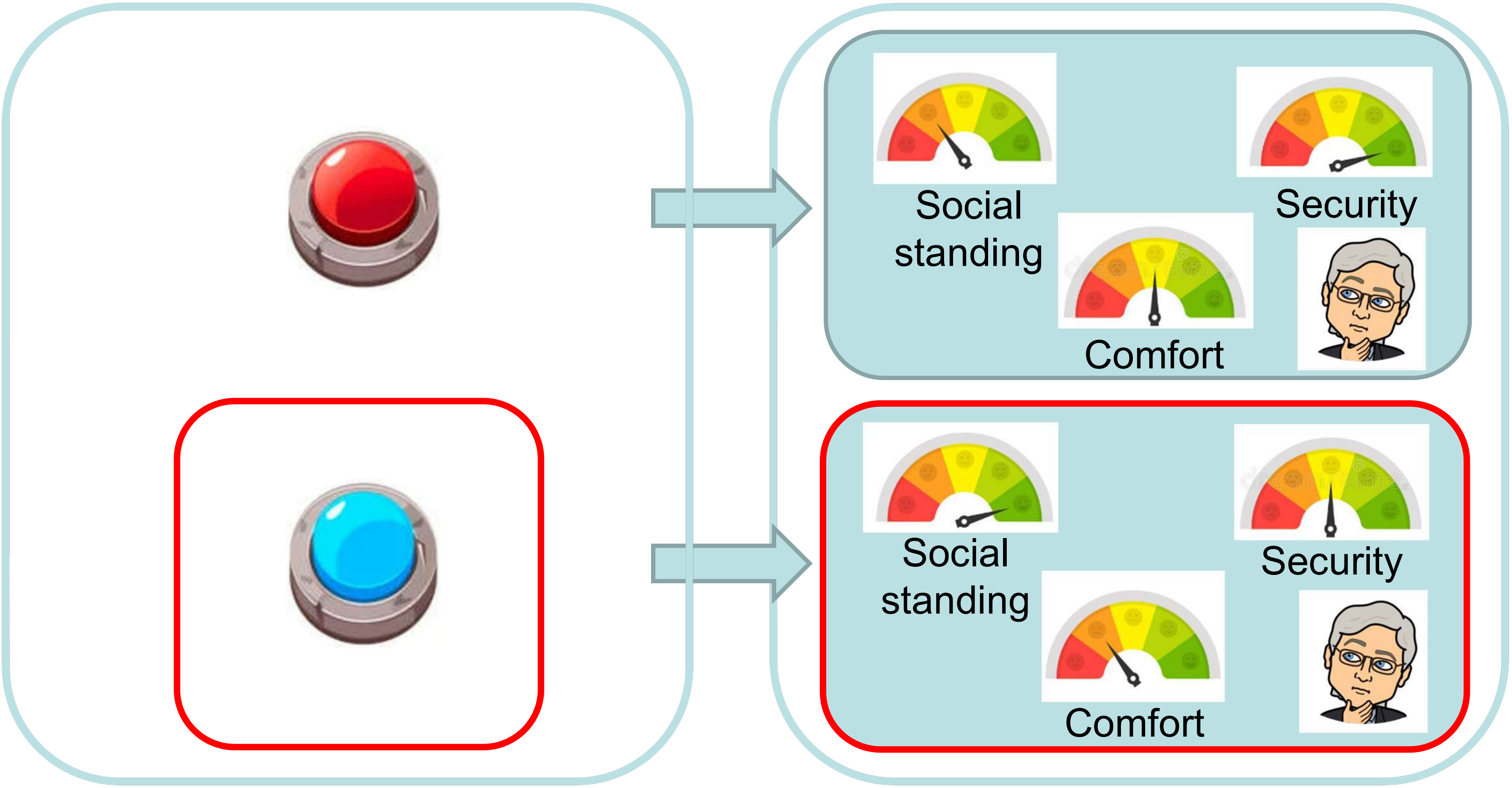
Security

Comfort

A Hybrid Method



A Hybrid Method



Social standing

Security

Comfort

Social standing

Security

Comfort

A Hybrid Method

Application #1: Dictator games

- One party chooses between two divisions of a fixed prize, one selfish, the other generous
- Key Result #1: The Non-Comparability Problem is real
  - People are better off with the prosocial option if someone chooses it for them than if they choose if for themselves (temptation/wistfulness)
  - People are better off with the payout-maximizing option if someone chooses it for them than if they choose it for themselves (guilt)
- Key Result #2: A Planner who mimics the DG choices in order to benefit the Dictator will be too generous

A Hybrid Method

Application #2: Opt-out games

- Each game gives the subject a choice between playing a DG and a fixed payment (where the receiver learns nothing if the fixed payment is chosen)

- The opt-out game is metachoice. It’s supposed to “price out” the dictator game.

- **Key result #3**: The net benefit of being assigned to play a DG is smaller than the net benefit of choosing to play a DG. Therefore, a Planner who evaluates DGs according to the opt-out valuations will create too many DGs.

A Hybrid Method

Extensions: This “mental states” approach offers a potential solution to other vexing welfare puzzles.

Example: how do we evaluate welfare in settings with *endogenous preferences*? (work in progress)

Start with the “chosen preferences” model of Bernheim et al. (*AER*, 2021):

- At time t, the individual holds a *worldview*, α_t. The utility derived from the consumption stream,$x_0, \dots, x_T$ is then $U(x_0, \dots, x_T, \alpha_t)$. The individual may (or may not) also place weight on $U(x_0, \dots, x_T, \alpha_{t'})$ for $t' \neq t$.

- Problem: which worldview(s) do we use to evaluate welfare?

A reformulation: The individual has a single objective function $W(m_0, \dots, m_t)$, where m_t, the vector of mental states at time t, is given by $m_t = M(x_0, \dots, x_T, \alpha_t)$. W is then an unambiguous welfare criterion.

Concluding Remarks

- **Challenge 1:** *Mistakes*
- **Challenge 2:** *Irreducible Inconsistency*
- **Challenge 3:** *Irreproducibilities*

Concluding Remarks

- **Challenge 1:** *Mistakes*
- **Challenge 2:** *Irreducible Inconsistency*

- Identify characterization failures
- Apply the unambiguous choice criterion

Concluding Remarks

- **Challenge 1:** *Mistakes*
- **Challenge 2:** *Irreducible Inconsistency*

- Identify characterization failures
- Apply the unambiguous choice criterion

- **Challenge 3:** *Irreproducibilities*

- Surrogate choices
- Statistically corrected hypothetical responses
- Recover preferences over mental states